

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Processing of dynamic ripple stimuli in the cat inferior colliculus

Permalink

<https://escholarship.org/uc/item/43h248qz>

Author

Escabí, Monty A.

Publication Date

2000

Peer reviewed|Thesis/dissertation

Processing of dynamic ripple stimuli in the cat inferior colliculus:
an ecological approach to sound processing.

by

Monty A. Escabí

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

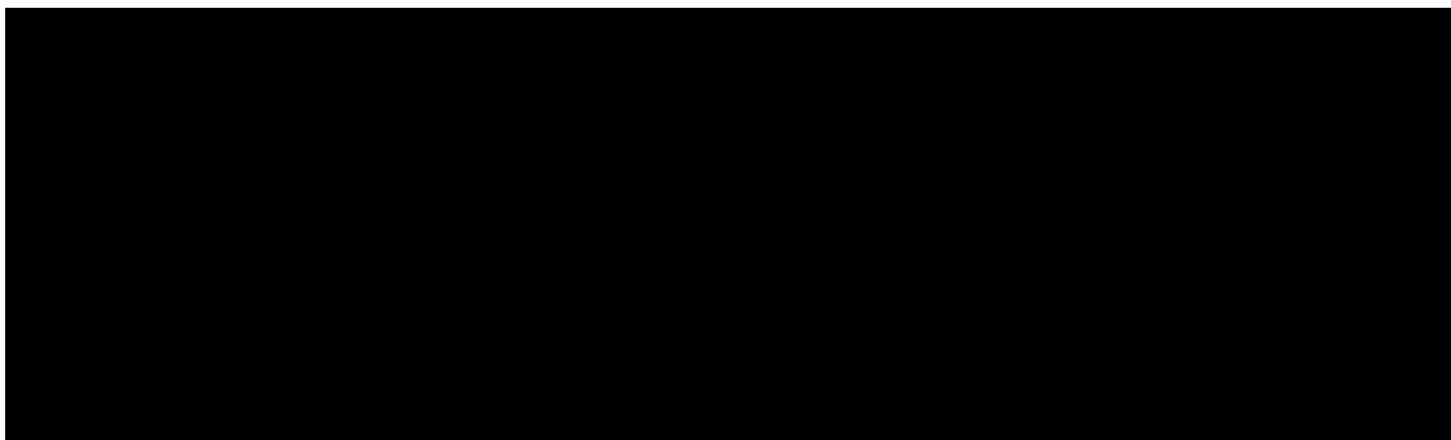
GRADUATE DIVISIONS

of the

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

and

UNIVERSITY OF CALIFORNIA BERKELEY



Date

University Librarian

Degree Conferred:

Copyright (2000)

by

(Monty A. Escabi)

UCSF LIBRARY

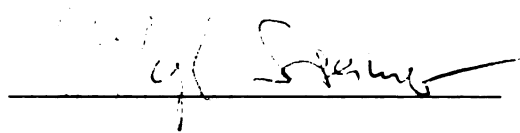
To my wife Lisa

UCSF LIBRARY

Abstract

Processing of dynamic ripple stimuli in the cat inferior colliculus: an ecological approach to sound processing. by Monty A. Escabí: Natural sounds, such as speech and vocalizations, are characterized by time-varying spectra that give rise to distinct temporal periodicities, frequency transitions, and spectral resonances. These structural features are decomposed by the primary sensory epithelium and give rise to a number of perceptual attributes. Given the complexity of the auditory neuronal network and the fact that the brain is in general extremely nonlinear, it is increasingly clear that simple acoustic stimuli (e.g. pure tones and noise) can not be used to identify natural sounds processing strategies. To understand how complex sound attributes are represented in the brain, statistical properties of the spectro-temporal envelope of natural sounds (including speech, vocalizations, environmental noise, and music) were studied in detail. Ensemble statistics are marked by robust spectrographic correlations, logarithmic contrast, and stimulus dynamics which are closely related to a number of perceptually relevant acoustic variables. Hypothetically, these higher-order stimulus attributes can be utilized by the auditory system for efficient sound processing and across-category discrimination. To test this hypothesis, neuronal recordings were performed in the central nucleus of the inferior colliculus (ICC) of cats using synthetic ripple stimuli that incorporate the observed statistical attributes. Using spectro-temporal receptive field (STRF) methods, it is found that ICC neurons efficiently utilize these higher-order stimulus attributes for sound processing. Populations of neurons are distinguished based on their degree of feature selectivity and their ability to time-lock to the spectro-

temporal envelope. A hierarchy of functionally distinct neuronal types is revealed based on three possible neuronal codes. Further evaluation reveals that the operating range of ICC neurons is physically matched to the spectro-temporal energy distributions observed in natural sounds. When tested with stimuli that mimic natural sounds, neurons show contrast tuning and improved spectro-temporal coding at time-scales comparable to the neuron's receptive field. These findings establish a link between acoustic ecology, acoustic sound structure, and neuronal processing. Such processing strategies make use of structural regularities in natural sounds and likely underlie human perceptual abilities.



Christoph E. Schreiner, Chair

UCSF LIBRARY

Table of Contents

1 Spectro–Temporal Attributes of Natural Sounds	1
Abstract	2
1 Introduction	3
2 Spectro–Temporal Stimulus Decomposition of Sounds	6
3 Spectro–Temporal Decomposition of Sounds Using an Auditory Filter Bank	7
4 Filter Selection and Design	13
5 Signal Decomposition and Envelope Extraction	17
6 Spectrographic Envelope	19
7 Low Order Stimulus Statistics – The Power Spectrum	23
8 Across Band Correlations of Natural Sounds	25
9 Spectro–Temporal Contrast of Natural Sounds	31
10 Contrast and Intensity Dynamics of Natural Sounds	37
11 Contrast and Intensity Ensemble Statistics	43
12 Discussion and Conclusion	53
13 References	61
2 Rippled Noise Stimulus Design: Theoretical and Ecological Considerations	65
Abstract	66
1 Probing the Auditory System with Simple Sounds	67
2 Nonlinear Auditory Processing	69

UCSF LIBRARY

3 Neuroethology Versus the Systems Approach	71
4 Stimulus Requirements for Deriving <i>STRFs</i>	78
5 Testing for Nonlinearity	84
6 Correlated Versus Uncorrelated Sounds	91
7 The Dynamic Ripple and Ripple Noise Stimuli	95
8 Ripple Stimulus Design	99
9 Design of the Dynamic Moving Ripple Envelope	101
10 Parameter Design for the Dynamic Ripple	104
11 Design of the Ripple Noise Envelope	110
12 Dynamic Ripple and Ripple Noise Spectro–Temporal Correlation Statistics	113
13 Dynamic Ripple Local Approximation	114
14 Dynamic Ripple Local Autocorrelation Function	117
15 Dynamic Ripple Global Autocorrelation Function	124
16 Ripple Noise Local Autocorrelation	129
17 Ripple Noise Local Autocorrelation: Effects of Finite Number of Dynamic Ripples ($L=16$) and Finite Window Size (σ_x and σ_r)	133
18 Ripple Noise Global Autocorrelation	144
19 Compressed Ripple Noise Autocorrelation Function (Effects of $f(x)$)	144
20 Dynamic Ripple Cross–Channel Correlations	150
21 Ripple Noise Cross–Channel Correlations	154
22 Dynamic Ripple and Ripple Noise Higher–Order Correlations	155
23 Dynamic Ripple and Ripple Noise Higher–Order Correlation Strength	159

24 Ripple Noise and Dynamic Ripple Stimulus Alterations	162
25 Harmonic Ripple	163
26 Comodulated Ripple	164
27 1/f Envelope Spectrum Ripple	166
28 Conclusion	167
29 References	170
3 Nonlinear Spectro–Temporal Processing and Feature Selectivity	176
Abstract	177
1 Introduction	178
2 Methods	181
3 Binaural Receptive Fields	186
4 Testing For Significance of the STRF.	191
5 STRF Comparisons – Moving Ripple versus Ripple Noise.	194
6 STRF Similarity for the Moving Ripple and Ripple Noise Stimulus	200
7 Response Strength – Moving Ripple versus Ripple Noise.	203
8 Quantifying Response Specificity to Structured Sound Patterns	214
9 Similarity Index Histogram of s– and f–Neurons.	223
10 Quantifying Feature Selectivity.	227
11 Binaural FSI Metric.	231
12 Distribution of FSI Values.	236
13 Independence of the Binaural SI Distribution.	238
14 The Ripple Transfer Function.	240

15 Non–phase–locked Neurons (C–Neurons)	247
16 Spectro–Temporal Population Statistics.	251
17 Discussion.	253
18 Conclusion.	261
19 References	264
4 Natural Contrast and Intensity Processing	272
Abstract	273
1 Introduction.	274
2 Contrast Statistics of Natural Sounds.	276
3 Contrast Versus Intensity Response Characteristics.	283
4 Independence of Response to Intensity and Contrast Cues	289
5 Effects of Envelope Statistics on Spectro–Temporal Coding	294
6 Spike Timing Precision and Response Reproducibility.	304
7 Discussion and Conclusion	309
8 Methods	317
9 References	328

List of Figures

Chapter 1

1 Cochlea frequency versus cochlear position function.	10
2 Cochleotopic filter bank.	16
3 Power spectrum of natural sounds.	24
4 Computing the across-channel correlation matrix.	26
5 Spectro-temporal correlations of speech and vocalizations.	28
6 Spectro-temporal correlations of environmental sounds.	29
7 Spectro-temporal correlations of music and white-noise.	30
8 Linear versus logarithmic spectro-temporal envelope.	34
9 Linear contrast statistics of natural sounds.	35
10 Decibel contrast statistics of natural sounds.	36
1 1 Constructing the time-dependent contrast distribution.	38
1 2 Time-varying contrast of environmental sounds.	41
1 3 Time-varying contrast of speech and vocalizations.	42
1 4 Time-varying contrast of music.	42
1 5 Parametrizing the time-dependent contrast distribution.	44
1 6 Intensity versus contrast statistics of natural sounds.	46
1 7 Contrast statistics.	47
1 8 Contrast dynamics.	48
1 9 Coherent contrast and intensity statistics.	52

Chapter 2

1 Relationship between the Voltera and Wiener kernels.	89
2 Dynamic ripple stimulus.	96
3 Ripple noise stimulus.	98
4 Ripple frequency and modulation rate trajectories.	102
5 Ripple phase time trajectory and distribution.	108
6 Dynamic ripple parameter statistics.	109
7 Amplitude distribution of the dynamic ripple and ripple envelope	112
8 Dynamic ripple local–autocorrelation function.	124
9 Ripple noise local–autocorrelation function	132
10 Ripple noise local–autocorrelation error.	143
11 One dimensional compressed ripple noise counterpart.	146
12 Autocorrelation and spectrum of the compressed one dimensional ripple noise	149
13 Compressed ripple noise autocorrelation function.	150
14 Correlation strength: dynamic ripple versus ripple noise.	162
15 Comodulated dynamic ripple	165
16 1/f ripple noise.	167

Chapter 3

1 Binaural receptive fields	189
2 <i>STRFs</i> obtained during bilateral stimulation.	190
3 Significant <i>STRF</i> distribution.	194

4 STRFs of s–neurons.	196
5 STRFs of f–neurons.	197
6 Similarity index distribution.	203
7 Firing rate and differential response parameters to the ripple noise and dynamic ripple stimuli.	209
8 Interrelation among the mean firing rate and differential response parameters for the dynamic ripple and ripple noise.	210
9 Rate and amplitude similarity index statistics.	213
10 Relationship between the covariance and the similarity index	219
11 Computing the similarity index histogram.	221
12 Similarity index distributions of s–neurons.	225
13 Similarity index distributions of f–neurons.	226
14 Feature selectivity index (<i>FSI</i>).	229
15 Binaural feature selectivity index.	234
16 Joint similarity index cumulative distribution.	235
17 <i>FSI</i> statistics.	237
18 Separability of the binaural similarity index histogram.	240
19 Ripple transfer function and histogram.	243
20 Ripple transfer functions	246
21 Non–phase locking neurons	250
22 Ripple density and modulation rate statistics	253

Chapter 4

1 Decibel and logarithmic spectrographic envelope.	278
2 Linear contrast statistics.	281
3 Logarithmic contrast statistics.	282
4 Liner ripple noise spectro-temporal envelope.	284
5 Logarithmic ripple noise spectro-temporal envelope	285
6 Contrast versus intensity response characteristics.	287
7 Firing rate reduction for contrast non-monotonic neurons	289
8 Separability of the contrast-intensity response function.	292
9 Separability index statistics of the contrast-intensity response function	293
10 Relationship between the contrast-intensity response function and the <i>STRF</i>	295
11 Population similarity index statistics	300
12 Population RSI and ASI statistics: 30 dB versus <i>Lin.</i>	301
13 Population RSI and ASI statistics: 60 dB versus <i>Lin.</i>	302
14 Population RSI and ASI statistics: 60 dB versus 30 dB.	303
15 Spike precision and reproducibility response rasters as a function of contrast.	307
16 Mutual information statistics.	308

Spectro-Temporal Attributes of
Natural Sounds

UCSF LIBRARY

Abstract

The time-varying spectrum of natural sounds and many man made sounds (e.g. music) is marked by spectral resonances, edges, temporal modulations, and frequency transitions, all of which give rise to distinct perceptual qualities. This study seeks to provide insight into the ensemble characteristics of natural sounds by analyzing high-order statistics of the time-varying spectrum of speech, animal vocalizations, music, and background sounds (wind, rain etc.). Low-order statistics such as the modulation spectrum and temporal contrast (Attias and Schreiner 1998a) of natural sounds show invariant statistics across various sound ensembles. Thus, such low-order descriptors can not be used directly to distinguish and classify sounds. We show that ensembles of natural sounds segregate if one takes into account cross-channel spectrographic correlations, spectrographic contrast, and stimulus dynamics. The presented findings allow us to define possible perceptually relevant acoustic variables and mechanisms for across-category discrimination of natural sounds.

UCSF LIBRARY

1.1 Introduction

In complex acoustic environments speech, vocalizations, and other competing sounds often do not occur in isolation. Despite this, the auditory system of humans and mammals is capable of distinguishing sounds in less than optimal conditions (Moore 1997). Mammals have evolved elaborate neural systems for analyzing the time-varying spectrum of natural sounds, classifying sounds, and for distinguishing distinct perceptual qualities present in natural sounds (e.g., pitch, timbre). Presumably, the evolved processing strategies have been evolutionary influences by ecological constraints, and are consequently, efficiently adapted for processing natural sounds (Rieke *et al.* 1995; Attias and Schreiner 1999b; Nelken *et al.* 1999).

Given this basic hypothesis, one approach of studying auditory function, is to first study in detail the structural characteristics of the acoustic ecology. This approach has been employed for naturally occurring visual scenes (Ruderman and Bialek 1994; Dong and Atick 1995; Ruderman 1997), for acoustically specialized mammals (Simmons, Howell, and Suga 1975), and to a much lesser degree for natural sounds in general (Attias and Schreiner 1999b; Nelken *et al.* 1999). In the case of the echolocating bat and songbird auditory systems (Simmons, Howell, and Suga 1975; Theunissen *Et. al.* 2000), the acoustic ecology which is most often considered is largely limited to a small set of highly stereotyped sounds that are prevalent in the animals vocal repertoire and that elicit a precise behavior. The relevant acoustic ecology of less specialized animals has only been studied to a small extent (Attias and Schreiner 1999a; Nelken, Rotman, and Yosef 1999) and it is, in general, not well understood.

In most animals the task of deciphering the relevant acoustic parameters in the

animals acoustic ecology is not a trivial task. Unlike the bat and songbird species, this is in part attributed to the fact that a direct link between physical properties of a sound, the animal's behavior, and physiology can not be easily established. Such is the case for acoustically nonspecialized animals such as the cat, rabbit, and possibly even for primates. In other animals, such as the echolocating bat and the barn owl, the search for relevant parameters is greatly simplified since these animals show a direct link between a sound and behavior.

Because of the general lack of understanding of natural sounds, attempts at understanding auditory function in most mammals is largely limited to studies that use narrow band acoustic stimuli. In the case of pure tones, these sounds excite only a small fraction of the primary sensory epithelium and, consequently, a small fraction of the auditory neuronal network. In the special cases where broadband stimuli are used these are essentially limited to white noise and clicks (e.g., Young and Brownell 1976; Yin, Chan, and Irvine 1986). Natural sounds are clearly not well described by these basic attributes. Instead, most natural sounds are broad-band, spectro-temporally complex, and nonstationary. Consequently, the excitation patterns produced by natural sounds on the cochlea and along the auditory neural network are significantly more complex than for the simple stimuli that are used to study the auditory system.

Given this general observation, we seek to understand the patterns of excitation that are produced by natural sounds at the level of the auditory periphery and at central auditory stations. This approach is dually motivated: first we seek to identify robust statistical characteristics which are prevalent in natural sounds and that allow one to distinguish between various classes of natural sounds (e.g. speech versus a background

sound such as running water). Secondly, we would like to identify structural characteristics of natural sounds that may be of perceptual relevance and which the auditory system may use for efficient sound encoding. Since low-order statistics of natural sounds show invariant modulation and contrast statistics (Attias and Schreiner 1998a), they alone are insufficient for distinguish among classes of natural sounds. Analysis of higher-order comodulation statistics, however, reveals that vocalizations can be distinguished from background sounds (Nelken, Rotman, and Yosef 1999).

In this study, the spectrographic statistics of natural sounds, including spectro-temporal correlations, contrast statistics, and stimulus dynamics, are therefore examined in detail. First, it is shown that natural sounds have strong spectro-temporal correlations across octave spaced frequency channels. Comparisons among speech, vocalizations, background sounds, and music reveals that sounds can be segregated based on the degree of correlation. Thus by analyzing spectro-temporal correlations, the auditory system can potentially distinguish among classes of natural sounds. Secondly, we show that the spectrographic representation of natural sounds has local amplitude fluctuations which span several orders of magnitude. Using perceptually relevant time-scales (for loudness perception) to model the dynamics and statistics of natural sounds, we show that these amplitude statistics can be use to distinguish among classes of natural sounds. These findings allow us to define possible perceptually relevant variables and allow us to identify structural characteristics of natural sounds which the auditory system may use for efficient sound encoding. These ideas are further tested and verified directly using *electrophysiologic* recording methods in chapters 3 and 4.

1.2 Spectro–Temporal Stimulus Decomposition of Sounds

The peripheral auditory system is characterized by a tonotopically arranged set of hair–cells which are individually tuned to a small range of frequencies (Lieberman 1982; Greenwood 1990). Upon arriving the cochlea, incoming sounds are decomposed by a bank of tonotopically arranged frequency channels into a complex spectro–temporal excitation pattern (Sachs and Young 1979; Delgutte and Kiang 1984; Shamma 1985; Carney and Geisler 1986; Geisler and Gamble 1989). Since this spectro–temporal decomposition defines the inputs for higher–order processing centers in the brain, it is useful to understand the basic parameters that are of possible relevance and the constraints that are imposed on the auditory neuronal network by incoming natural sounds.

Time–frequency representations, such as the spectrogram, have a long history in engineering, the physical, and biological sciences. Initially, the spectrogram was motivated by the need to devise physiologically plausible models of speech production and perception during the advent of telephone and other communications systems. A key aspect of the cochlear transformation is the conversion of a one dimensional acoustic pressure waveform, by an array of bandpass filters (i.e. the cochlea), into a spectro–temporal excitation pattern. The resulting neuronal discharge pattern describes the changes in the stimulus spectrum as a function of time, much like time–frequency representations used to analyze dynamic signals (Cohen 1995). This spectro–temporal neuronal discharge pattern is relayed by the eighth nerve to the cochlear nucleus, which serves as the inputs for higher–order processing centers in the brain. Thus, a key question in auditory neuroscience deals with trying to understand how such inputs are utilized by

the brain for efficient sound encoding, sound recognition, and source segregation. We specifically ask: what are the relevant spectro-temporal parameters for this stimulus representation? And how is this complex excitation pattern further processed by the brain? To understand this, it is first useful to understand statistical characteristics of this excitation pattern in detail.

1.3 Spectro-Temporal Decomposition of Sounds Using an Auditory Filter Bank

The spectro-temporal decomposition of sounds performed by cochlea is characterized by octave spaced filters of nearly equal resolution (Lieberman 1982; Kiang *et al.* 1965). Numerous physiologically motivated auditory filter bank models have been designed which mimic the acoustic stimulus decomposition performed by the cochlea. These spectro-temporal decomposition are used in a variety of application, ranging from design of auditory filter models (Carney 1993; Wang and Shamma 1995a 199b; Jenison *et al.* 1991), to sound compression, and sound analysis algorithms (Picone 1997).

Here, an alternative spectro-temporal filter bank decomposition is designed that satisfies two essential properties of the cochlear filter decomposition. It is required that the component filters have 1) logarithmically spaced center frequencies and 2) constant quality factor. The latter requirement essentially demands that the component filters have equal resolution (bandwidth) on an octave frequency axis much like cochlear filter resolutions. For completeness a slightly more refined filter model is used. This model takes into account the fact that the frequency spacing and filter bandwidths along the basilar membrane deviate slightly from this ideal logarithmic scenario at frequencies

below about 1000 Hz (Liberman 1982; Greenwood 1990). The cochlear center frequency versus cochlea position equation provided by Greenwood (Greenwood 1990) is used to model the spacing of the auditory filter bank at low and high frequencies.

Along the cochlear partition, the inner hair cell (IHC) center frequency is provided by (Greenwood 1990)

$$f = A(10^{ax} - k) \quad (1.1)$$

where x is the normalized cochlear distance (normalized between zero and one) from the apex (stapes) to the base of the basilar membrane. The constants A , a , and k are species dependent. For humans $A=165.4$, $a=2.1$ ($a=0.06$ if x is expressed in millimeters, total length of about 34 mm), $k=0.88$ whereas for the cat $A=456$, $a=2.1$, and $k=0.8$. At intermediate frequency values, the frequency versus position curves for the human and cat are for the most part identical (Fig. 1A) with a fixed offset along the cochleotopic, x , axis. They only differ at the extremities where the lower and upper frequency limit are determined by the constant A . For humans the lower and upper frequency limits are 20 Hz and 20 kHz respectively whereas for the cat they are 90 Hz and 60 kHz. Comparing the curves in an intermediate range of frequencies, say 90 Hz through 20 kHz, one notices that they are identical (with a fixed offset along the cochlear, x , axis). Since the presented data analysis is confined for acoustic signals with a frequency range of 100 Hz to 20 kHz, we arbitrarily use the parameters for humans in the filter bank design since it nicely accommodates this range of frequencies.

Although such a filter bank design is appealing because its spacing and resolution

is matched to that of the human cochlea, it is not intuitive since the variable x (cochlear distance) is not commonly used to describe auditory filter models. For example, it is not clear what the spatial resolution along the cochlea, Δx , is required to achieve a given spectral resolution, ΔX , of say 1/3 octave. In most instances it is convenient to think of spectral filtering and spectral bandwidths using an octave frequency convention where the frequency doubles with each octave. Thus, the cochlear distance variable needs to be related to the more conventional description of octave frequency. We define the octave frequency axis by

$$X = \log_2(f/f_r) \quad (1.2)$$

where f_r is a lower reference frequency, f is the frequency along the cochlear partition, and X is an octave (logarithmic base two) spaced frequency axis. Substituting the inverse of Eq. (1.2), $f = f_r 2^X$, into Eq. (1.1) and solving for X gives

$$X = \log_2[A/f_r] + \log_2[10^{\alpha} - k] \quad (1.3)$$

It is expected that for high frequencies the frequency variable, f , be precisely logarithmically spaced. For high frequencies we note that $10^{\alpha} \gg k$ is strictly satisfied and so allowing $k \approx 0$ results in

$$X = \log_2 \left[A/f_r \right] + x a \log_2 [10] \quad (1.4)$$

At high frequencies above about 1000 Hz, the octave frequency axis and cochlear partition distance are therefore linearly related (Fig. 1B). These variables can, therefore, easily be related for most of the hearing range using Eq. (1.4). Knowing this, the cochleotopic resolution, Δx , which is necessary to achieve a given or desired spectral resolution of ΔX (in octaves) is expressed as

$$\Delta x = \frac{\Delta X}{a \log_2 [10]} \quad (1.5)$$

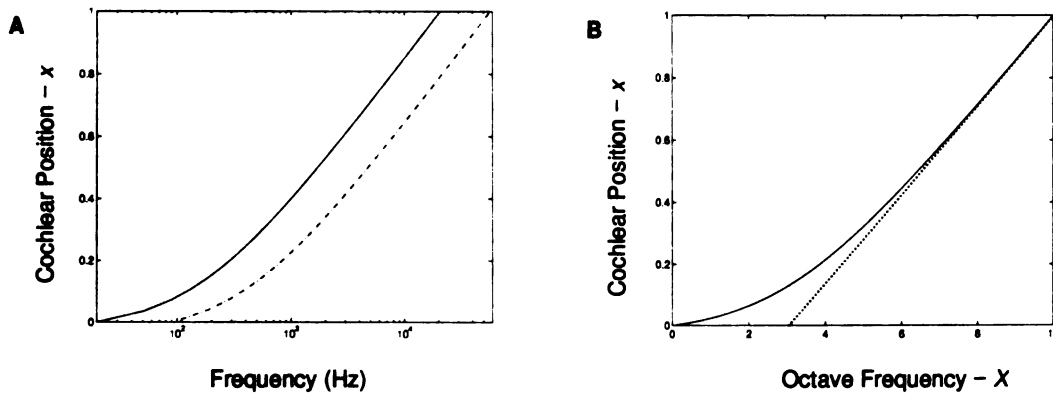


Figure 1: (A) Cochlea frequency versus cochlear position function for the cat (dotted) and human (continuous). Both curves have identical shape and differ only by a fixed offset which determines the minimum and maximum frequencies. (B) Human cochleotopic position function versus octave frequency function (continuous) and linear fit obtained using the octave frequency representation of Eq. (1.2). Note that the curves deviate at low frequencies where the cochleotopic curve flattens.

For the desired analysis, two filter banks are designed each with an equivalent spectral resolution of $\Delta X=1/4$ and $\Delta X=1/8$ octave. This corresponds to a cochleotopic resolution of $\Delta x=0.0358$ ($\Delta x=1.254$ mm) and $\Delta x=0.0179$ ($\Delta x=0.672$ mm) normalized units respectively. These two filter banks only differ in the spectro-temporal trade off which is a direct consequence of the uncertainty principle (Cohen 1995). The high resolution filter bank ($\Delta X=0.125$) can accommodate a periodic spectral oscillations or a ripple frequency of up to, Ω , of 4 cycles/octave yet has low temporal resolution. Alternately, the low resolution filter bank ($\Delta X=0.25$) accommodates a ripple frequency, Ω , of up to 2 cycles/octave but allows for slightly higher temporal resolution.

Although auditory filter bank models generally use filter bandwidths that adhere to the perceptually based filter bandwidth, i.e. the critical band ($1/3$ octave, $\Delta x=1.672$ mm) (Picone 1997), this convention is not used since the auditory system is actually sensitive to spectral frequencies beyond this range (up to ~ 8 ripples per octave; Van Veen and Houtgast 1985; Supin *et al.* 1999; Chin *et al.* 1999). Analysis of vowel sounds, for example, has shown that vowels can have spectral modulations (denoted by the ripple density, Ω) of up to 4 cycles per octave. Yet it appears that humans are most sensitive to spectral modulations of up to 2.5 cycles per octave (Van Veen and Houtgast 1985) which would require a filter bank resolution of about $1/5$ octave to appropriately sample these signals (as suggested by the Nyquist sampling theorem).

Secondly, most of the filter banks used for auditory analysis are constructed for simulating auditory neuronal responses and for understanding the perceptual and psychophysical limits of the auditory system. Our purpose here is not necessarily to simulate the auditory system, but instead to thoroughly characterize the spectro-temporal content of natural sounds. Hence we seek to find out "what stimulus content the auditory system is being exposed to?" How and if the auditory system makes use of this information is a separate problem which must be solved as well.

We use Eq. (1.1) (Greenwood 1990) to design a linear filter bank with L independent sub-bands. The cochleotopic axis is first discretized to a resolution of Δx (equivalent resolution of ΔX). The center frequency (on the logarithmic cochlear axis x), f_l , of the l^{th} filter component are expressed as

$$f_l = A(10^{ax_l} - k) \quad (1.6)$$

where $x_l = x_0 + l\Delta x$ and $l = 0 \dots L-1$. For each filter the 3 dB cutoff frequencies, denoted by f_l^* and f_{l+1}^* for the l^{th} filter, are

$$f_l^* = A(10^{ax_l} - k) \quad (1.7)$$

where $x_l^* = x_0 + (l-0.5)\Delta x$ and $l = 0 \dots L$. The linear bandwidth of the l^{th} filter is

therefore $\Delta f_l = f_{l+1}^* - f_l^* = A(10^{ax_{l+1}^*} - 10^{ax_l^*})$. A two octave segment of the auditory filter

bank is shown in Fig. 2 A. On a linear frequency axis the filter bandwidths, Δf_i , are dilated (much like for a wavelet filter bank) with increasing frequency. Unlike a wavelet decomposition, however, the amplitude of the filters do not scale and have a constant gain. When displayed on the cochleotopic axis, x , the filters are effectively identical having the same resolution, Δx .

1.4 Filter Selection and Design

Selection of filters for auditory models is generally based on choosing a filter prototype function which captures the physiologic properties of the cochlea and eighth nerve auditory responses. Filters which model the steep high frequency rolloffs and smooth low frequency transitions of eighth nerve auditory filter, such as the gamma-tone filter (Lyon 1982), are often used. More refined filter models have been designed using non-parametric methods which estimate the auditory nerve filter transfer function by fitting experimental data (Jenison *et al.* 1991). Here, filter criteria and design considerations for decomposing natural sounds into a spectro-temporal representation are outlined. Although the chosen filter prototypes (B -spline) are not physiologically motivated, they nonetheless offer several advantages over using physiologically derived filter shapes and more conventional precision filters (i.e. Kaiser, Dolph-Chebyshev, etc.). In particular, these filters are chosen since they are spectro-temporally compact. They provide superior stopband and passband attenuation properties over all other filter types, thereby preventing signal leakage from adjacent bands.

The chosen B -spline lowpass filter (Roark and Escabí 1998) has an impulse response

$$h[n] = \frac{\omega_c}{\pi} \frac{\sin(n\omega_c)}{n\omega_c} \left(\frac{\sin(\alpha n\omega_c/\rho)}{\alpha n\omega_c/\rho} \right)^\rho \quad (1.8)$$

where $n = -N, \dots, N-1, N$, $\omega_c = f_c/F_s$ is the discrete-time filter cutoff frequency (units of radians), f_c is the desired filter cutoff frequency (in Hz), F_s is the sampling rate (in Hz), $2N+1$ is the filter order (the number of coefficients), and α and ρ are filter parameters which control the filter transition width and the stopband and passband attenuations. The frequency domain lowpass filter prototype function (i.e. for $N \rightarrow \infty$) is given by

$$H(\omega) = 1 - \frac{1}{\rho!} \sum_{k=0}^{\rho} (-1)^k \binom{\rho}{k} \left[\frac{\rho}{2} \left(\frac{|\omega| - \omega_c}{\alpha \omega_c} + 1 \right) - k \right]_+^\rho \quad (1.9)$$

where $[x]_+ = \max(0, x)$. This filter can be thought of conceptually, as a spectral convolution between the ideal lowpass filter transfer function and a ρ^{th} order B -spline window of width $\alpha\omega_c/\rho$.

The B -spline filter design has several advantages over other commonly used precision filters such as the Kaiser, Saramakii, and Dolph-Chebyshev. First, the filter

prototype is effectively temporally (temporal convergence factor of $1/N^{p+1}$) and spectrally compact. This property requires that filter transfer function, $H(\omega)$, and its corresponding impulse response function, $h[n]$, be zero outside some range of values (for example $|n| > N$ and for $|\omega| > \omega^*$). Secondly, unlike most precision filters which generally have a constant attenuation throughout the passband and stopband, the B -spline filter transfer function has an exponentially decreasing stopband error (ATT) at frequencies away from the filter cutoff frequencies. An examples of the B -spline filter and a Kaiser window with similar design criteria are shown in Fig. 2 B. Note that the stopband error for the B -spline filter decreases at frequencies away from the filter cutoff frequencies. Signals which pass through these filters are therefore effectively bandlimited since these filters can achieve much higher attenuation than other filters with identical design specification. This is particularly important to prevent strong signals at adjacent frequency bands from leaking into bands which have very little energy. Such an artifact, for example, would show up as correlated activity across frequency bands despite the fact that these bands may not have any common signal. This would be a significant problem if one where to estimate spectro-temporal content using physiologically derived filters (since these filters generally have shallow rolloffs).

To construct the filter bank with desired cutoff frequencies as described in section 1.3, the lowpass filter impulse response of Eq. (1.8) is adapted so that it adheres to the bandpass filter specifications (Oppenheim and Schaffer 1989) in the filter bank design. The impulse response for the l^{th} bandpass filter is given by

$$b_i[n]=h_{i+1}[n]-h_i[n] \quad (1.10)$$

where $h_{i+1}[n]$ and $h_i[n]$ are the impulse response of the component lowpass filters with cutoff frequencies $\omega_{i+1}=2\pi f_{i+1}^*/F_s$ and $\omega_i=2\pi f_i^*/F_s$. The transition width of the i^{th} filters is chosen as $TW_i=(f_{i+1}^*-f_i^*)/4$. For all filters the minimum stopband and passband attenuation is set to 60 dB so that signal leakage is prevented. An example of such a bandpass filter is provided in Fig. 2 B. Equations for choosing the parameters α and ρ are provided by Roark and Escabí (1998).

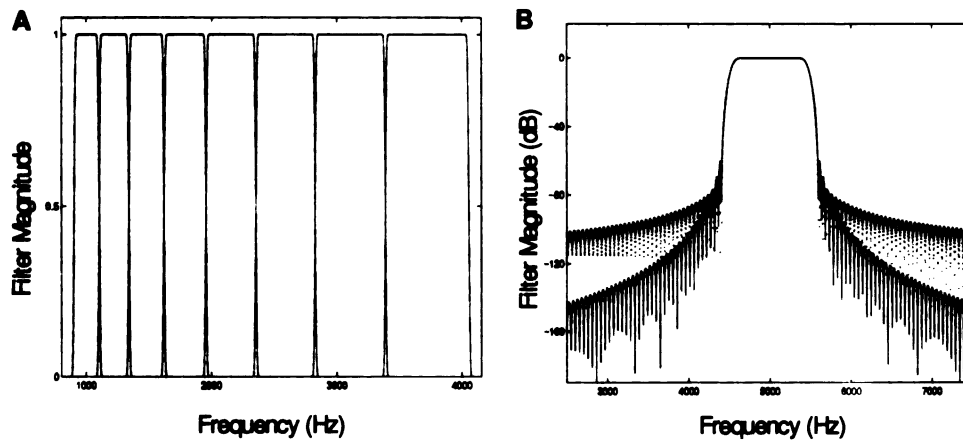


Figure 2: Cochleotopic filter bank used for spectro-temporal sound decomposition (A) shown for a two octave segment. The filter bandwidths grow with increasing center frequency. (B) Example B -spline filter (continuous line) used for spectro-temporal decomposition compared to a Kaiser filter (dotted line) using identical filter order ($N=848$). The Kaiser filter has a higher first and second sidelobe attenuation (72 dB) than the B -spline filter (60 dB). The B -spline filter, however, has a decreasing stopband attenuation at frequencies away from the filter cutoff frequency whereas the

Kaiser filter stopband attenuation levels off at 100 dB. This property helps reject signal leakage from adjacent filter bands.

1.5 Signal Decomposition and Envelope Extraction

Using the filter bank design of sections 1.3 and 1.4, acoustic signals, $x[n]$, where decomposed by filtering each sounds with the filter impulse response $b_l[n]$ using the discrete time convolution operator

$$y_l[n] = \sum_{k=-N}^N b_l[n-k]x[k] \quad (1.11)$$

For each input signal, $x[n]$, a series of L outputs, $y_1[n], y_2[n], \dots, y_L[n]$, is therefore produced. Since the acoustic signals, $x[n]$, were often extremely long (tens of minutes and therefore tens of mega samples) this operation was performed using the overlap save method which partitions the signals into blocks of a small fixed size upon performing this operation (Jackson 1989). This guarantees that one does not exceed the memory requirements of the computer. This method does not introduce any error to this operation.

For each band, the envelope was extracted using the Hilbert transform operator. To do this we approximated the analytic signal representation using (Cohen 1995; Oppenheim and Schaffer 1989)

$$z_l[n] = b_l[n] + H[b_l[n]] = a_l[n] e^{(j\phi_l[n])} \quad (1.12)$$

where $z_l[n]$ is the discrete time analytic signal, $H[x[n]] = \sum_{m=-\infty}^{\infty} h[n-m]x[m]$ is the 90 degree phase shifter operation, otherwise known as the discrete time Hilbert transformer (Cohen 1995; Oppenheim and Schaffer 1989), and

$$h[n] = \begin{cases} \frac{2 \sin^2(\pi n/2)}{\pi n}, & n \neq 0 \\ 0, & n = 0 \end{cases} \quad (1.13)$$

is the corresponding impulse response for the 90 degree phase shifter. Using this formulation, the envelope for the l^{th} band is given by $a_l[n] = |z_l[n]|$. Since the filter bandwidth, Δf_l , of the l^{th} output is dilated with increasing l , the bandwidth for each corresponding envelope, $a_l[n]$, is also dilated and can be approximated by Δf_l (Oppenheim and Schaffer 1989; Cohen 1994). As a consequence, the l^{th} temporal envelope has a maximum modulation rate of $\Delta f_l/2$. It is desired that the maximum modulation rate for each band be uniform so that their temporal properties can be compared across bands. This is achieved by lowpass filtering each band using a B -spline lowpass filter, $h[n]$, with cutoff frequency of 100 Hz (parameter for the filter are:

$$\alpha = 0.1037, \quad \rho = 2.3391, \quad N = 4973, \quad ATT = 60, \quad TW = 20)$$

$$e_l[n] = \sum_{k=-N}^N h[n-k] a_l[k] . \quad (1.14)$$

The lowpass filtered temporal envelope, $e_l[n]$, therefore accommodates the same range of temporal modulations for all bands.

Since acoustic features and sound perception are generally well described using a decibel intensity description, we also consider the zero mean decibel spectro-temporal envelope

$$e_l^{dB}[n] = 20 \log_{10}(e_l[n]) - \mu_{dB} . \quad (1.15)$$

where μ_{dB} is the mean value of $20 \log_{10}(e_l[n])$ and the expectation is taken across all time, n , and along the spectral axis, l . The mean normalization is performed to facilitate comparison and analytic assessment across frequency channels and across stimulus ensembles.

1.6 Spectrographic Envelope

To characterize linear spaced features of natural sounds the short-time Fourier transform signal representation is used (Cohen 1995; Oppenheim and Schaffer 1989). The discrete time version of this transform is given by

$$X[n, \omega_k] = \sum_{m=-N}^N x[n+m] w[m] e^{-j\omega_k m} \quad (1.16)$$

where as before n and ω_k are the discrete time and discrete frequency variables,

$w[n]$ is a time limited window sequence, and $x[n]$ is the discrete time sampled signal. As for the filter bank design of sections 1.3 and 1.4, the corresponding B -spline window function is used

$$w[n] = \left(\frac{\sin(\pi \alpha n / \rho)}{\pi \alpha n / \rho} \right)^\rho \quad (1.17)$$

where $n = -N, \dots, N$, N is the window order (number of coefficients), α is a parameter which controls the window bandwidth (Δf), and ρ control the window attenuation (ATT). This window is chosen for the same set of reasons and design considerations as for section 1.4. The spectrogram is obtained by evaluating the magnitude

$$S[n, \omega_k] = \sqrt{X[n, \omega_k] \cdot X[n, \omega_k]^*} \quad (1.18)$$

of the short-time Fourier transform. Here $X[n, \omega_k]^*$ is the complex conjugate of the short-time Fourier transform.

The stimulus spectro-temporal envelope, $\bar{S}[n, \omega_k]$, obtained by dividing the spectrogram by a detrending function $S[\omega_k]$

$$\bar{S}[n, \omega_k] = \frac{S[n, \omega_k]}{S[\omega_k]} = 1 + \frac{\Delta S[n, \omega_k]}{S[\omega_k]} \quad (1.19)$$

The quantity $\Delta S[n, \omega_k] = S[n, \omega_k] - S[\omega_k]$ is the difference spectrogram about the detrending function. The detrending function is obtained by applying a linear fit (in mean square sense) of general form $A\omega_k + B$ to the stimulus mean ensemble decibel power spectrum, $20 \log_{10}(E[S[n, \omega_k]])$ (expectation taken with respect to n). The detrending function is therefore expressed as

$$S[\omega_k] = 10^{(A\omega_k + B)/20} \quad (1.20)$$

Note that after combining terms from Eqs. 1.19 and 1.20 the overall decibel spectro-temporal envelope is conveniently expressed as

$$\bar{S}_{dB}[n, \omega_k] = 20 \log_{10}(S[n, \omega_k]) - A\omega_k - B \quad (1.21)$$

Although linear trends are subtracted from the stimulus decibel spectrogram using this procedure, the detrended stimulus is not white. Note that in general strong spectral oscillations are still present. Examples are shown in Figs. 8 and 12-14.

The outlined detrending procedure is applied for several reasons. First note that *natural* sounds generally have very little energy at high frequencies. On a logarithmic

(decibel) plot the power spectrum is usually strongly biased at low frequencies despite the fact that relevant stimulus components are also present at high frequencies. This procedure therefore removes spectral trends which are characteristic of natural sounds. Note that the auditory system effectively performs a similar detrending operation, since frequency tuning and integration bandwidths in the sensory epithelium of the cochlea are logarithmically spaced (e.g. Kiang *et al.* 1965; Evans 1972; Liberman 1982; Greenwood 1990). Because of this, similar detrending procedures are often employed for speech modeling and in speech recognition systems (Picone 1997). Secondly, this transformation is crucial for quantifying contrast statistics of natural sounds in sections 1.9–1.10. Unlike the spectrogram which depicts absolute energy variations of the stimulus, the defined spectro–temporal envelope depicts relative energy variations along time and frequency. This is not an unreasonable descriptor since it is arguable that relative quantities are far more important for the auditory processing than absolute quantities (for example Weber’s law). Note that similar reasoning is also applied to visual processing since visual contrast is likewise defined as a relative quantity ($C = (I_{Max} - I_{Min}) / (I_{Max} + I_{Min})$).

As for the spectro–temporal envelope of Eq. (1.15), we also consider the zero mean logarithmic amplitude spectro–temporal envelope

$$\bar{S}_{dB}[n, \omega_k] = 20 \log_{10}(\bar{S}[n, \omega_k]) - \mu_{dB} \quad (1.22)$$

where μ_{dB} is the mean of $20 \log_{10}(\bar{S}[n, \omega_k])$. This descriptor is used since the *perception* of intensity differences is ordered on a logarithmic space (Miller 1947; Harris

1963; Viemeister and Bacon 1988) and since temporal fluctuations of natural sounds are likewise logarithmically distributed (Attias and Schreiner 1998a).

1.7 Low-Order Stimulus Statistics – The Power Spectrum

For all soundscapes we estimated the stimulus power spectrum using a Welsch average periodogram (Hayes 1996). Prior to estimating the periodogram, each sound sequence was normalized as $s[n]/\sigma_s$. Here σ_s is the stimulus standard deviation. This normalization is performed so that all sounds have unity standard deviation therefore allowing for ease of comparison. The spectral resolution, Δf , was set to 86 Hz. The ensemble power spectrum was then estimated by averaging over all sounds in the ensemble using the equation

$$P_{ens}[\omega_k] = \frac{1}{N} \sum_{n=1}^N P_n[\omega_k] \quad (1.23)$$

Here, $P_n[\omega_k]$ is the Welsch average periodogram for a particular sound in the ensemble. Upon computing the ensemble periodogram, a least-squares linear fit of the form $S[\omega_k] = A\omega_k + B$ was applied to each ensemble in order to obtain descriptive parameters.

Fig. 3 shows results obtained for five sound ensembles (human conversational speech (A), environmental background sounds (B), animal vocalizations (C), pop music (D), and classical music(E)). In all instances, the power spectrum had a decreasing trend

as a function of increasing frequency. The constants A and B are given in Table 1.

	A (dB/kHz)	B (dB)
Human Speech	-2.68	7.1
Animal Vocalizations	-1.41	8.9
Background Sounds	-1.71	7.2
Pop Music	-2.29	7.9
Classical Music	-2.87	5.5

Table 1: Power spectrum statistics for five natural sound ensembles. Mean slope, A , and y -intercepts, B . All sounds had negative slopes and positive intercepts.

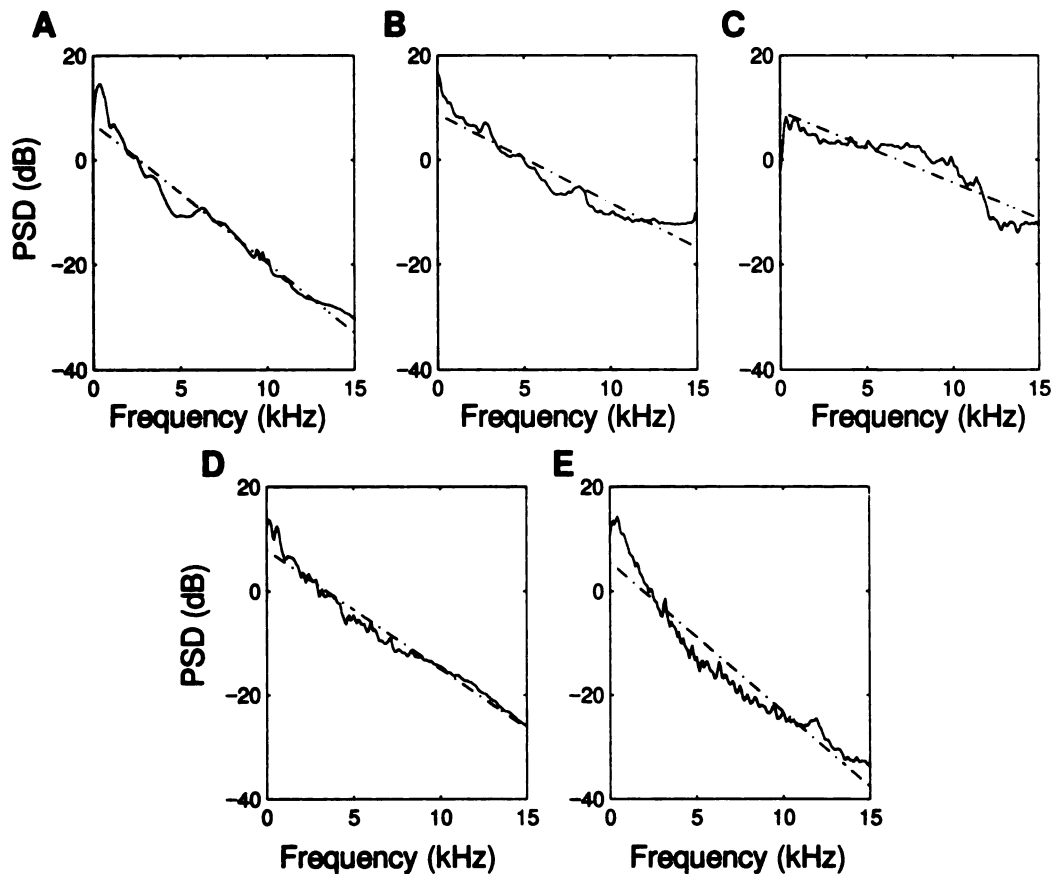


Figure 3: Power spectrum (continuous line) for five natural sound ensembles show a decreasing trend in energy as a function of frequency with similar intercepts and slopes

(see Table 1). The least-squares regression fit is shown as dotted line for all ensembles. Shown for: human speech (A), environmental background sounds (B) (e.g. wind, running water, etc.), animal vocalizations (C) (both primate and non-primate sources), pop music (D), and classical music (E).

1.8 Across Band Correlations of Natural Sounds

Throughout the remainder of this chapter we consider a generic spectro-temporal envelope variable, $s_k[n]$. For any of the described spectro-temporal measure (the linear and logarithmic spectro-temporal envelopes as well as the logarithmic and linear filter bank envelopes described in sections 1.5 and 1.6 respectively) can be substituted for $s_k[n]$. Specifics as to which envelope is used are noted in the figure legends and throughout the text.

For all natural sounds the crossband correlation was estimated using the correlation coefficient. For the k^{th} and l^{th} envelope outputs, $s_k[n]$ and $s_l[n]$ respectively (or the k^{th} and l^{th} spectrogram channels), the correlation coefficient is computed as

$$\rho_{kl}^2 = \frac{1}{\sigma_l^2 \sigma_k^2} E \left[\bar{s}_l[n]^2 \bar{s}_k[n]^2 \right] = \frac{1}{\sigma_l^2 \sigma_k^2} \frac{1}{M} \sum_{n=1}^M \bar{s}_l[n]^2 \bar{s}_k[n]^2 \quad (1.24)$$

where the time average expectation, $E[\cdot]$, is taken with respect to n ,

$\bar{s}_l[n] = s_l[n] - \mu_l$ is the zero mean spectro-temporal envelope, μ_l is the mean value of

the l^{th} channel envelope taken across all time, and σ_k and σ_l are the corresponding standard deviations for the k^{th} and l^{th} channels respectively. This measure quantifies the amount of redundancy or similarity that exists across frequency channels. The procedure for computing the across-channel correlation matrix is depicted in Fig. 4. The temporal envelope for each channel is first extracted, at which point a channel by channel comparison is performed using the correlation coefficient. Regimes in the correlation map with high correlation coefficient values designate channel combinations that show highly correlated temporal modulations.

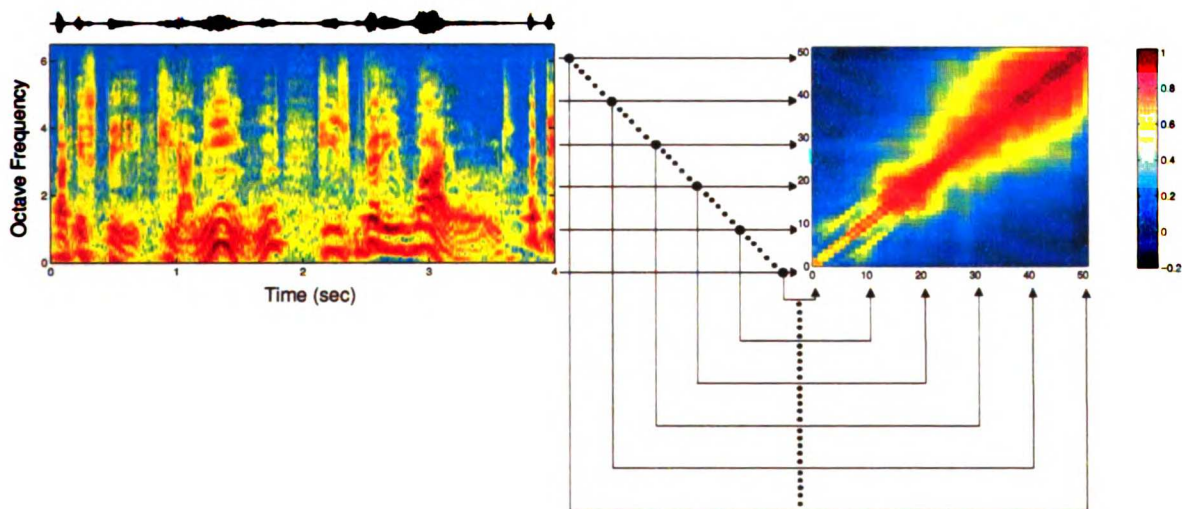


Figure 4: Computing the across-channel correlation matrix, ρ_{kl} . The stimulus waveform (top left) is decomposed into a spectro-temporal stimulus representation (bottom left). The spectro-temporal envelope (shown for human speech) is then used to construct the across-channel correlation matrix. For each frequency-channel (total of 52 channels), the temporal envelope is extracted and compared with all of the other temporal envelopes. This comparison consists of computing the correlation coefficient between any two channels. This measure provides an unbiased estimate of the degree of

corresponding channels are highly correlated whereas values near zero (blue) indicate that the temporal envelopes for the compared channels are highly dissimilar.

Correlation coefficient matrices, ρ_{kl} , were computed for all sounds in the chosen ensembles using both the decibel and linear amplitude spectro-temporal envelope. Likewise, across-channel correlation matrices were also computed for the spectrographic and the octave filter bank spectro-temporal decompositions. In all instances the results were qualitatively similar. Results are therefore presented only for the octave filter bank design decibel spectro-temporal envelope.

Fig. 5-7 depicts typical across-correlation matrices, ρ_{kl} , for the different sound ensembles. Clear and distinct trends were observed across the different ensembles. Of all the sound categories, vocalization sounds had the most diverse range of correlation matrices, ρ_{kl} . Across-channel correlation matrices for vocalizations have highly structured oscillatory patterns, indicative of complex patterns of correlation across distinct frequency channels (Fig. 5 D and F). Likewise, speech sounds also had a complex pattern of spectrally correlated channels although, unlike the vocalization sounds, overall pattern of the correlation matrix was homogenous for the different sound segments used (compare the correlation matrices for Fig. 5 B and Fig. 4). By comparison, white noise has no across-channel correlations (Fig. 7 F).

By far the weakest correlations were observed for environmental background sounds. Examples are shown for running water (Fig. 6 A and B), wind (Fig. 6 C and D), and shuffling leaves (Fig. 6 E and F). These sounds generally showed very weak correlation patterns. For the shuffling leaves example (Fig. 6 F) a high degree of

correlation was observed among the high frequency channels. This finding is evident in its spectro-temporal envelope (Fig. 6 E) which shows a series of sharp broad-band features that are most prominent at high frequencies.

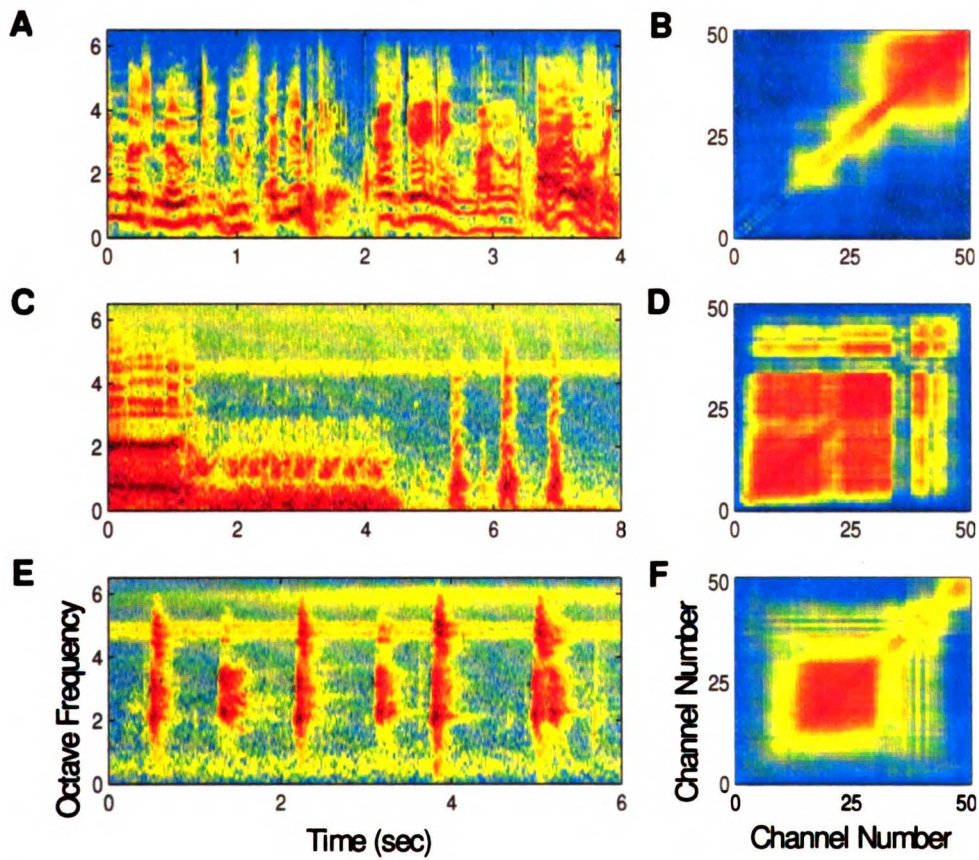


Figure 5: Spectro-temporal correlations for speech and primate vocalizations. Spectro-temporal envelope segment (A) and across-channel correlation matrix of human speech segment (B). Across-channel correlation matrix for primate vocalizations (D) and (F) and a short segment of the corresponding spectro-temporal envelopes (C) and (E) respectively. Both speech and animal vocalizations showed significant and highly structured patterns of across-channel correlations, indicative of complex interactions across spectral channels.

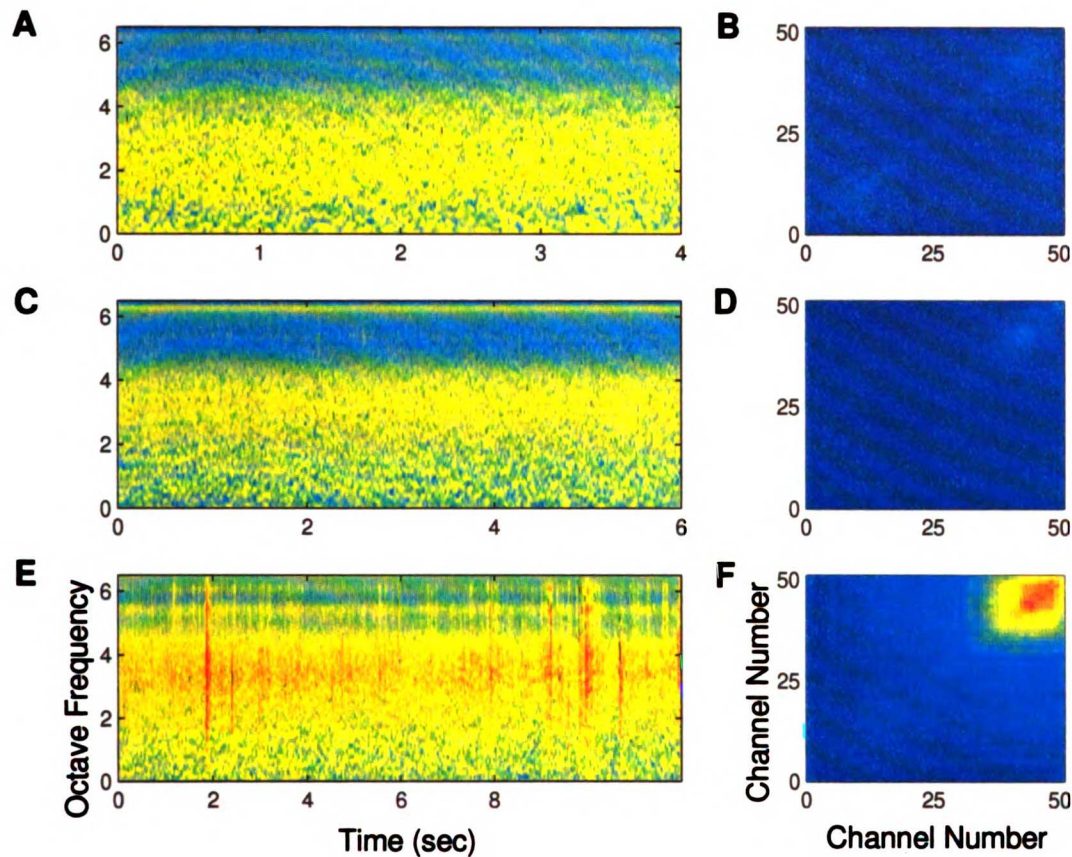


Figure 6: Spectro-temporal correlations for environmental background sounds.

Spectro-temporal envelope segments for the sounds emanating from a moving stream (A), wind (C), and shuffling leaves (D) and the corresponding across-channel correlation matrices (B, D, and F respectively). Both the moving stream and the wind have very little across-channel correlations. The shuffling leaf sounds have significantly higher correlations. This was most obvious at high frequencies where transient broadband click-like sounds create comodulated temporal components.

As for speech and vocalizations, both pop and classical music showed a high degree of correlation across frequency channels (Figs. 7 B and D), although the observed patterns for the different sound segments did not show pronounced differences. As for speech, correlations were strongest at high frequencies. This is evident from the spectro-

temporal envelopes (Fig. 7 A and C) which show comodulated components at high frequencies.

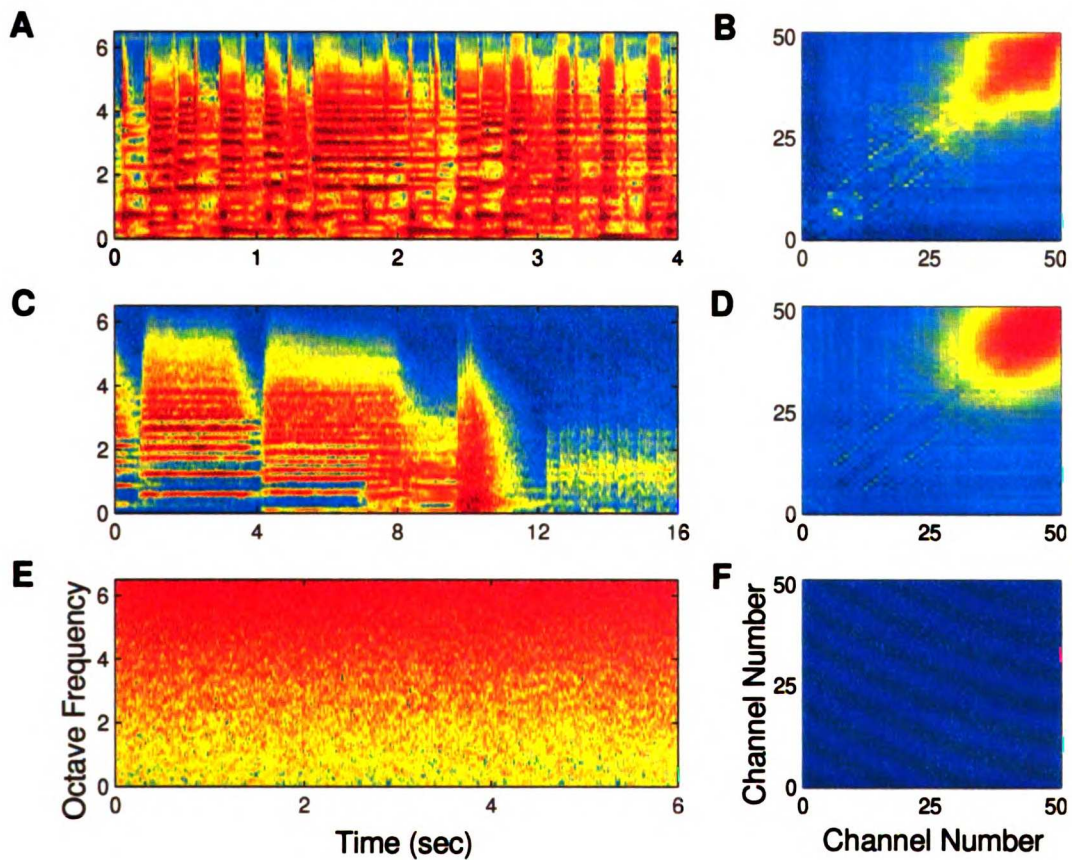


Figure 7: Spectro-temporal correlations for music and white noise. Spectro-temporal envelope segments for classical music (A) and pop music (C) show a high degree of structure. By comparison white noise (E) has little spectro-temporal structure. The corresponding across-channel correlation matrices for classical music (B) and pop music (D) show significant wide scale correlations. These are most prominent at high frequencies likely because of broadband temporally comodulated components. The across-channel correlation matrix for white noise (F) shows no correlation across frequency channels.

1.9 Spectro–Temporal Contrast of Natural Sounds

Visual contrast is defined as the percent deviation relative to the mean intensity of a spatial sinusoid grating. Mathematically it is expressed as $C=(I_{\max}-I_{\min}) / (I_{\max}+I_{\min})$ where I_{\max} and I_{\min} correspond to the maximum and minimum stimulus intensities (Albrecht 1995; Nordmann, Freeman, and Casanova 1992; Troy *et al.* 1998). In the auditory literature the analogous quantity is the modulation depth or modulation index, $\beta=(I_{\max}-I_{\min}) / I_{\max}$. Such a description suffices for the case of sinusoidal, square wave, and other simple stimulus gradations since these waveforms are fully specified by their minimum and maximum intensities. For natural signals, where the amplitude gradations can cover several orders of magnitude, such descriptions fail to fully characterize amplitude fluctuations since they only take into account the minimum and maximum envelope intensities. They do not tell us anything about intermediate values and higher–order amplitude statistics of the modulation signal. To overcome this we adopt a more general definition of contrast to denote the probability distribution of the relative amplitude gradations.

A large ensemble of natural sounds was analyzed which included human speech (Excerpts from Hamlet), music (pop and classical), environmental sounds (wind, rain, thunder, etc.), animal vocalization (primate, bird, cat, crickets etc.) and mixtures of the latter two. These sounds were taken as representative examples of the vast acoustic biotope (Smolders *et al.* 1979) which mammals and humans are typically exposed to. For comparison, white noise was included in this analysis as a control. For all sounds the relative spectro–temporal envelopes of Eq. (1.19) and (1.21), $\bar{S}(t, f_k)$ and $S_{dB}(t, f_k)$,

were computed and the corresponding envelope contrast distributions, $C = p(\bar{S})$ and

$C_{dB} = p(\bar{S}_{dB})$, were estimated for thousands of sound segments.

Fig. 8 shows the decibel and linear amplitude spectro-temporal envelopes for a human speech segment. The linear amplitude spectro-temporal envelope (Fig. 8 A) shows little detail and largely consists of amplitude values near zero (blue). The measured linear modulation depth for this speech segment is exceptionally high (99.9994%), whereas the measure standard deviation, σ , is relatively small (0.019 normalized amplitude units for an amplitude range that spans 0 to 1). Together these two descriptors provide a conflicting and misleading description of the envelope fluctuations. The large modulation index suggests that the sound components for this segment span a large range of the 0 to 1 linear amplitude dimension, whereas the small standard deviation suggests that it only covers a small portion of this linear amplitude space. By comparison, the decibel amplitude spectro-temporal envelope (Fig. 8 B) shows significant more structure. A close inspection of the logarithmic decibel envelope,

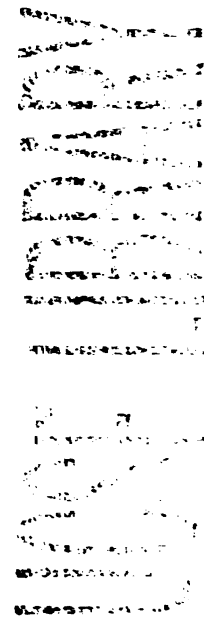
$\bar{S}_{dB}(t, f_k) = 20 \log_{10}(\bar{S}(t, f_k))$, reveals that the speech signal has spectral and temporal amplitude fluctuations that span several orders of magnitude (roughly 50 dB, Fig. 8 B).

To quantify these observations, we computed the linear and decibel contrast distributions for all sounds by collapsing all pixels values of the linear and decibel spectro-temporal envelopes respectively into a probability histogram. These are shown collectively for all sound ensembles in Figs. 9 and 10. The linear amplitude distribution was obtained by normalizing the spectro-temporal envelope so that it has a maximum value of unity, $\bar{S}_{Lin}(t, f) = \bar{S}(t, f) / \max(\bar{S}(t, f))$, therefore obeying the general

convention used to define a modulation signal (Cohen 1995). For all natural sounds the linearly defined envelope has a skewed amplitude distribution such that loud (near unity) sound segments are sparse whereas soft segments (near zero) are much more common (Fig. 9). In contrast, white noise (Fig. 9 F) has a linear amplitude distribution which is broadly distributed and partially symmetric. Upon performing a logarithmic decibel transformation of the envelope to construct the decibel contrast distributions,

$C_{dB} = p(\bar{S}_{dB})$, the relative amplitude gradations of natural sounds are roughly symmetric, have an average standard deviation of 10.9 dB, and span an overall range of more than 25 dB (Fig. 10) for the natural sounds ensembles. Traditional definitions of contrast, such as the modulation depth or the envelope standard deviation, fail to characterize such higher-order statistics associated with the shape and the overall range of the envelope gradations.

The transformed logarithmic decibel amplitude (\bar{S}_{dB}) magnifies the soft and moderately loud sound segments relative to the very loud sounds. Thus one can discern the fine detail in the amplitude distribution over several orders of magnitude. This descriptor is perceptually motivated since the perception of loudness and intensity discrimination thresholds are ordered on a decibel space (Miller 1947; Stevens 1957; Harris 1963; Stevens 1972; Jesteadt, Wier, and Green 1977; Viemeister and Bacon 1988). For all sounds the distribution of logarithmic-contrast is broadly distributed. To quantify the range of relative amplitudes we measured the average spread of the distribution, σ_{dB} . With the exception of the background sounds, all natural sounds had relatively large standard deviation values: 11.0 dB for speech, 13.3 dB for vocalizations, 7.4 dB for background sounds, 11.2 dB for pop-music and 11.8 dB for the classical



music ensemble. By comparison, the white noise control ensemble has a small standard deviation of only 5.6 dB.

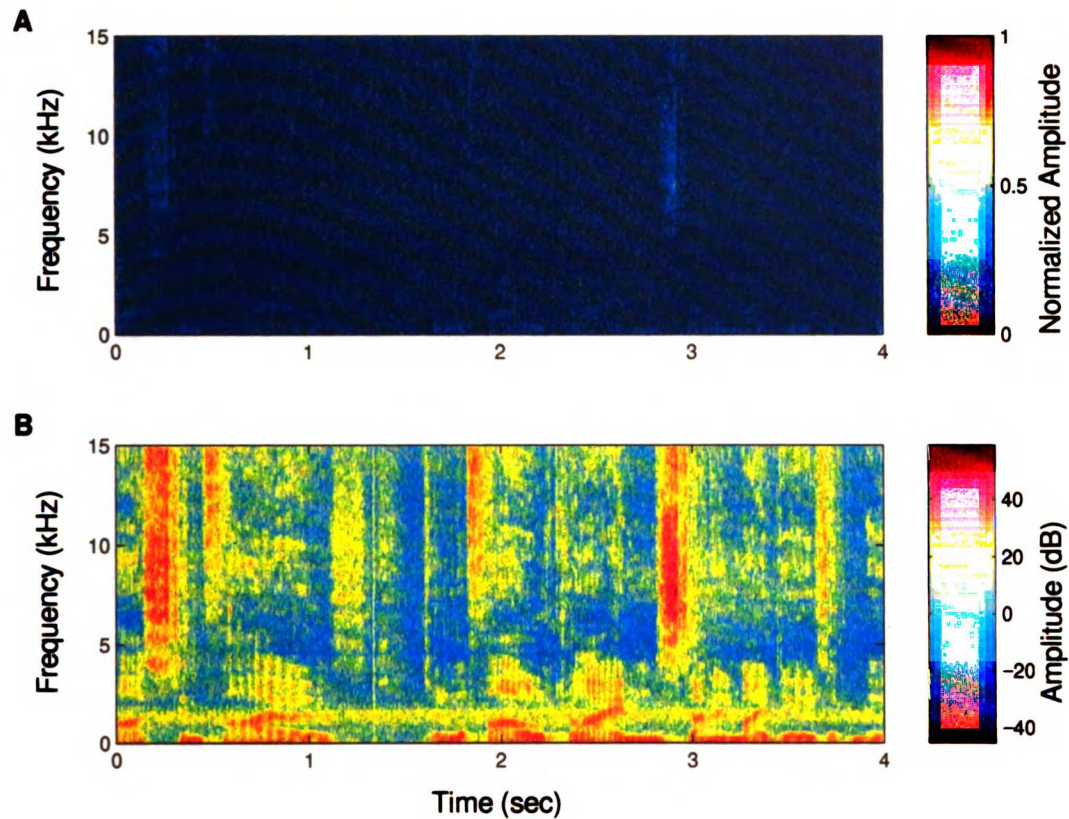


Figure 8: Detrended spectrographic envelope for a short speech segment. Shown using a

linear amplitude, $\bar{S}_{Lin}(t, f_k)$, and a decibel amplitude convention

$$\bar{S}_{dB}(t, f_k) = 20 \log_{10}(\bar{S}(t, f_k)) - \mu_{dB}$$

. The linear amplitude spectro-temporal

envelope, shows little detail and most of the signal values are concentrated near zero.

The decibel spectro-temporal envelope has significant more detail and has amplitude fluctuations which span a large dynamic range of more than 50 dB.

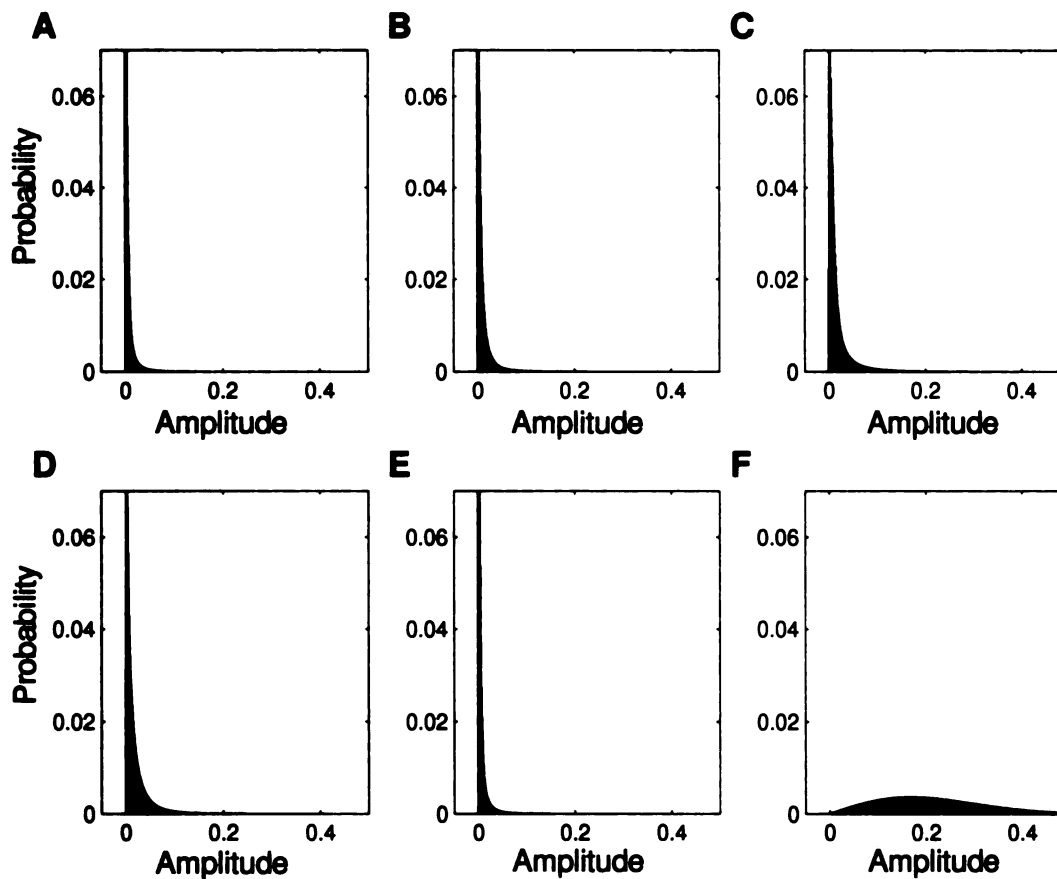


Figure 9: Linear contrast statistics for natural sound ensembles. The linear amplitude distribution, $p(\bar{S})$, for speech (A), animal vocalizations (B) (both primate and non-primate sources), background sounds (C) (e.g. wind, running water, etc.), pop-music (D) classical music (E) and white noise (F). All sounds are normalized so that they have a maximum amplitude of unity. Natural sound ensembles have a highly skewed exponential-like linear amplitude distributions. The spectro-temporal envelope of natural sounds has a significantly larger proportion of soft to loud sound components. By comparison, white noise has a broad distribution (F).

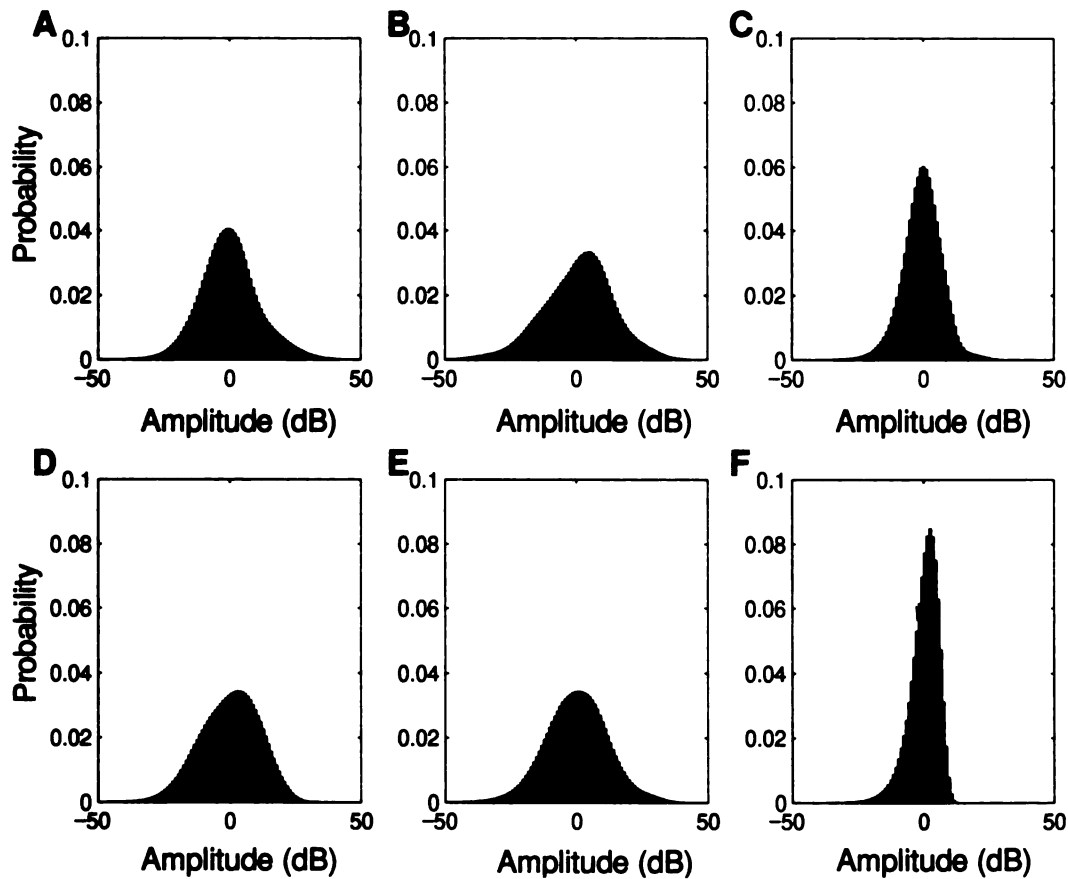


Figure 10: Decibel contrast statistics for natural sound ensembles. The decibel amplitude distribution, $p(\bar{S}_{dB})$, for speech (A), animal vocalizations (B) (both primate and non-primate sources), background sounds (C) (e.g. wind, running water, etc.), pop-music (D) classical music (E) and white noise (F). All natural sound ensembles have normal-like decibel distributions. Of these, environmental sounds has the narrowest distribution indicating that the overall range spectro-temporal fluctuations are significantly smaller than for speech, vocalizations, and music. By comparison, white noise has the narrowest distribution indicative of a narrow range of spectro-temporal amplitude fluctuations (F).

The statistical homogeneity of the *shape* of contrast distribution across the four

natural sound ensembles suggests that logarithmic amplitude fluctuations are an invariant acoustic property across natural stimuli (Attias and Schreiner 1998a). Natural sounds are therefore characterized exponential-like amplitude distributions and normal-like logarithmic contrast which extends over a dynamic range of 14–25 dB (i.e. $2\sigma_{dB}$). This fundamental property of natural sounds closely resemble natural image statistics which show similar spatial amplitude fluctuations (Ruderman and Bialek 1994; Dong and Atick 1995; Ruderman 1997).

1.10 Contrast and Intensity Dynamics of Natural Sounds

Although such a description gives us insight into the *global* amplitude statistics of sensory signals, it nonetheless presents us with a static picture of the acoustic world which has been averaged for a large ensemble over all time. In reality, natural signals such as speech are time-varying and non-stationary. It therefore makes sense to consider the dynamic behavior of these signals at time-scales which are relevant for neuronal and perceptual integration . A realistic model of contrast therefore takes into account time dependencies that arise from multiple sound sources which radiate in and out of the acoustic scene.

To characterize such time dependencies we defined a time-dependent contrast distribution, $C_{dB}(t) = p(\bar{S}_{dB}|t)$. This statistic was computed by discretizing the time axis of the spectrographic signal representation into 47 msec frames (Fig. 11) and computing the contrast distribution for each frame. A frame size of 47 milliseconds is chosen since intensity perception has a maximum integration time-scale which is slightly

larger (in the order of 200–500 msec) (Hughes 1946; Garner and Miller 1947). Thus by choosing this time-scale we can sample and track the dynamic behavior of intensity fluctuations within a perceptually relevant time-scale.

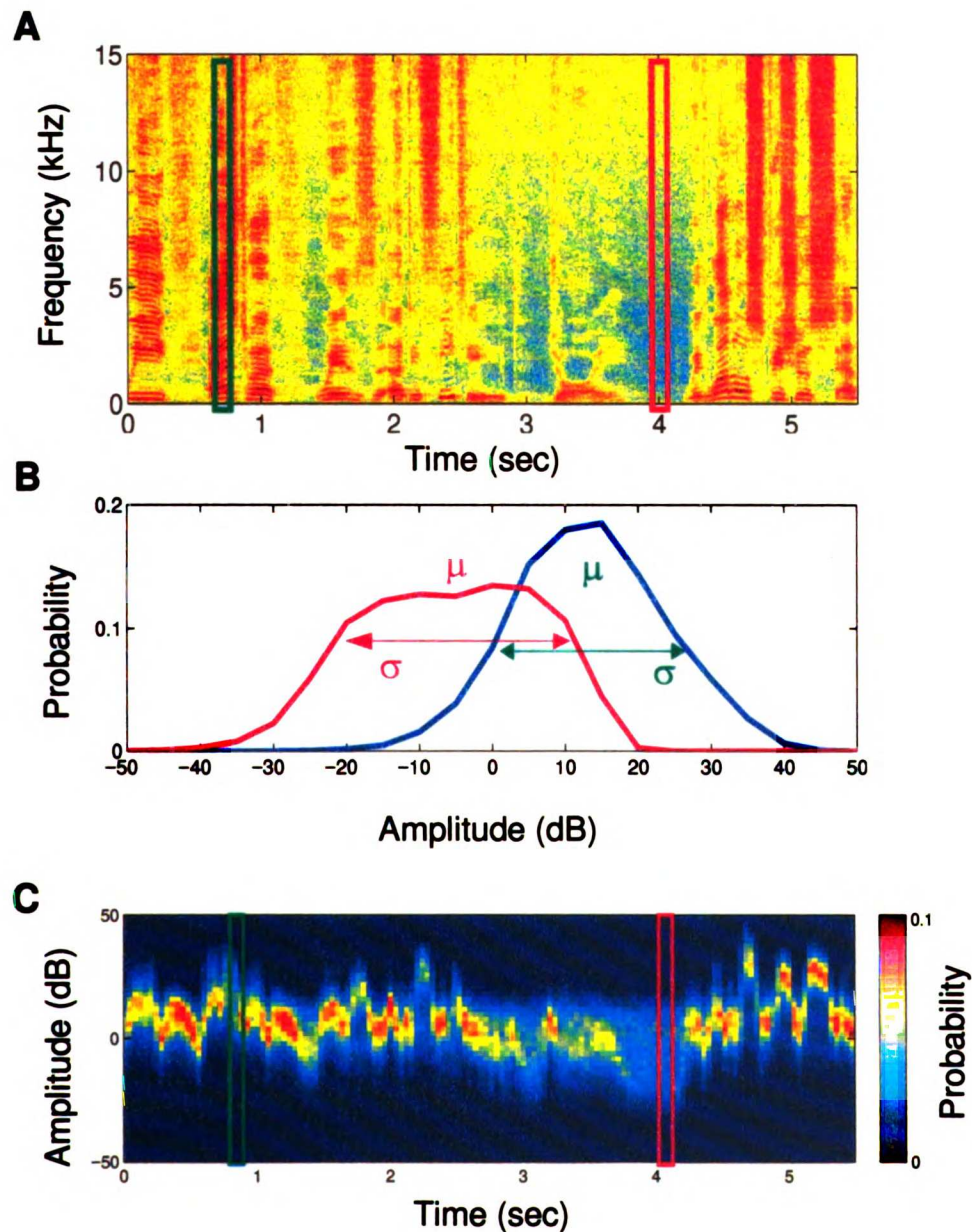


Figure 11: Constructing the time-varying contrast distribution, $C_{dB}(t) = p(\bar{S}_{dB}|t)$.

The sound's spectro-temporal envelope (shown for human speech) is broken up into

overlapping frames of 47 msec width. Two such non-overlapping frames are shown at a separation of roughly 3 sec (A). For each frame, the local contrast distribution is estimated by collapsing the pixels in the chosen frame into a probability histogram (B). The corresponding contrast distributions are shown in (B) for the green and red frames of (A). Note that the shape of the distribution (including the mean and standard deviation) differ from frame to frame. For all time instants, the probability distributions are collapsed into a three-dimensional plot (C) where the colorscale denotes the relative probability. This plot depicts the progression of the contrast distribution with time. The resulting time-varying contrast distribution is non-stationary, changing both in its mean value and its standard deviation as a function of time.

The time-dependent contrast distribution, $C_{ab}(t) = p(\bar{S}_{ab}|t)$, was computed for all soundscapes in the chosen ensembles. Examples of each are provided in Figs. 12–14. For most environmental background sounds the shape of the contrast distribution was globally stationary. An example is shown for the sounds emanating from a waterfall in Fig. 12 A. The shape of the contrast distribution is constant throughout the sound segment. Analogous properties are observed for white noise (Fig. 12 B).

Speech, mixtures of vocalizations, and music, on the other hand, had the character where the mean value, $\mu_{ab}(t) = E[\bar{S}_{ab}|t]$, and the standard deviation,

$$\sigma_{ab}(t) = E\left[\left(\bar{S}_{ab} - \mu(t)\right)^2 | t\right],$$

of the decibel contrast distribution (see Fig. 11 and Fig. 13) were time-dependent and largely determined by the specific sound which dominates the acoustic scene. Fluctuations in the mean of the contrast distribution reflect changes in the mean intensity of the sound whereas fluctuations in the standard deviation reflect the local variability of the amplitude gradations within a 47 msec sound segment. Note that

the intensity fluctuations associated with the mean are themselves a form of contrast (on a large time-scale) which reflects the fact that the contrast distribution is a function of the time-scale over which it is defined.

Vocalization and speech sounds are characterized by non-stationary / time-dependent contrast distributions. Examples are provided in Fig. 11 and 13. The speech segments of Figs. 11 and 13 A oscillate between loud (high μ_{dB}) and soft sound segments (low μ_{dB}) in a time-dependent manner. Furthermore, the width of the contrast distribution (σ_{dB}) also varies with time. Thus the dynamic range of the local spectro-temporal gradations (within the 47 msec analysis frame) change in a time-dependent manner.

Mixtures of vocalizations and environmental noises likewise followed non-stationary contrast statistics. Fig. 13 B shows such an example for an animal vocalization (giant anteater *Myrmecophaga tridactyla*) superimposed on mixture of a background noise. The contrast distribution oscillates between two states which are individually determined by the properties of the noise and the vocalization. In this particular example, the time-dependent mean and standard deviation covaried with each other in a negatively correlated fashion – although this was not always the case. During the vocalization (loud portion of the stimulus) the contrast distribution is narrowest (small $\sigma_{dB}(t)$) whereas during the background sound (soft segment, low $\mu_{dB}(t)$) it is significantly narrower (large $\sigma_{dB}(t)$). For this example this trend resulted from departure of $\bar{S}(t, f_k)$ about the detrending function $\bar{S}(f_k)$.

As for the speech and vocalization sounds, both classical and pop music have non-stationary contrast distributions, $C_{dB}(t)$, with a time-dependent mean value, $\mu_{dB}(t)$, and standard deviation, $\sigma_{dB}(t)$. Examples are provided in Fig. 14. Classical music has contrast statistics which appear as random oscillations of the contrast distribution. By comparison the oscillations of $C_{dB}(t)$ for the pop music ensemble appear to be significantly more structured. For this example, oscillations of the contrast distribution are quasi-periodic and locked to the rhythmic pattern of the music, as evident from the spectrographic representation. Furthermore, fluctuations in the mean and standard deviation were generally much slower for classical music.

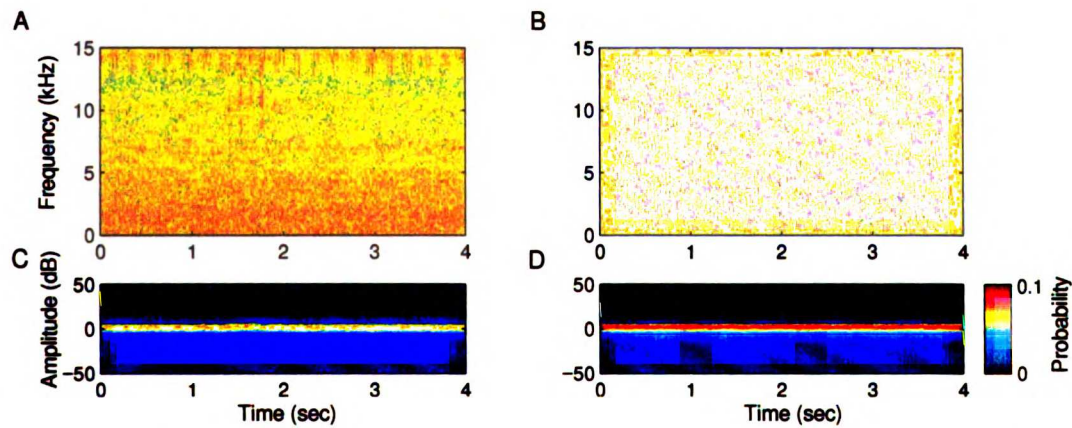


Figure 12: Time-varying contrast distribution for environmental background sound and white noise. The sounds emanating from a waterfall (A) have stationary contrast statistics (C). The time-dependent contrast distribution for this sounds is homogenous for all time (C). Likewise, white noise (B) has stationary contrast statistics (D).

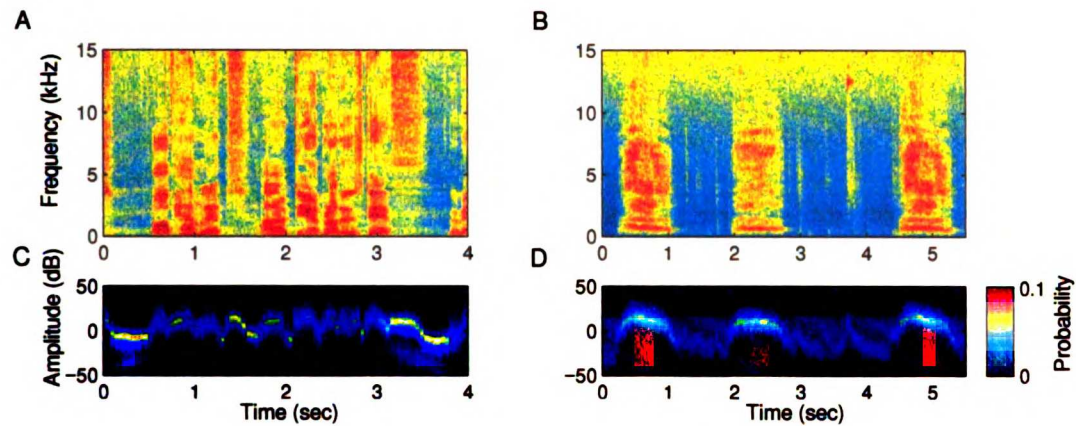


Figure 13: Speech (A) and animal vocalizations (B) have non-stationary (time-dependent) contrast statistics with complex dynamics. The time-dependent contrast distributions, (C) and (D), for the corresponding segments of (A) and (B) oscillates wildly between loud and soft sound segments. Furthermore, the width of the contrast distribution, $C(t)$, also oscillates in a time-dependent manner suggesting that the dynamic range of the local spectro-temporal fluctuations varies with time.

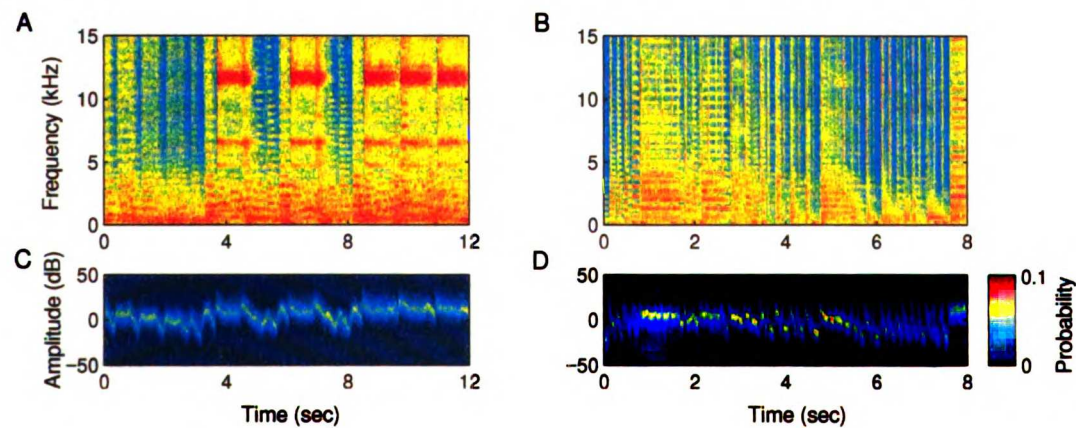


Figure 14: Classical (A) and pop (B) music have non-stationary contrast statistics with complex dynamics. As for speech and vocalization sounds (Fig. 13), the time-dependent contrast distributions, (C) and (D), oscillate between loud (high μ_{dB}) and soft segments (low μ_{dB}). The dynamic range of the local contrast statistics (denoted

by σ_{dB}) are also time-varying.

1.11 Contrast and Intensity Ensemble Statistics

To quantify the observed contrast dynamics for the various sound ensembles, the time-dependent contrast distribution $C_{dB}(t)$ was parametrized by computing its time-dependent mean value $\mu_{dB}(t)$, and its standard deviation $\sigma_{dB}(t)$, (Fig. 15). For all sounds in a given ensemble the joint histogram for these quantities was computed. The joint histogram was normalized so that its cumulative sum gives unity probability. This descriptor approximates the joint distribution function, $p(\mu_{dB}, \sigma_{dB})$, and characterizes the statistical dependence and the relative occurrence of these parameters at time-scales of 47 msec. Ensemble histograms for both parameters are shown in Fig. 16 for human speech, animal vocalizations (primate and non-primate sources), environmental noise sounds (rain, running water, wind, etc.), classical music, pop music, and white noise.

Human speech and animal vocalizations have the character where the relative intensity fluctuations, designated by μ_{dB} , and the local contrast fluctuations, designated by σ_{dB} , are significantly broader and span a larger range of values than environmental noise sounds. This is evident in the speech and vocalization examples of Fig. 11 and 13 where the contrast distribution oscillates wildly in its mean value and its overall width. The parameter σ_{dB} is significantly larger (t-test, $p < 10^{-15}$) for speech and vocalizations than for environmental sounds indicating that the local spectro-temporal fluctuations in these sounds are broader than for environmental sounds. Thus, relative intensity fluctuations and the local contrast statistics present in environmental noise sounds are

relatively homogenous when compared to vocalization sounds and music. Likewise the white noise control stimulus shows little fluctuations in these parameters when measured at time-scales of 47 msec.

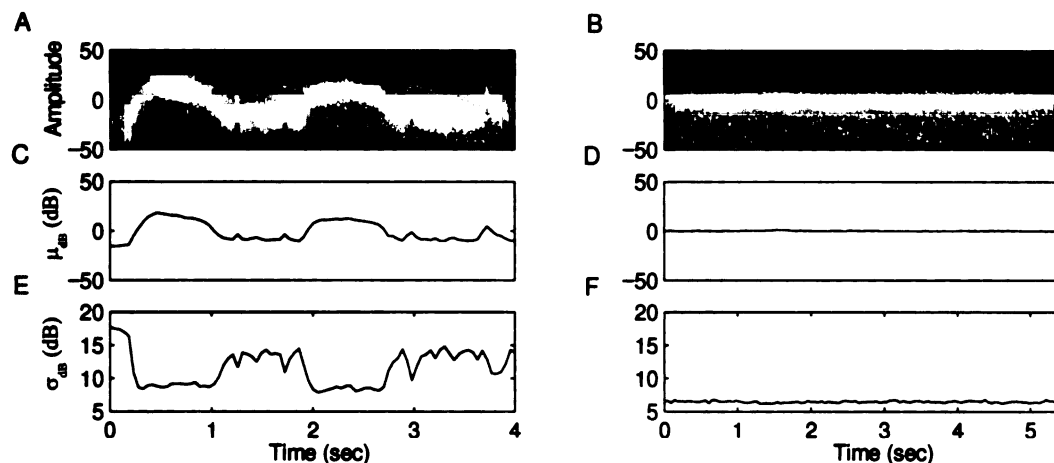


Figure 15: Parametrizing the contrast distribution into its time-varying parameters,

$\sigma_{dB}(t)$ and $\mu_{dB}(t)$. Shown for the vocalization sound of Fig. 13 B (A, C, and E)

and the running water sound of Fig. 12 A (B, D, and F). The time varying mean,

$\mu_{dB}(t)$, designates the instantaneous relative intensity of the sound. The time-

varying standard deviation, $\sigma_{dB}(t)$, is determined by the instantaneous width of the

contrast distribution and is therefore representative of the instantaneous dynamic range

of the sound. For the vocalization example the mean (C) and standard deviation (D)

parameters oscillate wildly as a function of time. By contrast, these parameters are

stationary and have no obvious fluctuations for the water sound (D and F).

For comparison a one dimensional histogram was computed for $\sigma_{dB}(t)$. This

is shown for the different sound classes in Fig. 17. Note that the distribution and the

corresponding mean values for speech and vocalizations are almost identical (mean value

of 9.3 dB for speech and 9.5 dB for vocalizations, t-test $p>0.9$) whereas the distribution for the environmental sound is significantly narrower and has a significantly lower mean value (mean of 7.6 dB, t-test $p<10^{-15}$). The distributions for classical music and pop music were slightly overlapped although the mean value was higher for classical music (10.0 dB versus 8.7 dB, t-test $p<10^{-15}$). To distinguish possible differences between primate and non-primate animal vocalizations, we additionally broke up the parameter signals into those arising from primate and non-primate sources. The distribution for

$\mu_{dB}(t)$ versus $\sigma_{dB}(t)$ were highly overlapped and covered a similar range of values.

As for human speech, the corresponding mean value for $\sigma_{dB}(t)$ were also not significantly different (mean of 9.6 dB for primates and 9.4 dB for non-primate, t-test $p>0.9$).

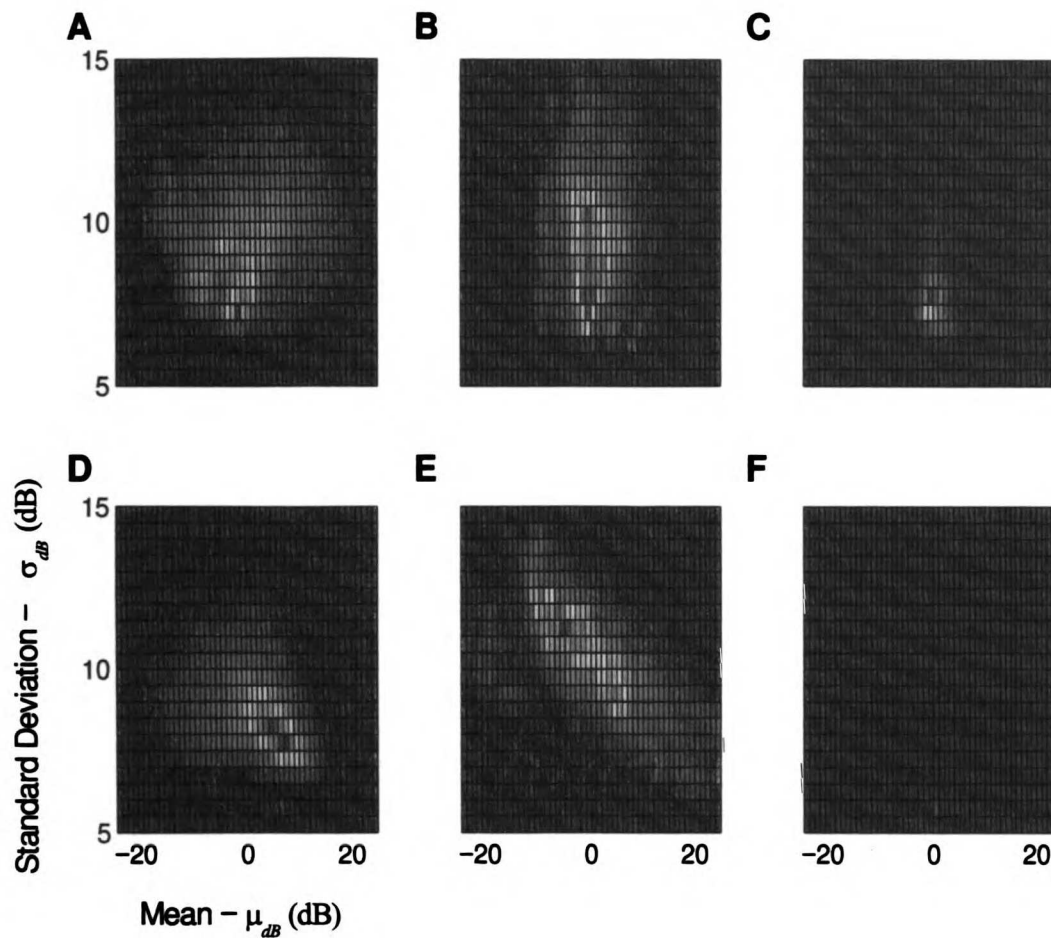


Figure 16: Intensity versus contrast statistics for (A) human speech, (B) animal vocalizations, (C) environmental sounds, (D) pop music, (E) classical music, and (F) white noise. The time-dependent trajectories for the mean and standard deviation of the contrast distribution (Fig. 15) are collapsed into a joint probability histogram. The standard deviation designates the local variability of the spectrographic signal within a 47 msec frame (Fig. 11). The mean designates the average intensity for each frame. Both speech (A) and vocalizations (B) cover a significantly broader range of values than environmental sounds (C) and white noise (F). Classical music shows a significantly broader range of values for σ_{dB} than pop music. In addition, the histogram for classical music is obliquely oriented indicative of a negative correlation.

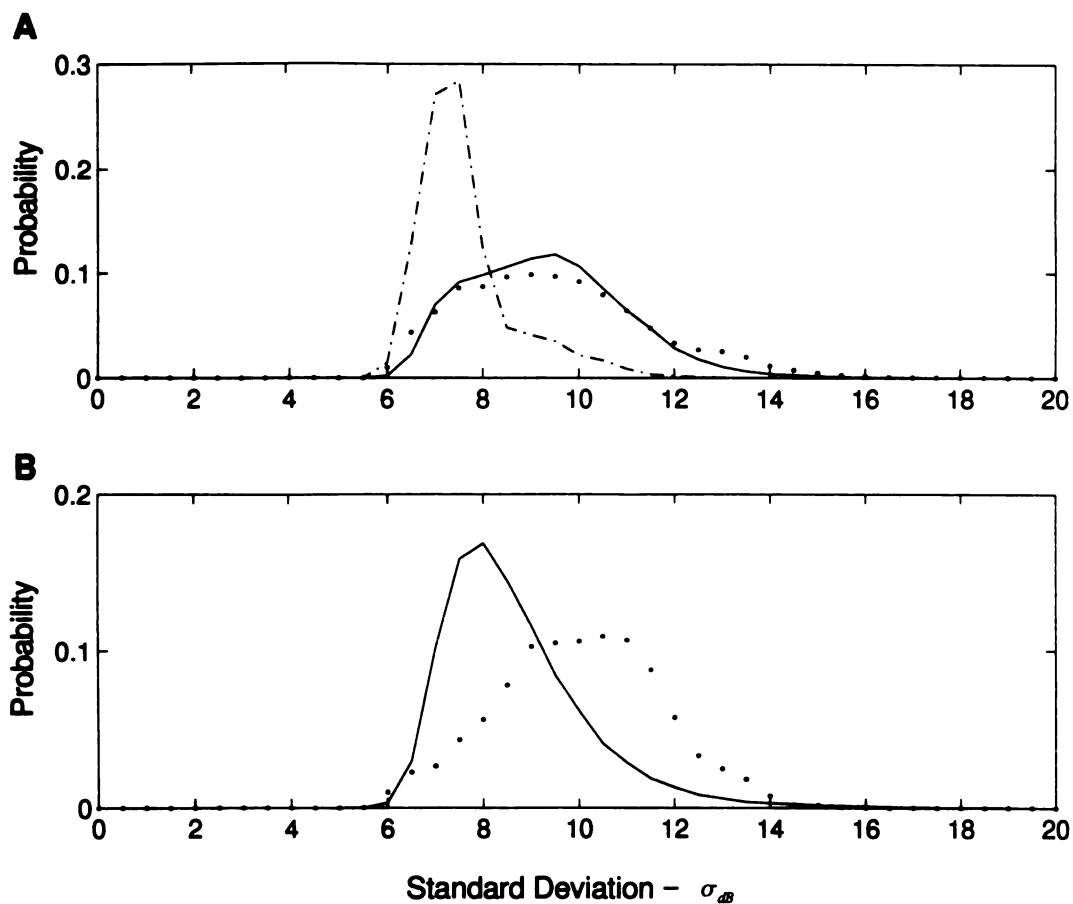


Figure 17: Contrast statistics for the sound ensembles of Fig. 16 (A–E). The standard deviation trajectory distribution is shown for speech (continuous), vocalizations (dotted), and environmental sounds (dashed) in (A). The distribution for speech is highly overlapped with the distribution for vocalizations (mean value of 9.3 and 9.5). The distribution for environmental sounds assumes significantly lower values (mean 7.6). (B) The standard deviation trajectory distribution for pop (continuous) and classical (dotted) music.

The dynamic behavior of these parameters was determined by computing the power spectrum of $\mu_{dB}(t)$ and $\sigma_{dB}(t)$ for each ensemble. The power spectrum for these parameters are shown in Fig. 18. In all instances, the power spectrum had a

decreasing trend as a function of frequency that followed a $1/f$ type functional relationship. The strongest fluctuations in these parameters for all sounds therefore occurred below 1 Hz. In the instance of pop music, the power spectrum also has a strong peak centered about 6 Hz. By listening to this music it was evident that all soundtracks used for this analysis had a strong rhythmic pattern near 6 Hz.

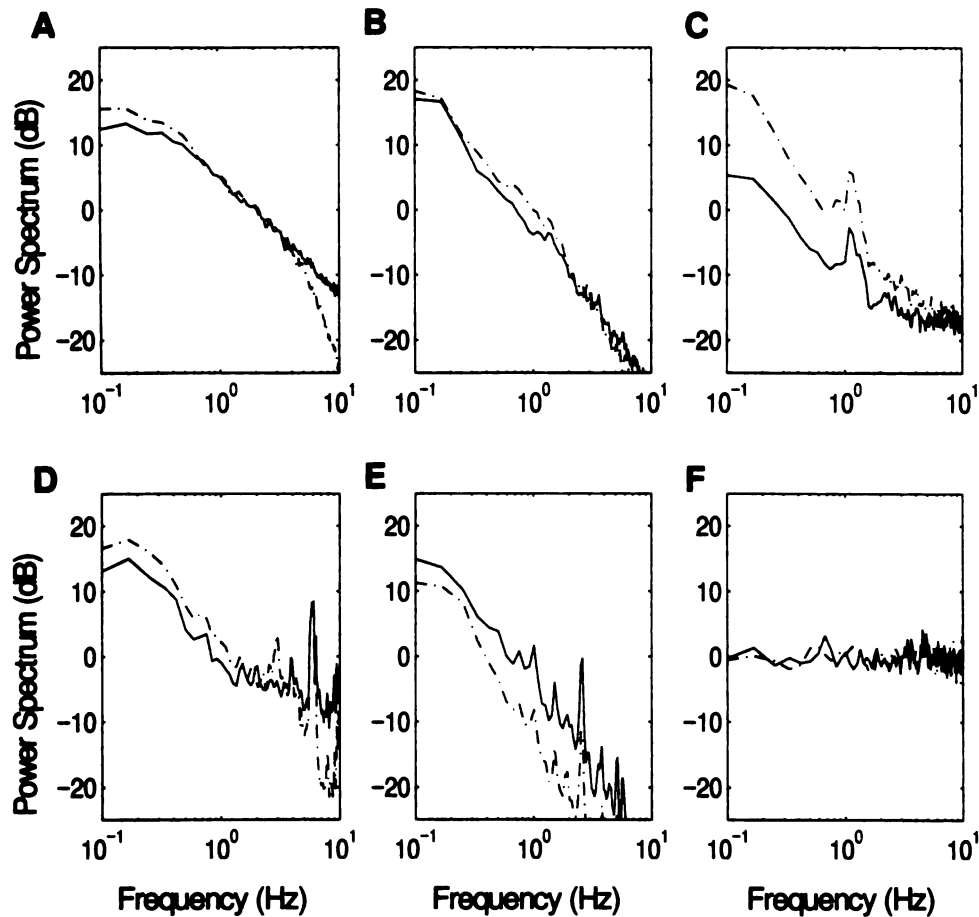


Figure 18: Power spectrum for the parameters $\sigma_{dB}(t)$ (continuous) and $\mu_{dB}(t)$ (dashed-dotted) shown for (A) human speech, (B) animal vocalizations, (C) environmental sounds, (D) pop music, (E) classical music, and (F) white noise. Both parameters have power spectrums with similar trends. With the exception of white

noise, all sounds have a $1/f$ like trend in energy as a function of increasing frequency.

For all ensembles, the power spectrum for $\mu_{dB}(t)$ and $\sigma_{dB}(t)$ were visually very similar (Fig. 18). Thus it is possible that both parameters are temporally correlated following similar trajectories. An alternate possibility, however, is that both parameters do not covary in time and the similarity in the power spectrum arises because the parameters follow similar statistics. To ascertain this possibility, we first computed the Pearson correlation coefficient (Zar 1999) between $\mu_{dB}(t)$ and $\sigma_{dB}(t)$. If both parameters follow similar trajectories it is expected that the correlation coefficient be near unity. If the parameters are temporally uncorrelated, the resulting correlation coefficient will be near zero. Correlation coefficients near negative one alternately indicate that the trajectories are temporally correlated but differ in polarity. Results are provided in Table 2.

Example trajectories for $\mu_{dB}(t)$ and $\sigma_{dB}(t)$ are provided in Fig. 15. For all ensembles, particular examples could be found that showed an anticorrelated ($r < 0$), positively correlated ($r > 0$), and uncorrelated ($r \approx 0$) relationship between these two parameters. Thus the described correlation coefficient provide the average statistics for the whole ensemble. Both the pop and classical music ensembles showed a significant negative correlation (bootstrap, $p < 10^{-10}$). Thus when the spectro-temporal intensity, $\mu_{dB}(t)$, was high the local contrast standard deviation, $\sigma_{dB}(t)$, was reduced and vice versa. In contrast the human speech, vocalization, and environmental sound ensembles have a small but significant positive correlation (bootstrap, $p < 10^{-10}$).

It is possible that the measured contrast standard deviation, $\sigma_{dB}(t)$, reflects a departure from the mean ensemble spectrum (as for the example of Fig. 15A), as opposed to local fluctuations about the mean spectrum. Recall that the detrended spectro-temporal envelope was obtained by performing a global detrending operation on the spectrogram by subtracting the best-fit linear trend of the ensemble spectrum (Eq. (1.21)). To determine if this is so, the *local* spectro-temporal envelope, $\bar{S}_{dB}[n, \omega_k]$, was further detrended using a linear spectral fit of the general form $A\omega_k + B - \mu_{dB}$ (estimated for a 47 ms frame). The detrended spectro-temporal envelope is then given by $\bar{S}_{dB}[n, \omega_k] = 20 \log_{10}(S[n, \omega_k]) - A\omega_k - B + \mu_{dB}$ where A and B now represent the linear regression coefficients for the *local spectrum* and μ_{dB} is the local mean. This procedure removes the local spectral trend but, unlike the detrending operation of Eq. (1.21), it preserved the local mean value, μ_{dB} . All of the presented statistics were reestimated using this procedure. Although specific instances were found where exceptionally high values of $\sigma_{dB}(t)$ were attributed to departure of the detrending function from the local spectrum this was not the general rule. In most instances the obtained results were qualitatively similar for the two detrending procedures although $\sigma_{dB}(t)$ was slightly smaller in value for the local detrending. This parameter followed similar trajectories for either of the performed detrending operations suggesting that fluctuations of the contrast standard deviation do not arise solely from departure between the *local* and the *global* ensemble spectrum. Instead, this result argues that a significant amount of the observed spectro-temporal variability arises from spectro-temporal oscillations about the mean

spectrum. Furthermore, the obtained population statistics of Figs. 16–18 were qualitatively identical indicating that either detrending procedure captures the essential statistical properties for the described ensembles.

	Pearson Correlation Coefficient (r)
Human Speech	0.141 ± 0.003
Animal Vocalizations	0.034 ± 0.004
Background Sounds	0.040 ± 0.008
Pop Music	-0.254 ± 0.003
Classical Music	-0.646 ± 0.003

Table 2: Ensemble correlation statistics between $\sigma_{dB}(t)$ and $\mu_{dB}(t)$. Instantaneous contrast and intensity parameters for human speech, classical music, and pop music are highly correlated across time. Animal vocalizations and background sounds show little covariation among these parameters.

To further understand the dynamic behavior of these higher-order stimulus parameters we computed the coherence function (Marmarelis and Marmarelis 1978; Hayes 1996; Bendat 1990). This descriptor measures the degree of linear association between two signals as a function of frequency. A value near unity for the coherence indicates a high degree of linear association whereas a value near zero is indicative of no linear association. Since the correlation coefficient averages over all sounds and over all temporal segments, information about the time-scales of interaction between these two parameters is discarded. Thus, the correlation coefficient measure can not identify what regime of the power spectral density (Fig. 18) is responsible for the temporal

covariations between the trajectory signals $\mu_{dB}(t)$ and $\sigma_{dB}(t)$.

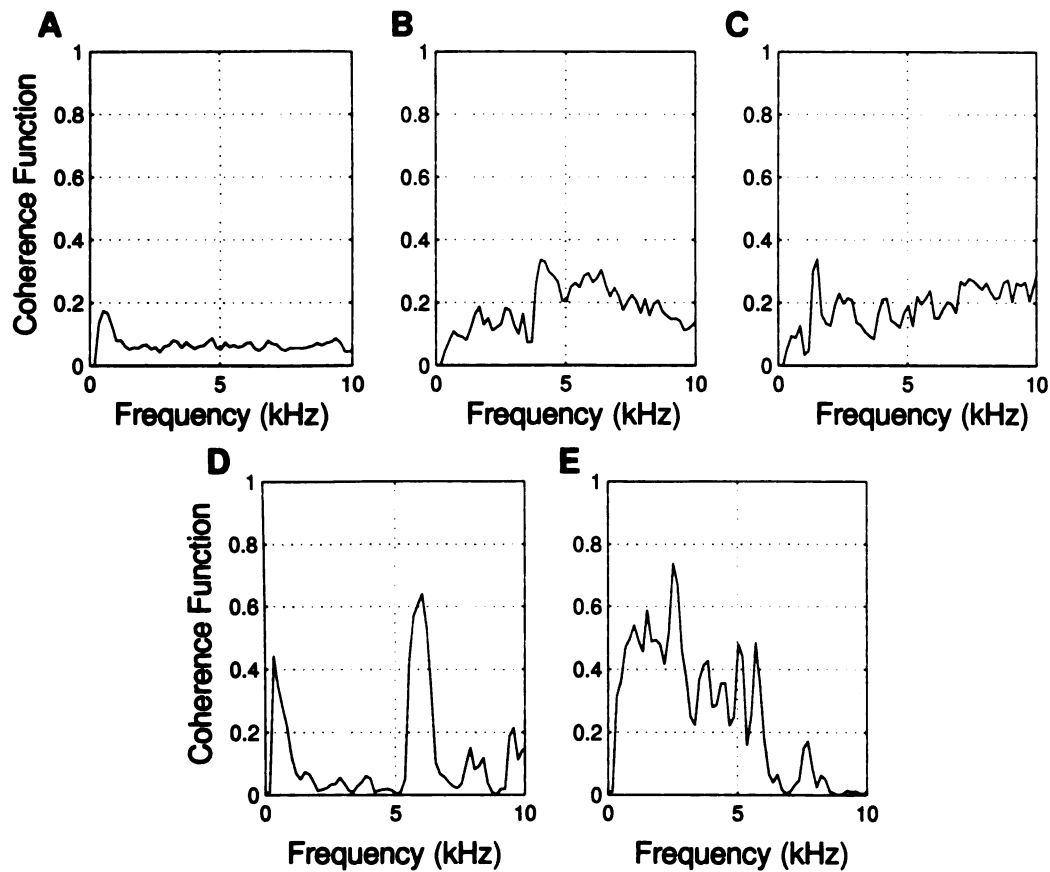


Figure 19: Coherence function between the parameters $\sigma_{dB}(t)$ and $\mu_{dB}(t)$ shown for human speech (A), animal vocalizations (B), environmental sounds (C), pop music (D), and classical music (E). Speech, animal vocalizations, and background sounds show a weak coherence between these two parameters consistent with the measured correlation coefficients of table 2. These parameters show a large amount of correlated signal activity in the vicinity of 6 Hz for pop music and at frequencies below 6 Hz for classical music.

The ensemble coherence functions between the mean and standard deviation

trajectory signals are depicted in Fig. 19. All signals showed a statistically significant coherence (bootstrap, $p < 0.05$) indicating some amount of temporal covariation for these parameters. Consistent with the measured correlation coefficients for the different sound ensembles, pop and classical music had the strongest coherence function. For both of these ensembles, the coherence was localized to a small regime of the frequency axis. Classical music had the strongest coherent oscillations between $\sigma_{dB}(t)$ and $\mu_{dB}(t)$ at frequencies below six Hz. Pop music alternately had coherent activity in the vicinity of six Hz. Vocalizations, speech, and background sounds alternately had weak coherence functions that were not localized along the frequency dimension.

1.12 Discussion and Conclusion

The spectro-temporal envelope of natural sounds is a mathematical construct which describes the spectro (spatio)-temporal neuronal excitation pattern produced by the acoustic sensory epithelium. Because of this it is thought to contain much of the pertinent acoustic information which the brain uses for complex sound analysis and encoding. To date a quantitative evaluation of the relevant statistical components of the spectro-temporal envelope of natural sounds is lacking. In this study, we analyzed a number of higher-order statistical characteristics for five natural sound ensembles. The presented data demonstrates that natural sound ensembles share a number of spectro-temporal characteristics but yet differ in terms of their associated dynamics and the degree of coherency across spectral channels.

Comparisons among natural sound ensembles show that vocalizations, speech, and music have a significant amount of correlated signal components across spectral

channels. These sounds therefore activate the auditory neuronal network with a complex spectro-temporal excitation pattern composed of redundant signal components. By comparison, the across-channel correlation matrices of environmental sounds show significantly lower levels of across-channel correlation and spectro-temporal redundancy. The role of redundant signal information for acoustic processing is in general not well understood although it may bestow the auditory with its robust characteristics for sound processing under a number of adverse conditions. Redundant acoustic information, for examples, may be necessary for detecting relevant acoustic signals in background noise and reverberant environments. How the brain uses such redundant information directly for complex signal analysis and source segregation still needs to be determined, although initial insights are provided by human psychoacoustics studies for speech perception.

Psychoacoustics studies support the observation that speech is highly redundant and that pertinent acoustic information is preserved across spectral channels. By performing a number of modifications of the speech waveform studies have demonstrated that speech contains a large amount of redundant information that is not necessary for detection and classification of speech. Perception of speech, for example, is robust to a number of spectral, temporal, and amplitude alterations. Filtered speech changes significantly in its overall quality when filtered above or below 1.8 kHz. Despite this, much of the necessary information for identifying and distinguishing speech segments is retained when such filtering is performed (Moore 1997). Other alterations include infinite peak clipping of the amplitude waveform which converts the speech waveform to a binary sequence. Despite the loss of amplitude information such highly

distorted speech can be understood by listeners (Moore 1997) who achieve word articulation scores of 80–90%.

Spectro–temporal correlations likely play a significant role in auditory grouping and auditory scene analysis. Psychoacousticians have demonstrated that sound components that are presented in temporal unison often "group" together to form an unified percept or an acoustic stream (Moore 1997). A classic example is the case where two pure tones at distinct frequencies are coherently modulated by a common envelope. The two sounds are perceived as a part of a whole and can not be distinctly identified. If the same pure tones are instead modulated by independent temporal envelopes, the two sounds segregate and are each perceived as a distinct entity. Given the observed spectro–temporal correlations and differences among natural sound ensembles, it is likely that such signal statistics are pertinent for sound source segregation. From a neuronal coding perspective it is plausible that the observed different levels of spectrographic correlations may be pertinent for signal detection, inter–category discrimination, and sound source segregation by the auditory neuronal network.

Physiologic studies on cats and songbirds have further demonstrated the importance of the spectro–temporal envelope and the inherent redundancy which exists across spectral channels (Theunissen and Doupe 1998) of natural sounds. Using procedures which degrade the spectral and/or temporal resolutions of a sounds spectrogram, these studies have demonstrated that neuronal responses of auditory neurons to natural sounds are robust under various adverse conditions. Neuronal responses appear to be extremely robust against spectral degradations but are significantly more sensitive to temporal modifications. These findings stress the relative

importance of temporal over spectral information for acoustic processing and further demonstrate notion that natural sounds contain a significant amount of redundant signal information.

A secondary acoustic property which may facilitate signal detection and the reliability of sensory coding is the signal contrast. Contrast is a fundamental property of all sensory signals including visual, somatosensory, and acoustic. In most instances the contrast of a sensory signal is specified by the signal's peak to minimum intensities or its standard deviation. In vision, for example contrast is generally specified by the ratio between the difference signal intensity and the mean level or luminance. Visual contrast is therefore specified by the equation: $C = (I_{Max} - I_{Min}) / (I_{Max} + I_{Min})$ where I_{Max} and I_{Min} designate the maximum and minimum signal intensities. In general such description are insufficient since they can't account for the intermediate values of the sensory signal which may be equally and possibly more physiologically relevant than the extremum values. We provide a more complete description of the spectro-temporal contrast or the amplitude gradations of natural stimuli by considering the complete probability distribution, as opposed to simpler descriptions such as the modulation index or the standard deviation.

The contrast distribution was examined for both the linear amplitude and the logarithmic (decibel) amplitude spectro-temporal envelope. In the first case, the amplitude distribution of natural sounds are skewed towards zero containing a high proportion of the signal's amplitude values at low levels. By comparison, the linear amplitude distribution for white noise is broadly distributed and therefore does not share **this** statistical attribute observed for natural sounds. The decibel spectro-temporal

envelope amplitude distribution of all natural sound ensembles was broadly distributed qualitatively resembling a normal distribution of amplitude values. The average range of values, as measured by the standard deviation, was broadest for vocalizations (13.3 dB), music (11.5 dB) and speech (11.0 dB) and narrowest for environmental sounds (7.4 dB). Not surprisingly the overall range of values spanned by white noise was significantly smaller (5.6 dB).

Given that the auditory system of humans has a dynamic range of more than five orders of magnitude it is of interest to determine how and if the observed decibel distributed amplitude fluctuations are utilized for efficient sound encoding and/or sound categorization. One hypothesis of sensory encoding asserts that the dynamic range of the input sensory stimulus must be physically matched to the operating range of the neural system in order to maximize the information transfer and encoding ability (Rieke *et al.* 1997). The work of Attias and Schreiner (1998 a and b) demonstrates the importance of the complete statistics of the amplitude signal and the effects on the neuronal encoding ability. From the presented data one interesting observation which is consistent with this hypothesis is the observation that the average range of values spanned by natural sounds ($2\sigma_{dB}$) and the 90th percentile range (roughly $3\sigma_{dB}$) is comparable in magnitude to the average dynamic range of peripheral auditory neurons which typically span a dynamic range 30–60 dB (Evans and Palmer 1980; Veimeister 1988). Furthermore since the rate–level dependencies of auditory neurons of the peripheral auditory system have a linear dependence with decibel intensity, an efficient probing sound would span a *the* decibel amplitude dimension. In fact, from an information theoretic perspective the *stimulus* which would most efficiently drive such a system (i.e. with a linear rate versus

level (SPL) dependency) would follow normally distributed decibel contrast statistics as is the case for natural sounds.

Although all of the studied natural sounds had logarithmic distributed contrast fluctuations and therefore shared a common attribute, the measured variability and dynamics were distinctly different across the five sound ensembles. By comparing these sounds at pertinent time-scales for neuronal and perceptual integration of intensity, it is shown that environmental sounds are time-invariant and have little spectro-temporal variability whereas vocalized sounds and music have non-stationary contrast statistics. Using a perceptually relevant time-scale of 47 ms the contrast distribution was decomposed into temporally disjoint segments. The running contrast distribution was then analyzed by computing the time-varying mean, $\mu_{dB}(t)$, and the contrast standard deviation, $\sigma_{dB}(t)$. These descriptive parameters describe the intensity fluctuations and the instantaneous contrast statistics of the stimulus respectively. Analogous to white noise, environmental background sounds (e.g. running water and wind) showed narrow distributions for both of these parameters suggesting that they are relatively homogenous over the analyzed time-scales. By comparison, the distribution of values for these parameters, $p(\mu_{dB}, \sigma_{dB})$, was significantly broader for vocalizations, speech, and music. Consequently these sounds show strong intensity and contrast fluctuations at the analyzed time-scales.

Further evaluation of the temporal dynamics and temporal covariation among *these* parameters for the different ensembles reveals that the intensity and contrast *fluctuations* have $1/f$ like spectrum with most of the parameter signal energy residing at

low frequencies (below 1 Hz). Similar observations have been described for the intensity fluctuations of speech and music (Voss and Clarke 1975; Voss and Clarke 1978) although a clear picture of the temporal covariations among the spectro-temporal contrast and the spectro-temporal stimulus intensity has not been described. Here we additionally show that the time-varying mean value, $\mu_{dB}(t)$ and standard deviation, $\sigma_{dB}(t)$, parameters obtain from the spectrographic envelope show significant amounts of covariation. By comparing the coherence function between these two parameters it is demonstrated that vocalized and environmental sounds have weak coherence functions whereas musical sounds have the strongest covariations among these parameters. Furthermore the strong covariations observed in musical sounds were most strongly isolated at particular frequencies below 10 Hz whereas for speech, vocalizations, and background sounds the temporal covariations appear to be less frequency specific.

The described spectro-temporal statistic show that natural signals have a wealth of information which can be feasibly used by the auditory system for stimulus coding and categorization. Currently, due to the limited knowledge of the general properties of natural sounds little is known as to weather these acoustic parameters are pertinent for sound perception and stimulus encoding in the central nervous system. Judging from human psychophysics data and the fact that the response of central auditory neurons is strongly affected by closely related stimulus parameters (including intensity, modulation depth, and spectral correlation), it is expected that these parameters may be pertinent for sound processing. The dependence of the neuronal response on the described acoustic parameters will be studied in the following sections. Chapter 3 addresses the issue of spectro-temporal correlations and their effect on the response of inferior colliculus

neurons. The dependence of the neuronal response as a function of the stimulus contrast is evaluated further in chapter 4.

1.13 References

- D.G. Albrecht. Visual cortex neurons in monkey and cat: Effects of contrast and spatial and temporal phase transfer function. *Visual Neurosci.* **12**, 1191–1210 (1995).
- H. Attias, and C.E. Schreiner. Temporal Low Order Statistics of Natural Sounds. *Advances in Neural Information Processing Systems* **10**, 27–33 (1998a).
- H. Attias, and C.E. Schreiner. Coding of Naturalistic Stimuli by Auditory Neurons. *Advances in Neural Information Processing Systems* **10**, 103–109 (1998b).
- J.S. Bendat. *Nonlinear systems analysis an identification from random data*. John Wiley and Sons, New York, 1990.
- L.H. Carney. A model for the responses of low–frequency auditory nerve fibers in cat. *J. Acoust. Soc. Am.* **93**, 401–417 (1993).
- L.H. Carney and C.D. Geisler. A temporal analysis of auditory nerve fibers responses to spoken stop consonant–vowel syllables. *J. Acoust. Soc. Amer.* **79**, 1896–1914 (1986).
- T. Chi, Y. Gao, M.C. Guyton, P. Ru, and S. Shamma. Spectro–temporal transfer functions and speech intelligibility. *J. Acoust. Soc. Am.* **106** (5), 2719–32 (1999).
- L. Cohen. *Time Frequency Analysis*. Prentice Hall, New Jersey, 1995.
- B. Delgutte and N.Y.S. Kiang. Speech coding in the auditory nerve. I. Vowel–like sounds. *J. Acoust. Soc. Am.* **75** (3), 867–879, 1984.
- D.W. Dong and J.J. Atick. Statistics of natural time–varying images. *Network: Computation in Neural Systems* **6** (3), 345–58, 1995.
- E.F. Evans. The frequency response and other properties of single fibers in the guinea–pig cochlear nerve. *J. Physiol.* **226**, 263–287 (1972).
- E.F. Evans and A.R. Palmer. Relationship Between the Dynamic Range of Cochlear Nerve Fibers and Their Spontaneous Activity. *Experimental Brain Research* **40**, 115–118, 1980.
- W.R. Garner and G.A. Miller. The masked threshold of pure tones as a function of duration. *J. Exp. Psychol.* **37**, 293–303, 1947.
- C.D. Geisler and T. Gamble. Responses of "high–spontaneous" auditory–nerve fibers to constant–vowel syllables in noise. *J. Acoust. Soc. Amer.* **85**, 1639–1652, 1989.
- D.D.** Greenwood. A cochlear frequency position function for several species – 29 years

- later. *J. Acoust. Soc. Am.* **87**, 2592–2605, 1990.
- J.D. Harris. Loudness discrimination. *J. Speech Hear. Disord. Monographs, Supplement* **11**, 1–63 (1963).
- M.H. Hayes. *Statistical Digital Signal Processing and Modeling*. Wiley & Sons, 1996.
- J. W. Hughes. The threshold of audition for short periods of stimulation. *Proc. R. Soc. B.* **133**, 486–490, 1946.
- L.B. Jackson. *Digital Filters and Signal Processing*. Kluwer Academic Publishers, 1989.
- R.L. Jenison, S. Greenberg, K.R. Klunder, and W.S. Rhode. A composite model of the auditory periphery for the processing of speech based filter response functions of single auditory–nerve fibers. *J. Acoust. Soc. Am.* **90** (1), 773–786, 1991.
- N.Y.S. Kiang, T. Watanabe, C. Thomas, and L.F. Clark. *Discharge Patterns of Single Fiber's in the Cat's Auditory Nerve*. Cambridge, MA: MIT Press (1965).
- W. Jesteadt, C.C. Wier, and D.M. Green. Intensity discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.* **61**, 169–177 (1977).
- C.E. Liberman. The cochlear frequency map of the cat: Labeling auditory–nerve fibers of unknown characteristic frequency. *J. Acoust. Soc. Am.* **72**, 1441–1449, 1982.
- R.F. Lyon. A computational model of filtering, detection and compression in the cochlea, ICASSP 82, Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing, 1982
- P.Z. Marmarelis, V.Z. Marmarelis. *Analysis of Physiological Systems Modeling. The White Noise Approach*, Plenum Press, New York, 1978.
- G.A. Miller. Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *J. Acoust. Soc. Am.* **191**, 609–619, 1947.
- B. Moore. *An introduction to the psychology of hearing*. Academic Press, New York, NY, 1997.
- J.P. Nordmann, R.D. Freeman, and C. Casanova. Contrast sensitivity in Amblyopia: Masking Effects of Noise. *Investigative Ophthalmology and Visual Science* **33**, 2975–2985 (1992).
- A. V. Oppenheim and R. W. Schaffer. *Discrete Time Signal Processing*. Prentice Hall, New Jersey. 1989.
- J. W. Picone. Signal Modeling Techniques in Speech Recognition. *Proc. IEEE* **8** (9), 1215–1247, 1997.

- D.L. Ruderman. Origins of scaling in natural images. *Vision Research* **37** (23), 3385–3398, 1997.
- D.L. Ruderman and W. Bialek. Statistics of Natural Images: Scaling in the Woods. *Physical Review Letter* **73** (6), 814–817, 1994.
- J.A. Simmons, D.J. Howell, N. Suga. Information content of bat sonar echoes. *American Scientist* **63**:204–215, 1975.
- J.W.T. Smolders, A.M.H.J. Aertsten, and P.I.M. Johannesma. Neural representation of the acoustic biotope: A comparison of the response of auditory neurons to tonal and natural stimuli in the cat. *Biological Cybernetics* **35**, 11–20 (1979).
- S.S. Stevens. On the psychophysical laws. *Psychol. Rev.* **64**, 153–181, 1957.
- S.S. Stevens. Perceived level of noise by Mark VII and decibels (E). *J. Acoust. Soc. Am.* **93**, 425–434, 1972.
- A.Ya. Supin, V.V. Popov, O.N. Milekhina, and M.B. Tarakanov. Ripple depth and density resolution of rippled noise. *J. Acoust. Soc. Am.* **106** (5), 2800–5 (1999).
- T.W. Troy, A.E. Krukowski, N.J. Priebe, and K.D. Miller. Contrast-invariant orientation tuning in cat visual cortex: Thalamocortical input tuning and correlation-based intracortical connectivity. *J. Neurosci.* **18** (15), 5908–5927 (1998);
- R.M. Roark and M.A. Escabí. B-spline design of maximally flat and prolate spheroidal-type FIR filters. *IEEE Trans. on Signal Processing* **47** (3), 701–716, 1998.
- F. Rieke, D. Warland, R. de Ruyter van Steveninck, W. Bialek. *Spikes: Exploring the Neural Code*. The MIT Press, 1997.
- M.B. Sachs and E.D. Young. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *J. Acoust. Soc. Am.* **66** (2), 470–479, 1979.
- S.A. Shamma. Speech processing in the auditory system I: The representation of speech sounds in the response of the auditory nerve. *J. Acoust. Soc. Am.* **78** (5), 1612–1621, 1985.
- F.E. Theunissen and A.J. Doupe. Temporal and spectral sensitivity of complex auditory neurons in the nucleus HVC of male zebra finches. *J. Neuroscience* **18** (10), 3786–3802, 1998.
- T.M. Van Veen and T. Houtgast. Spectral Sharpness and Vowel Dissimilarity. *J. Acoust. Soc. Am.* **77** (2), 628–634, 1985.

N.F. Viemeister and S.P. Bacon. Intensity discrimination, increment detection, and magnitude estimation for 1-kHz tones. *J. Acoust. Soc. Am.* **84**, 172–178 (1988).

N.F. Viemeister. Intensity coding and the Dynamic Range Problem. *Hearing Research* **34**, 267–274, 1988.

R.V. Voss and J. Clarke. 1/f noise in music: Music from 1/f noise. *J. Acoust. Soc. Am.* **63**, 258–263 (1978).

R.V. Voss and J. Clarke. 1/f noise in music and speech. *Nature.* **258**, 317–318 (1975).

K. Wang and S.A. Shamma. Auditory analysis of spectro-temporal information in acoustic signals. *IEEE Engineering in Medicine and Biology* **14** (2), 186–194, 1995a.

K. Wang and S.A. Shamma. Spectral shape analysis in the central auditory system. *IEEE Transactions on Speech and Audio Signal Processing* **3** (5), 382–395, 1995b.

T.C.T. Yin, J.C.K. Chan, and D.R.F. Irvine. Effects of interaural time delays of noise stimuli on low-frequency cells in the cat inferior colliculus. I. Responses to wideband noise. *J. Neurophysiol.* **55**, 280–300, 1986.

E.D. Young, and W.F. Brownell. Responses to tones and noise of single cells in the dorsal cochlear nucleus of unanesthetized cats. *J. Neurophysiol.* **39**, 282–300, 1976.

J.H. Zar. *Biostatistical Analysis*. Prentice Hall, New Jersey, 1999.

Rippled Noise Stimulus Design: Theoretical and Ecological Considerations

Abstract

Complex acoustic stimuli, such as speech and music, have time-varying spectrum that give rise to rapidly changing frequency transitions and temporal periodicities. Despite this, central auditory representations are most often probed using simple acoustic stimuli which lack many of the structural components of natural sounds. Given the complexity of the auditory neuronal network and the fact that the brain is in general extremely nonlinear, it is increasingly clear that simple acoustic stimuli can not be used directly to identify many of the processing schemes which the auditory system uses for complex sounds analysis. Thus the question arises: should one use natural sounds directly to study central auditory representations? Or, should one use complex synthetic stimuli that are specifically tailored for a particular application?

In this chapter we consider how complex acoustic stimuli can be systematically tailored to incorporate basic attributes present in natural sounds and how these can be used to identify nonlinear processing abilities of auditory neurons. Throughout, we outline a number of necessary experimental, ecological, psychoacoustical, physiological, and theoretical considerations which should be taken into account when designing such complex stimuli. We proceed by designing two sounds that incorporate a number of low-order and high-order characteristics of natural sounds, are parametrically accessible, and are theoretically compatible with reverse correlation procedures. Analogous to natural sounds, these stimuli are broad-band, spectro-temporally complex, and are particularly well suited for studying various nonlinear transformations that may exist along the auditory pathway. The usefulness of this approach and its applicability for physiological systems is verified in chapters 3 and 4.

2.1 Probing the Auditory System with Simple Sounds

Much of our understanding of central auditory function is derived from studies which use simple sounds to probe neuronal sensitivities. With the exception of the bat and songbird auditory systems, ethologic considerations have had only a minor impact in our general understanding of central auditory function. However, the neuroethologic approach used in the bat and songbird has taught us that simple sounds can not reveal many of the neural specializations which the auditory system uses for natural sound processing (Suga *et al.* 1975; Suga and Jen 1976; Margoliash 1983; Olsen and Suga 1991a 1991b; Margoliash and Fortune 1992; Ohlemiller *et al.* 1996; Razak *et al.* 1999). With the exception of a handful of studies (Aersten 1980 1981; Schreiner and Calhoun 1994; Kowalski, Depireux, and Shamma, 1996a 1996b; Attias and Schreiner 1998a 1998b; Nelken, Rotman, and Yosef 1999), the use of natural sounds and complex stimuli which incorporate statistical and structural sound features has not been readily adopted in other mammalian species. Although such an approach may be advantageous it has nonetheless eluded much of the auditory community.

By far the most widely used acoustic stimulus for studying central auditory representations is the pure tone. This stimulus is commonly used to map the frequency response area of neuron. The pure tone has a basic appeal to most auditory physiologists since peripheral and central auditory neurons respond to a restricted range of frequencies and since the lemniscal auditory pathway is organized with respect to frequency in a tonotopic fashion (Liberman 1982; Greenwood 1990; Fay and Popper 1992). Hence it is not uncommon to think of the auditory pathway as performing a Fourier-like decomposition of incoming acoustic sounds. Since a pure tone excites a restricted portion

of the primary sensory epithelium, pure tones allow one to investigate and map local neuronal sensitivities.

Unfortunately natural sounds are seldom narrow-band and they rarely resemble pure tones. Instead, natural sounds are often broad-band, spectrally complex, and time-varying with rapidly changing onsets and offsets. To understand how such characteristic features are represented in the brain, auditory scientists have used a vast number of simple sounds which independently probe each of these stimulus dimensions. Temporal preferences, for examples, are most often studied using sinusoidal amplitude modulated tones (e.g., Schreiner, Urbas, and Mehrgardt 1983; Rees and Moller 1987; Langner and Schreiner 1988) or repetitive clicks trains (e.g. Eggermont 1999). These sounds can test the ability of a neuron to follow rapidly changing sound transitions and periodic events. Sounds such as frequency modulated sweeps are additionally used to investigate neuronal responses to transient events with time-varying frequency transitions (e.g., Rees and Moller 1987; Mendelson *et al.* 1993).

Spectral selectivity of auditory neurons are alternately tested using various broad-band stimuli and combinations of narrow-band stimuli. White noise and clicks, for example, provide a simple complement to the pure tone which allow one to characterize neuronal responses to broad-band sounds (e.g., Ehret and Moffat 1985; Young and Brownell 1976; Yin, Chan, and Irvine 1986). Two tone response tuning curves are often constructed to probe excitatory and inhibitory neuronal response characteristics. Ripple noise stimuli were introduced by Houtgast (Houtgast 1977) to study the psychophysical limits of spectral filtering and lateral inhibition of the auditory system. Recently these sounds have been used to thoroughly characterize spectral

response sensitivities and lateral inhibition of primary auditory cortex neurons (Schreiner and Calhoun 1994; Calhoun and Schreiner 1998; Kowalski *et al.* 1996a 1996b).

2.2 Nonlinear Auditory Processing

Although all of these stimuli can provide valuable insight into the workings of auditory neural networks and their spatial arrangements, results using such sounds can not be easily compared or extended to more complex and dynamic stimulus scenarios. This is in part due to the fact that the brain is highly nonlinear. If the brain were to perform a linear decomposition of incoming sounds, then the responses to complex stimuli (i.e. vocalizations, speech, sound mixtures etc.) could be understood by simply observing responses to its constituent components. The superposition principle of a linear system guarantees (Marmarelis and Marmarelis 1978) that the systems behavior for complex stimuli can be extrapolated by studying the systems behavior for simpler stimuli.

A complex stimuli $x(t)$, for example, can be decomposed using some basis set of local stimulus features, $x_k(t)$, which occur at delays, τ_l . The complex stimulus is therefore represented by

$$x(t) = \sum_{k=1}^N \sum_{l=1}^{L_k} x_k(t - \tau_l) \quad (2.1)$$

where L_k is the number of occurrences of the k^{th} feature at temporal delays τ_l and N is the number of acoustic features which the sound is decomposed into. The response of

a linear time invariant system to such a stimulus is given by

$$r(t) = \sum_{k=1}^N \sum_{l=1}^{L_k} r_k(t - \tau_l) \quad (2.2)$$

where the response, $r(t)$, of a complex stimulus is simply the sum of responses to its individual components, $r_k(t)$.

For a nonlinear system, Eq. (2.2) will in general not hold and the true responses to a complex sound will deviate from Eq. (2.2). The amount of departure from Eq. (2.2) depends on the nature of the nonlinearity. In general, three types of response components contribute to this departure: response interaction terms in which the response of one feature, $x_k(t)$, can be strongly affected by a nearby component, $x_m(t)$. Response gain terms can alternately magnify the response of a single component by a nonlinear gain factor causing large departures from Eq. (2.2). Such a nonlinear level dependence is a common feature of nonlinear systems and the central auditory system as a whole (Ehret and Merzenich 1988; Eggermont 1989). Thirdly, dynamic nonlinearities, which are prevalent in neural systems, can alter the shape of the systems nonlinearity and filtering characteristics in a time dependent manner (Smirnakis *et al.* 1997). Consequently, although the complex stimulus can be decomposed as a superposition of many simpler stimuli (Eq. (2.1)), the response of a nonlinear system to such can not be decomposed into the sum of the individual response components. Hence the functional rules which the brain uses for natural sound processing can not be easily and fully extrapolated using simple stimuli such as pure tones, clicks, modulated tones, etc.

2.3 Neuroethology Versus the Systems Approach

Linear system theoretic approaches are commonly used in conjunction with such **sounds** to characterizing auditory neuronal responses. Of these, the transfer function **method** (Rees and Moller 1983; Langner and Schreiner 1988; Schreiner and Calhoun 1994; Calhoun and Schreiner 1998; Kowalski *et al.* 1996a 1996b) is by far the most **widely** used functional descriptor. This is attributed to the fact that this methodology is **theoretically** well defined, easy to interpret, and experimentally tractable (i.e. sounds are **easy** to design and the data analysis is simple). More recently the linear spectro-temporal **receptive field** (impulse response estimates) (Aertsen *et al.* 1980 1981; Hermes *et al.* 1981; Yeshurun, Wollberg, and Dyn 1987; deCharms, Blake, and Merzenich 1998; Theunissen *et al.* 2000; Klein *et al.* 2000) has also been used for studying central auditory neuronal response properties.

Although such methods do reveal quasi linear processing characteristics of auditory neurons, either of these methods generally lack the ability to discern "hard" nonlinear response characteristics which may be prevalent in central auditory neurons (Young 1998). In part this is attributed to the types of spectro-temporal sound ensembles (spectro-temporal m-sequences, randomly distributed tone pip ensembles, modulated tones, click trains etc.) which are used to study neuronal responses with such methods. For *STRF* methods, for example, the stimuli which are most prevalent generally have spectro-temporal white-noise like properties but the envelope spectrum, otherwise known as the characteristic function (Cohen 1995), is significantly reduced. This is done so that the range of modulation frequencies and spectral periodicities is matched to the

range for which the specific brain region is responsive. Although such "white-noise" like **stimuli** are in principle well suited for such analysis, occurrences of higher-order **acoustic** features and combinations, which may be necessary to drive highly nonlinear (**selective**) neurons, are rare at high power levels within the limited experimental **recording** time.

In the bat auditory system, for example, nonlinear processing provides a **substrate** for processing behaviorally relevant sounds. Central auditory neurons for such **species** can show strong selectivity to behaviorally relevant sounds (Suga, Simmons, and Jen 1975; Suga and Jen 1976; Margoliash 1983), combination sensitivity to conjunctions of **biologically** important acoustic stimuli (Suga, O' Neil, and Manabe 1978; Margoliash and Fortune 1992; Olsen and Suga 1991a 1991b; Doupe 1997) and context dependent response characteristics (Ohlemiller, Kanwal, and Suga 1996; Razak, Fuzessery, and Lohuis 1999). All of these response characteristics clearly arise from highly nonlinear phenomena such as neural inhibition, thresholding, and adaptive response mechanisms which are common to neural systems (Casseday, Ehrlich, and Covey 1994; Kuwada *et al.* 1997; Spiro, Dalva, and Mooney 1999; Bringuier *et al.* 1999). When tested with simpler stimuli such as pure tones and white noise, these neurons often show little or no response (Theunissen *et al.* 2000). Hence such neurons are not easily characterized with simple stimuli and the linearizing approaches associated with them.

Because of this, system theoretic approaches are not widely used to study species with highly specialized acoustic behaviors. Instead neuroethology has been the prominent driving force for scientist who study bat echolocation and avian song recognition. Ecological considerations are generally employed in the selection of the search stimuli

and features that are used to probe neuronal responses. For such species this task is **relatively** simple largely because of the stereotyped calls which these animals use during **their** vocalizing behavior. Although these approaches have proved noteworthy **identifying** nonlinear response characteristics and mechanisms, they are not infallible and **can** easily lead to misconceptions and oversimplifications of the neural capabilities of **these** animals. Given that such systems are generally extremely nonlinear, any knowledge **gained** with a given sound about the systems properties will likely hold true only for that **sound** and experimental condition. An example of such, is provided by the same **investigators** which previously showed that the bat auditory cortex possesses a high **degree** of neural specialization for echolocation tasks. FM-FM neurons of the mustache bat **respond** selectively to combinations of FM segments in that species' echolocation **calls** (Suga, O' Neil, and Manabe 1978). Recently Suga and his collaborators **demonstrated** that these neurons also respond selectively to a variety of communication **sounds** and therefore serve an important secondary function (Ohlemiller *et al.* 1996). In **certain** instances, the sounds that are used to study auditory processing for such species **are** therefore much too constrained to fully characterize more general response attributes.

As for the bat and songbird, similar principles have also been employed for **studying** the audio-vocal behavior of primate species (Winter and Funkenstein 1973; **Glass** and Wolberg 1983; Ploog 1981 ;Wang *et al.* 1995) although these have not **revealed** similar neuronal specializations. Studies in the squirrel monkey have shown that **nearly** all neurons respond to simple sounds (pure tone, pips, and noise) and to species **specific** vocalizations although they respond to the latter unselectively (Winter and **Funkenstein** 1973). Unlike the bat and songbird neurons which can respond almost

exclusively to a single sound, primate cortical neurons generally respond to numerous **sounds** and the neuronal responses appear to be correlated with low-order features of the **driving** stimulus (i.e. spectral energy distribution, temporal structure, etc.) (Winter and **Funkenstein** 1973; Glass and Wolberg 1983; Ploog 1981; Wang *et al.* 1995). This may **reflect** operating principles for primates that are vastly different than those for the bat and **songbird** species where neurons can respond exclusively to a single sound component or to **combinations** of such. However, these primate studies were largely conducted in primary **auditory** cortex whereas the specialized processing in bats is most clearly expressed in **stations** outside of AI.

In mammals in which the auditory system lacks obvious specializations and **relevant** acoustic behavioral paradigms (i.e. many terrestrial mammals such as the cat and **possibly** including primates), neuroethology has had little impact, since the set of **biologically** relevant stimuli is enormous. Consequently, linearizing methods which use **engineering** principles (such as reverse correlation and the transfer function method) by **testing** neuronal responses to a wide range of simple sounds are most often employed. A **handful** of studies have shown that nonlinear phenomena underlie the ability of central **auditory** neuron to extract information inherent in natural sounds (Nelken *et al.* 1999; **Attias** and Schreiner 1998b). Aside from these, however, much of central auditory **physiology** in the cat has proceeded by using stimuli which are for the most part **ethologically** uninteresting.

Hence a clear dichotomy is established between neuroethology and the more **general** systems approach used to characterize auditory neuronal responses in the cat. **While** the systems approach utilizes a vast collection of simple sounds to probe multiple

stimulus conditions and operating points, the ethological approach uses a biased and **highly** restricted stimulus set which only probes a limited operating regime. At least two **major** limitations are anticipated, the first arising from the methodology and the second **arising** from the types of stimulus used to characterize the system.

In the systems approach, the types of stimuli used may not provide enough **driving** force, especially if the neuron's nonlinearities are specifically adapted ("hard" **nonlinearities**) for processing a given stimulus feature and/or combinations. Such is the **case** for feature selective neurons in the bat and songbird and for bat combination **sensitive** FM-FM neurons. Hence this leads us to the notion that to "see" or characterize **such** a nonlinearity you must first provide sufficient driving force along the appropriate **stimulus** dimension. Simple sounds often lack many of the higher-order statistics and **correlations** necessary to properly drive such auditory neurons. Hard nonlinearities that **are present** for specific scenarios of sound processing are therefore not easily **characterized** using such sounds. In many instances, especially scenarios where the **system** has a "soft" or weak nonlinearity, the systems approach can be advantageous **since** it offers a simple parametric description of the stimulus-response relationship and **since** each stimulus dimension can be explored quasi-independently.

The ethological approach likewise has advantages and disadvantages. As **previously** mentioned, application of neuroethologic principles has shared large amounts **of success** for studying hard nonlinearities in the bat and songbird species. This is largely **attributed** to the fact that identifying the "right" stimulus is a fairly simple task for **animals** which actively vocalize because of the obvious behavioral needs and biological **importance** of their vocalizations. In mammals which rely heavily on passive listening,

e.g. for hunting and avoiding predators, this passive mode hampers the identification of **relevant** sounds and acoustic features. Consequently for such animals (i.e. the cat) finding **the** right stimulus is not a trivial task since the set of biologically relevant stimuli is very **large**. For these animals behaviorally relevant stimulus paradigms have not been **revealed**. Hence although this methodology is readily applicable for the bat and songbird **species**, it is not easily applicable for nonspecialized and acoustically passive terrestrial **mammals**.

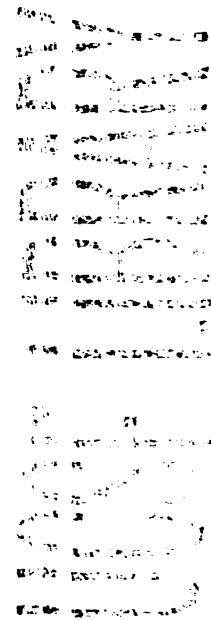
In theory one can imagine taking a huge collection of natural sounds and playing **them** continuously to an animal in order to overcome this limitation. Such attempts (**Smolders et al.** 1979), however, have not revealed similar specializations that arise from **hard** nonlinearities. This is partly due to the fact that for such sound schemes one must **contend** with the high dimensionality of the stimulus. When using natural sounds one **inevitably** probes the auditory system with many physical dimensions that include **carrier** structure (e.g. harmonic vs. inharmonic), spectral and temporal envelope (first, **second**, and higher-order statistics), intensity, binaurality (both interaural intensity and **temporal** differences). Likewise one may conceive of this process as probing the system **with** the corresponding perceptual dimensions which include comodulation, pitch, **rhythms**, loudness, streaming, and timbre (which itself is a multidimensional percept **depending** on the spectral envelope, temporal changes in time, and whether the sound is **harmonic** or inharmonic (Plomp 1969 1970). Either way, it is clear that there is a **geometric** explosion of stimulus parameters and response possibilities. This makes it **particularly** difficult from the analysis point of view since it ultimately increases the **complexity** of the analysis required to identify hard nonlinearities and to disassociate

neuronal responses and mechanisms.

Another limitation for this approach is that for natural sounds all of the physical **stimulus** dimensions are already highly biased and these may be different for different **stimulus** ensembles. For example, the temporal modulation spectrum has a $1/f$ **characteristic** (Attias and Schreiner 1998a 1998b; chapter 1) as does the spectral envelope (**chapter 1**) and intensity fluctuations of the stimulus (Voss and Clarke 1975). **Consequently**, in many natural sounds low frequencies predominate over higher **frequency** components for these stimulus parameters. Although one can in principle **circumvent** these problems by detrending (via deconvolution) the stimulus bias when **using** reverse correlation methods (Theunissen *et al.* 2000), this may not be feasible for **the full** range of sounds especially if the signal power for a relevant parameter is so **small** that the response signal to noise ratio is below chance. Under such conditions the **stimulus** correlation matrix will be non-invertible and the actual transfer function for the **given** parameter can not be determined reliably. Practically, this situation is quite **common** due to the high modulation index of natural sounds (Van Veen and Houtgast 1985), and the effects of background noise and environmental acoustic (Schroeder, Gottlob, and Siebrasse 1974).

For experimental paradigms where one is interested in identifying neuronal **mechanisms**, one instantly realizes the bottleneck and limitations that can arise from **using** continuous soundscapes and the simple stimuli used for the systems approach. For **sounds** which are designed to be compatible with reverse correlation methods (such as **spectro-temporal** m-sequences), relevant higher-order acoustic features (e.g. spectral **resonances**, FM sweeps) which are necessary to efficiently drive auditory neurons are

commonly underrepresented. The use of natural soundscapes can partly overcome this by **limiting** structural sound components to those which are likely more relevant. For certain **natural** sound components (e.g. high frequency envelope components), however, this **may** not be reasonable since these can be relatively scarce (because of the stimulus bias **and** high dimensionality of natural sounds) preventing the observer from achieving **sufficient** statistical power. Moreover, because of the geometric complexity of natural **sounds**, statistical bias, and large number of acoustic dimensions it is exceedingly **difficult** to dissociate responses arising from a single acoustic feature and/or dimension. **One** must therefore jointly consider the limited amount of experimental recording time **which** is available (this is ultimately determined by the electrode stability of the **experimental** setup and it is usually in the range of tens of minutes to several hours) and **the stimulus** space of interest. Either way, when using such sounds to derive *STRFs* one **may** ultimately be wasting precious recording time by exploring only a small subspace of **the target** objective or parameter, while driving the neuron or system (for most of the **recording** epoch) with numerous other sound features that are not of immediately **interest**. Although natural sounds may be efficient stimuli to drive the auditory system as **a whole**, the statistical bias of their spectro-temporal composition and high **dimensionality** can prohibit a clear understanding of the underlying neuronal principles.



2.4 Stimulus Requirements for Deriving *STRFs*

The spectro-temporal receptive field (Aertsen *et al.* 1980 1981; Hermes *et al.* 1981; Yeshurun *et al.* 1987; Nelken *et al.* 1997; deCharms *et al.* 1998; Theunissen *et al.* 2000; Klein *et al.* 2000) provides a linear model for characterizing the response area of

auditory neurons. In the visual system, the analogous functional descriptor is the spatio-temporal receptive (Jones and Palmer 1987; Deangelis, Ohzawa, Freeman 1993; Anzai *et al.* 1999; Reich *et al.* 2000) field. This linear descriptor has been successfully used to describe the response areas of visual neurons along the space-time dimensions. Conceptually, the *STRF* can be thought of as the optimal linear descriptor which jointly characterizes a neuron's spectral (spatial for the visual system) and temporal preferences.

In general, the *STRF* serves as a linear model which can be used to predict neuronal responses to arbitrary stimuli. For a quasi linear neurons, the *STRF* serves as an invaluable tool since it retains much of the neuron's transfer function characteristics which are necessary for predicting neuronal responses. For a nonlinear neuron, however, the *STRF* may not fully generalize and will often fail at describing the neuron's transfer function attributes. The ability to characterize highly nonlinear neurons therefore depend strongly on the neurons operating point and on the driving stimulus used to derive the *STRF*. Here we consider the stimulus requirements which are necessary for deriving auditory spectro-temporal receptive fields.

We consider a multi-input single output linear filter bank (Marmarelis and Naka 1974) as a model representation for auditory neuronal filtering. This representation consists of a set of N adjacent linear filters tonotopically arranged along the primary sensory epithelium (e.g. the cochlea). This representation is motivated by the fact that the primary sensory epithelium performs a spectro-temporal decomposition of incoming sounds and consequently all further processing along the auditory system is constrained by this output pattern.

Given a spectro-temporal representation for a sound, $S(t, X_k) = s_k(t)$, the signal

$s_k(t)$ describes the temporal modulations for the k^{th} input channel (tonotopically arranged). We use a filter model to describe the neuronal integration and temporal dynamics of response for a given channel. The spectro-temporal filter bank model consists of a set of N octave spaced linear filters, $[h_1(\tau), h_2(\tau), \dots, h_N(\tau)]$, where

$h_k(\tau) = h(\tau, X_k)$ is the impulse response of a linear filter centered about the frequency band X_k and τ corresponds to the temporal lag of the filter. Here X_k corresponds to the center frequency of the k^{th} filter in units of octaves. Taken together, the spectral array of N filters describe the spectro-temporal integration dynamics for a single neuron.

For such a model neuron the overall response output, $r(t)$, is obtained by summing the response for each of the tonotopically arranged frequency channels

$$r(t) = r_0 + \sum_{k=1}^N r_k(t) \quad (2.3)$$

where r_0 is the neuron's mean firing rate (zeroth-order kernel),

$$r_k(t) = \int s_k(t-\tau) h_k(\tau) d\tau + e_k(t) \quad (2.4)$$

is the output for the k^{th} frequency channel, $s_k(t) = s(t, X_k)$ is the input of the k^{th} filter channel, and $e_k(t)$ is a noise term that arises from measurement error and the neuron's

internal noise. For practical reasons, we assume that $e_k(t)$ has zero mean and standard deviation denoted by σ_e . Furthermore, $e_k(t)$ is statistically independent of the input signal $s_k(t)$. The response of the k^{th} frequency band corresponds to the linear temporal convolution between the k^{th} input, $s_k(t)$, and the k^{th} impulse response, $h_k(\tau)$, as described by Eq (2.4). Note that the input stimulus, $s_k(t)$, varies along the temporal and spectral axis and therefore corresponds to a spectro-temporal stimulus representation, more commonly referred to as the spectro-temporal envelope (Kowalski *et al.* 1996a; Klein *et al.* 2000).

In practical applications, it is desired to estimate the spectro-temporal receptive field of a neuron using the reverse correlation procedure. This procedure consists of performing a crosscorrelation between the neuronal response and input driving stimulus.

Unlike the one dimensional stimulus case, where the response difference output,

$r(r) - r_0$, is crosscorrelated with a single input, the described spectro-temporal

representation requires that the response be crosscorrelated with each of the N inputs. For

the linear model neuron this procedure is expressed as

$$E[(r(t) - r_0) \cdot s_i(t + \sigma)] = \sum_{k=1}^N E[r_k(t) s_i(t + \sigma)] = \quad (2.5)$$

$$\sum_{k=1}^N \int E[s_k(t - \tau) s_i(t + \sigma)] h_k(\tau) d\tau + E[e_k(t) s_i(t + \sigma)] \approx$$

$$\sum_{k=1}^N \int R_{ss}(\tau - \sigma, X_k - X_i) h_k(\tau) d\tau$$

where $E[\cdot] = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} \cdot dt$ is the time average operator, $l=1, \dots, N$, and $R_{ss}(\tau, \zeta)$

is the stimulus spectro-temporal autocorrelation function. For a sufficiently large

recording period, T , the error crosscorrelation $E[e_k(t)s_l(t+\sigma)]$ approaches zero since

$e_k(t)$ and $s_l(t+\sigma)$ are statistically independent and both have zero-mean. If the spectro-temporal autocorrelation of the stimulus has the unique property that it has

impulse like characteristics, that is $R_{ss}(\tau, \zeta) = \sigma_s^2 \delta(\tau, \zeta)$, then the spectro-temporal crosscorrelation between the stimulus and the output simplifies to

$$E[(r(t) - r_0) \cdot s(t + \sigma, X_l)] = \sigma_s^2 \sum_{k=1}^N \int \delta(\tau - \sigma, X_k - X_l) h(\tau, X_k) d\tau = \sigma_s^2 h(\sigma, X_l) \quad (2.6)$$

Here σ_s is to the standard deviation of the stimulus spectro-temporal envelope. The

spectro-temporal receptive field, $h(\sigma, X_l)$, for the model neurons is instantly derived

as

$$h(\sigma, X_l) = \frac{1}{\sigma_s^2} \cdot E[(r(t) - r_0) \cdot s(t + \sigma, X_l)] \quad (2.7)$$

Thus the linear neuron's spectro-temporal impulse response (i.e. its *STRF*) can be

estimated directly by performing a crosscorrelation between the neuron's response,

$r(t)$, and each of its N individual inputs, $s_k(t)$, for $k=1, \dots, N$.

For a neuronal spike train, $r(t) = \sum_i \delta(t-t_i)$ of M neuronal event times, t_i ,

Eq. (2.7) can be easily expanded as a spike triggered average

$$h(\sigma, X_i) = \frac{1}{\sigma_s^2 T} \sum_{i=1}^M s(t_i + \sigma, X_i) . \quad (2.8)$$

In practice, T corresponds to the experimental recording period which is in all instances a finite quantity. Hence, from Eq. (2.6) and (2.8), the estimation of the linear *STRF* is greatly simplified by considering a spectro-temporal stimuli with an impulsive spectro-temporal autocorrelation function. One implication of this result is that the only prerequisite for deriving the *STRF* for a linear model neuron via Eq. (2.8) is that the grand average spectro-temporal autocorrelation function have impulse like properties, regardless of whether the stimulus is stationary or non-stationary (Eq. (2.5) and (2.6)). Consequently, one can consider classes of acoustic stimuli that retain the global requirements necessary for deriving *STRFs* (i.e. impulsive global autocorrelation function), but yet are ethologically derived. We will consider a class of nonstationary sounds with strongly biased instantaneous correlation statistics. In particular, such stimuli may be of particular interest for studying various classes of nonlinear auditory neurons which often do not respond efficiently to the white noise and m-sequence type stimuli that are commonly used for reverse correlation procedures. The goal of such stimuli, as will be described subsequently, is to provide increased nonlinear driving force using acoustic stimulus features that are known to efficiently drive auditory neurons.

2.5 Testing for Nonlinearity

A common procedure for characterizing and determining the relative degree of **nonlinearity** of a neuronal systems is to estimate its higher-order system kernels. Using **such** a procedure for estimating the nonlinear contributions of a system is analogous to **fitting** a nonlinear function by a Taylor series expansion. The main distinction between **the** Taylor expansion and Volterra systems representation is that the system's Volterra **kernels** describe a nonlinear filtering transformations, whereas the elements of the Taylor **expansion** (i.e. $f_1(x)=x$, $f_2(x)=x^2$, $f_3(x)=x^3$ etc.) describe a nonlinear **transformation** without any filtering.

Most often this approach of characterizing system nonlinearities is generally not **extended** beyond second-order due to experimental limitations which limit the amount of **recorded** data. Although such descriptors are indeed useful for describing subtle **nonlinearities**, they are nonetheless faced with practical limitations since they require **large** amounts of data, are computationally intensive, and are often difficult to interpret. **Given** the finite experimental recording time of neurophysiologic experiments, the **estimation** of higher-order kernels is further confounded by the fact that white noise like **stimuli**, which are prerequisite for deriving higher-order kernels, often do not provide **sufficient** nonlinear driving force (e.g. higher-order correlations are weak) to activate **very** nonlinear system elements. Furthermore, for many neural systems relevant aspects **of** the system transformation are best described by dynamic nonlinearities (Smirnakis *et al.* 1997) which are often not easily described using a Volterra/Wiener series **representation**.

The described multi-input/single-output representation for the linear model neuron of Eq. (2.3) and (2.4) can be formally extended to include nonlinear elements and across-channel nonlinear interactions. The spectrographic multi-input nonlinear representation for a model neuron can be expressed as

$$r(t) = r_0 + \sum_{k=1}^N r_k(t) + \sum_{k=1}^N \sum_{l \neq k} r_{kl}(t) + \dots \quad (2.9)$$

where

$$r_k(t) = \sum_{n=1}^{\infty} r_{k,n}(t) \quad (2.10)$$

is the Volterra expansion of the k^{th} input channel. The n^{th} order term

$$r_{k,n}(t) = \int \dots \int x_k(t-\tau_1) \dots x_k(t-\tau_n) h_{k,n}(\tau_1, \dots, \tau_n) d\tau_1 \dots d\tau_n \quad (2.11)$$

describe the nonlinear filtering contributions to the neuron's firing rate that is produced

by the k^{th} input channel. Here $x_k(t)$ is the input to the k^{th} filter channel and

$h_{k,n}(\tau_1, \dots, \tau_n)$ is the n^{th} order nonlinear kernel for this channel. The n^{th} order kernel describes the nonlinear filtering transformation between the input and the output of this channel. For the special case of a linear model neuron (Eq. (2.3) and (2.4)) the kernels exist only for $n=1$.

In the third term in the series of Eq. (2.9)

$$r_{kl}(t) = \sum_{n=1}^{\infty} r_{kl,n}(t) \quad (2.12)$$

corresponds to a sum of the n^{th} order interaction products between the k^{th} and l^{th} input channels. These terms describe the functional interactions between any two channels. As an example one can consider the second-order interaction product between the k^{th} and l^{th} input channels. This is described by a second-order convolution

$$r_{kl,2}(t) = \int \int h_{kl,2}(\tau_1, \tau_2) x_k(t-\tau_1) x_l(t-\tau_2) d\tau_1 d\tau_2 \quad (2.13)$$

between the inputs $x_k(t)$ and $x_l(t)$ and the second-order nonlinear cross-kernel,

$h_{kl,2}(\tau_1, \tau_2)$, which describes the second-order nonlinear filtering function between

the k^{th} and l^{th} input channels. The output, $r_{kl,2}(t)$, corresponds to the firing rate

contribution that is produced by the nonlinear interaction between these two input

channels. All of these operations can be extended to include higher-order interactions

products between any number of input channels.

The procedure for identifying the linear kernel of the system outlined in section 2.4 can be extended directly for identifying the higher-order and cross kernels of the nonlinear model neuron of Eq. (2.9) (Marmarelis and Marmarelis 1978). This approach requires identification of the higher-order terms of the series expansion via a higher-order reverse correlation procedure analogous to Eq. (2.7). Although this approach is in theory well suited for rigorously identifying the higher-order nonlinear attributes of the

system under study, it is in general cumbersome, computationally intensive, and requires large amounts of data. Thus in practical applications this method is not feasible and is not extended beyond a second-order analysis of the system's kernels (Yeshurun, Wollberg, and Dyn 1987). Furthermore, unlike the linear model neuron scenario, where the reverse correlation procedure extracts the systems linear impulse response directly, the measured *STRF* is no longer identical to the systems linear spectro-temporal impulse response. Instead the estimated kernels are now a composite functions of the linear and the nonlinear elements of the system. Also, the experimentally measured *STRF* is now a function of the driving stimulus used to characterize the system.

For the described nonlinear neuron, the essential relationship between the neuron's Volterra kernels and the measured *STRF* is (if one ignores cross-channel interactions and uses white noise) (Marmarelis and Marmarelis 1978, Eq. 4.50, pg. 150)

$$STRF_w(t, X_k) = w_k(\sigma) = \quad (2.14)$$

$$\sum_{m=0}^{\infty} \frac{(2m+1)! \sigma_s^{2m}}{m! 2^m} \times \int_0^{\infty} \cdots \int_0^{\infty} h_{2m+1}(\tau_1, \tau_1, \dots, \tau_m, \tau_m, \sigma_1) d\tau_1 \cdots d\tau_m$$

where the experimentally measured kernels for each channel, $w_k(\sigma)$, are now referred to as Wiener kernels. The label *STRF_w* is used to denote the Wiener kernel derived *STRF*. Note that unlike the linear model neuron scenario, where the reverse correlation procedure produces the systems linear kernels directly, the derived *STRF_w* is now a sum of projections of the odd-order Volterra kernels, h_{2m+1} , and a functions of the stimulus

power σ_s^2 .

In many instances, the derived Wiener kernel *STRF* is advantageous since it contains linear and nonlinear stimulus-response characteristics in one descriptor. Note that the higher-order nonlinear projections are progressively weaker for higher-order nonlinear elements (because of the overpowering denominator term in Eq. (2.14); also see Fig. 1 for illustration) so that the Wiener-derived *STRF* largely captures linear response characteristics. This descriptor is optimal in the sense that it provides maximal information about the systems transfer characteristics (since it combines linear and nonlinear information). In fact, it is possible to derive an *STRF_w* for a nonlinear system even in the absence of linear system elements. The drawback of this descriptor is that it needs to be reestimated for each stimulus condition and operating point in order to preserve its optimal properties. Alternately, the Volterra *STRF* representation is advantageous in that all of the terms are distinct and invariant as a function of any stimulus parameter (e.g., stimulus power and other high-order stimulus characteristics) and, consequently, they never requires reestimation.

Given these basic properties we devise a scheme for identifying complex nonlinearities that may be pertinent for neuronal encoding. As described, it is theoretically possible to identify nonlinear response characteristics by estimating the system's higher-order kernel using a higher-order reverse correlation procedure. However, experimental and practical limitations prevents us from doing so. An indirect

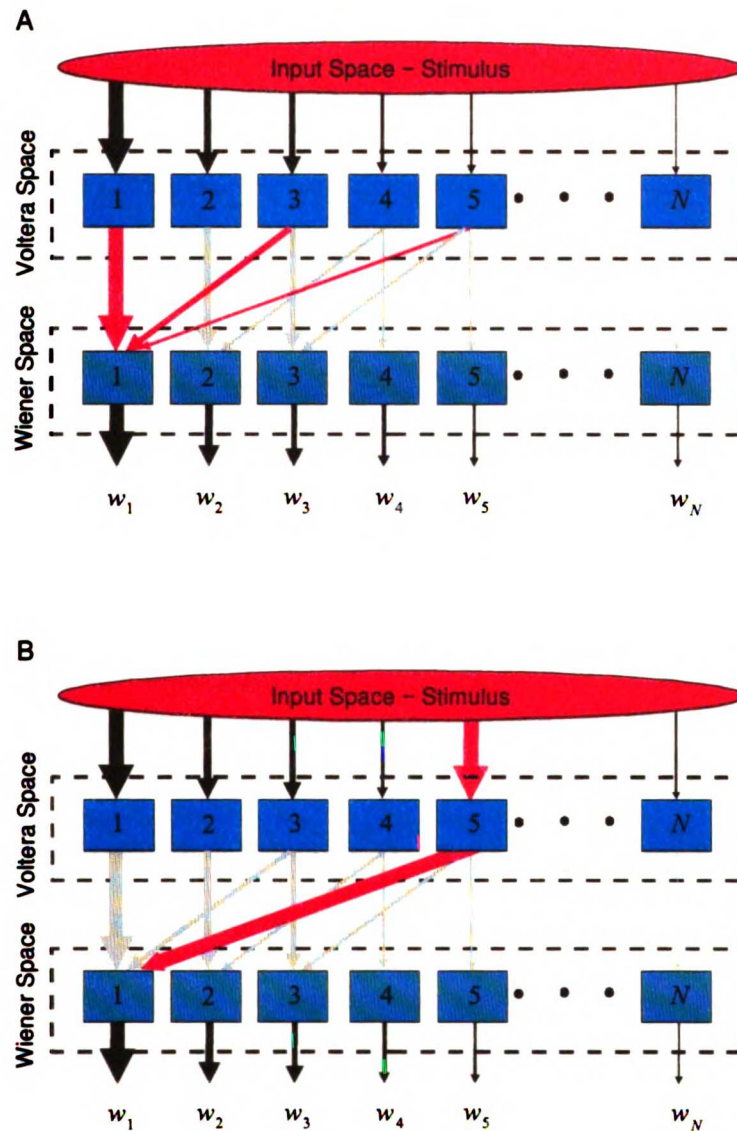


Figure 1: (A) Relationship between the input stimulus, Volterra system kernels, and the Wiener system kernels. The input stimulus is represented by a sequence of higher-order stimulus correlations. These are depicted as distinct inputs to the system or, equivalently, the Volterra space. The Volterra space can be thought of as the physical elements of the system where the order designates the order of the described nonlinearity. For most reverse correlation stimuli, these higher-order inputs get progressively weaker (with increasing order). The Volterra kernels project onto the

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

measured Wiener kernels, w_k . For the 1st-order Wiener kernel (e.g. the *STRF*) the projections arise from all of the odd-order Volterra kernels (depicted in red). As for the input patterns of the stimulus, these projection patterns get progressively weaker with increasing order. (B) Altering the projection pattern onto the first-order Wiener kernel by altering the fifth-order input of the stimulus (altered input depicted in red).

approach around this problem, is to systematically alter the stimulus higher-order spectro-temporal correlations, so that the effective projection pattern from nonlinear terms (Eq. 2.14) is altered. This procedure is schematized in Fig. 1 for a single channel of the nonlinear filter bank model.

We consider two input signals, $S_A(t, X)$ and $S_B(t, X)$ and use these to perform an *A/B* comparison of the neuron's response. By design the two signals are chosen so that their first-order autocorrelation functions $R_{ss}(\sigma, \zeta)$ are identical. Only the higher-order correlation functions are different and these can be chosen by the experimenter based on a priori knowledge of the higher-order correlations that may be pertinent. As an example consider the projection arising from the fifth-order Volterra element of Fig. 1. We can magnify this projection by magnifying the fifth-order correlations of $S_B(t, X)$ (Fig. 1 B) while keeping them intact for $S_A(t, X)$. For the linear model neuron of Eq. (2.3) and (2.4) it is expected that the derived *STRF* be identical for both sounds since for a linear neuron, the derived *STRF* is only a function of the first-order autocorrelation function (i.e. only dependent on the projection arising from the first Volterra element of Fig. 1A), $R_{ss}(\sigma, \zeta)$ (see Eq. (2.5)). For a nonlinear neuron, however, the derived

$STRF_w$ is a function of the higher-order correlations of the stimulus (Eq. (2.14)), which are in this case distinctly different for $S_A(t, X)$ and $S_B(t, X)$. In this case, the projection arising from fifth-order Volterra element is magnified and this is reflected directly in the $STRF_w$ for $S_B(t, X)$ (Fig. 1B). More generally, this procedure can be extended by performing any higher-order alteration of interest. Thus for a neuron that has significant higher-order nonlinearities, the obtained $STRFs$ for $S_A(t, X)$ and $S_B(t, X)$ reflect differences that can be attributed directly to the specific alteration performed on the stimulus and its nonlinear interaction with the system.

2.6 Correlated Versus Uncorrelated Sounds

Little is known as to how the central auditory system of non-specialized mammals decompose and processes complex stimuli that are common in natural environments. Clearly not all natural sounds are alike, and it is of interest to understand how different classes of natural sounds are represented and processed by the central auditory system. As an example consider speech and vocalization sounds (Chapter 1: Fig. 5). Such sounds often have coherently activated spectral resonances, temporal modulations, and FM sweeps which together give rise to distinct perceptual qualities. Harmonicity and fast temporal periodicities give rise to the unified percept of pitch whereas slower temporal modulations that occur from disjoining speech segments and word transitions are perceived as discrete auditory objects or acoustic rhythms (Plomp 1967 1983). The perception of timbre (Plomp 1970; Pols, Kamp, and Plomp 1969; Van

Veen and Houtgast 1983), on the other hand, is largely dominated by spectral shape and spectral resonances that arise in speech from postural adjustments of the vocal tract and oral cavity.

In contrast, environmental noise sounds often do not share the same physical and perceptual attributes inherent to vocalizations and speech. The water sounds emanating from a small stream and the sound of ruffling leaves (Chapter 1: Fig. 6), for example, have randomly modulated spectro-temporal envelope and lack most of the distinct spectral and temporal cues that are common to vocalizations. Among these, environmental sounds often do not have strong coherent spectral resonances and temporal periodicities. Since such sounds generally do not arise from vibrating media and air columns, such as for vocal fold vibrations and the vocal tract in human speech, they therefore also lack harmonic components.

A common determinant of the perceptual and physical qualities of natural sounds are therefore determined by the level of correlation or redundancy that is present in the acoustic signal. Vocalization sounds, for example, are locally highly structured and have spectro-temporal envelopes which are highly redundant (Attias and Schreiner 1998a; Nelken, Rotman, and Yosef 1999; also see chapter 1). This is usually the result of repetitive temporal periodicities, comodulation, and spectral resonances which generally do not occur in isolation and are all the result of the constraints imposed by the voice generating mechanisms. The non-speech sound arising from shuffling leaves or running water lacks this high local correlation, likely because of the erratic patterns of air and fluid flow that give rise to such sounds. The running stream also lacks many of the complex dynamics present in the speech sound. The time-varying envelope of the

running stream preserves statistically similar acoustic properties for all times. A snapshot of the spectrogram for this sound at two distinct time instants would look largely the same. This is in marked contrast to vocalizations and speech which have local correlation properties that are continuously changing and markedly different for distinct time epochs.

Such spectro-temporal characteristics were quantified and explored in detail in Chapter 1. Here it suffices to note that these sounds represent two qualitatively different and extreme scenarios of auditory processing. With this in mind, we would like to understand how such stimulus characteristics are represented and processed by individual auditory neurons and how these are ultimately represented in the spatio-temporal neural discharge activity at various stations of the auditory system (e.g. the inferior colliculus, auditory cortex). Although we will not use natural sounds directly to achieve this (for the reasons mentioned in section 2.3) our motivation is strictly guided by neuroethologic principles. The remainder of this chapter focuses on the acoustic stimulus design. Two acoustic stimuli are designed that incorporate the following key attributes of natural sounds:

- 1) **Dynamic** – As with natural sounds, the probing stimulus should be dynamic so that it prevents response adaptation, activates dynamic nonlinearities, and so that its statistical structure changes with time. This is closely related to the notion of non-stationarity which requires that the autocorrelation function (here we consider only the spectro-temporal autocorrelation function) be time-varying (Hayes 1996; Marmarelis and Marmarelis 1978).
- 2) **Spectro-temporally complex** – It is desired that the driving stimulus be sufficiently

complex so that it embodies key spectro-temporal features of natural sounds. Some of these include FM sweeps, spectral resonances, and temporal modulations.

- 3) Globally Unbiased – In order to provide a complete and statistically sound characterization of neuronal responses the long term spectro-temporal statistics should be unbiased.
- 4) Locally correlated – As for speech and vocalizations, one sound will be designed to explore responses to sounds that are local structured and biased. The global statistics for this sound should nonetheless satisfy requirement 3.
- 5) Locally uncorrelated – This property is used to explore responses to sounds that are qualitatively similar to the babbling brook example. As for the sound used in 4, this sound is also globally unbiased.
- 6) Biologically plausible – This is our main source of motivation which is closely tied to requirements 1 and 2.
- 7) Persistently exciting – This term is often used in the engineering literature (Ljung 1987) to refer to the amount of driving force. A persistently exciting stimulus should continuously provide excitatory drive within the integration limits of the system. In neuroscience terms this requirement demands that the stimulus should also provide sufficient excitatory and inhibitory drive. Hence the stimulus should continuously probe neuronal responses up to and above the relevant neural integration limits. This is accomplished by designing acoustic stimuli that contain spectro-temporal acoustic features (onsets, offsets, resonances, FM sweeps, etc.) which continuously drive the auditory system.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

2.7 The Dynamic Ripple and Ripple Noise Stimuli

Two broadband stimuli (Fig. 2 and 3) were designed to mimic some of the spectral and temporal features characteristic of two classes of natural sounds. Although these stimuli do not capture the complete range of perceptual and acoustic properties (i.e. comodulation, harmonicity, $1/f$ modulation spectrum etc.), they nonetheless capture essential properties of their spectro-temporal envelope. Here, we sought to preserve the local correlation properties of the spectro-temporal envelope of natural sounds because these determine important perceptual qualities such as timbre (Plomp 1967; Pols, Kamp, and Plomp 1969; Plomp 1970; Van Veen and Houtgast 1983 1985).

The dynamic ripple stimulus (Fig. 2) is motivated by the ripple spectrum noise used in human psychophysics studies (Houtgast 1977) to study lateral inhibition and more recently for studying spectral and temporal receptive fields in the ferret and cat auditory cortex (Schreiner and Calhoun 1994; Kowalski, Depireux, and Shamma 1996a 1996b). The instantaneous spectrum for this sound is a sinusoidal grating on a log-frequency and log-intensity axis. It is analogous to spatial sinusoidal gratings used in visual experiments to investigate neural sensitivities (Victor and Purpura 1998; Girman, Sauve, and Lund 1999). A key characteristic of the dynamic ripple is evident upon examining its local statistics. Note that the spectro-temporal envelope is locally highly structured (having distinct temporal modulations, spectral resonances, and FM sweeps) much like the features found in vocalizations and speech. Similar to animal vocalizations and speech, this envelope shows nonrandom spectral resonances and temporal modulations at a characteristic spectral and temporal frequencies. This sound has strong short term correlations which are locally determined by its instantaneous stimulus

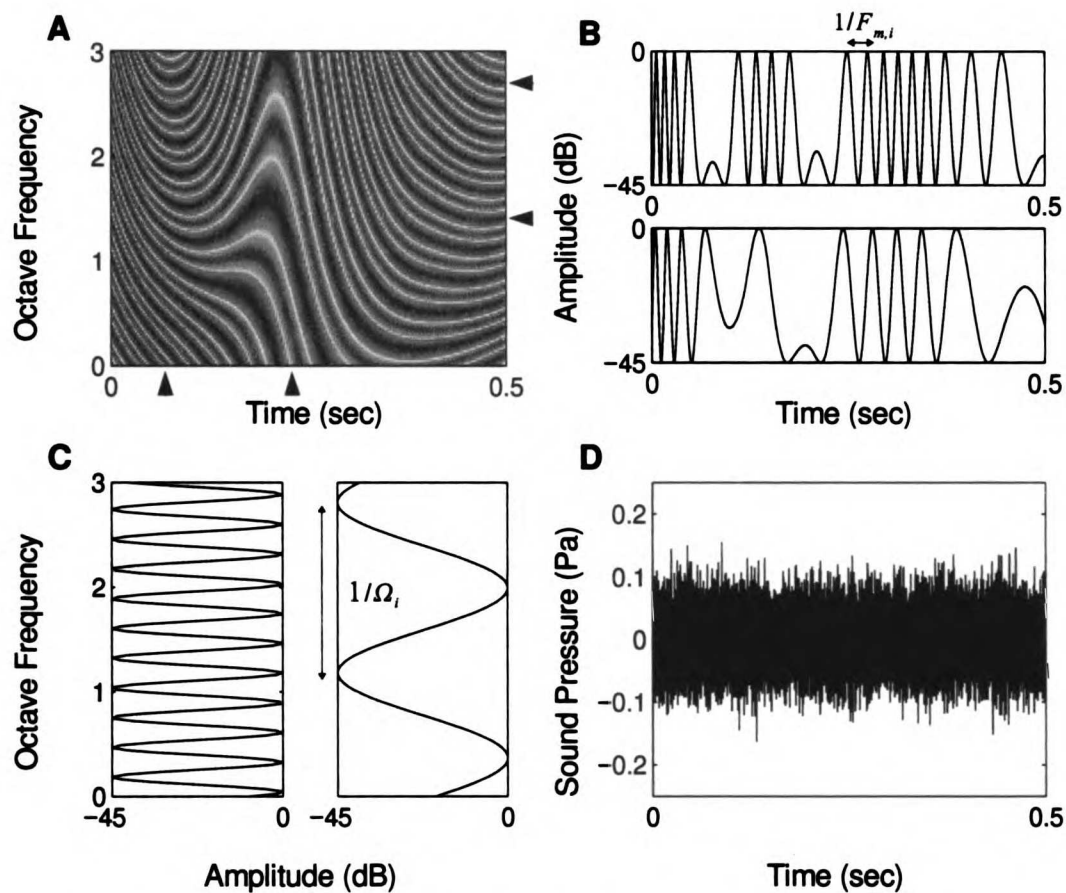


Figure 2: (A) The dynamic ripple spectro-temporal envelope has complex dynamics, spectral resonances, and temporal modulations which coexist along the spectral and temporal axis. (B) Temporal cross sections of the spectro-temporal envelope shown at locations marked by arrows. Note that the instantaneous modulation rate, $F_{m,i}$, changes dynamically with time (bandlimited to 1.5 Hz). (C) Spectral cross-section shown at locations marked by arrows. At a given time instant, the stimulus envelope has a sinusoidal shape on a logarithmic frequency – logarithmic amplitude axis where the instantaneous ripple frequency, Ω_i (bandlimit frequency 3 Hz), determines the number of resonances (cycles / octave) along the spectral axis. The acoustic pressure waveform (D) for the dynamic ripple envelope of (A) has a noisy character similar to white noise.

Shown for $\Omega_{Max} = 4$ cycles/octave and $F_{Max} = 70$ Hz.

parameters (temporal modulation rate and ripple density).

Fig. 2a shows the spectro-temporal envelope of a segment of the dynamic moving ripple stimulus. At a fixed time instant (Fig. 2c) the spectral envelope has a sinusoidal shape on a logarithmic-amplitude and logarithmic-frequency axis where the envelope frequency, $\Omega(t)$ (units of cycles per octave), varies dynamically with time. Along the temporal axis (Fig. 1b), the envelope turns on and off dynamically so that the temporal modulation rate, $F_m(t)$ (units of Hz), varies as a function of time. The dynamic ripple stimulus is of particular interest since it mimics the dynamic spectral profiles created by formants (spectral resonance of the vocal tract) in speech production and animal vocalizations.

A second stimulus was designed which has weak local correlations and therefore has complementary local statistics to the dynamic ripple envelope. Unlike the dynamic ripple, the ripple noise (Fig. 3a) envelope is locally weakly correlated (unstructured) resembling background and environmental noise like wind and rain (see chapter 1). Hence, this sound is equivalent to traditional reverse correlation stimuli. Spectral and temporal cross sections for this envelope are shown in Fig. 3b and c. Unlike the cross sections for the dynamic ripple, which have spectral and temporal oscillations at a characteristic frequency, the ripple noise cross sections are noisy and resemble a bandlimited uniformly distributed noise signal. Hence, this sound lacks the high redundancy which is present in animal vocalizations, speech (Attias and Schreiner 1998; Nelken, Rotman, and Yosef 1999; see chapter 1), and the dynamic ripple envelope. Despite these local properties the ripple noise sound probes the same range of temporal and spectral modulations as the dynamic ripple sound.

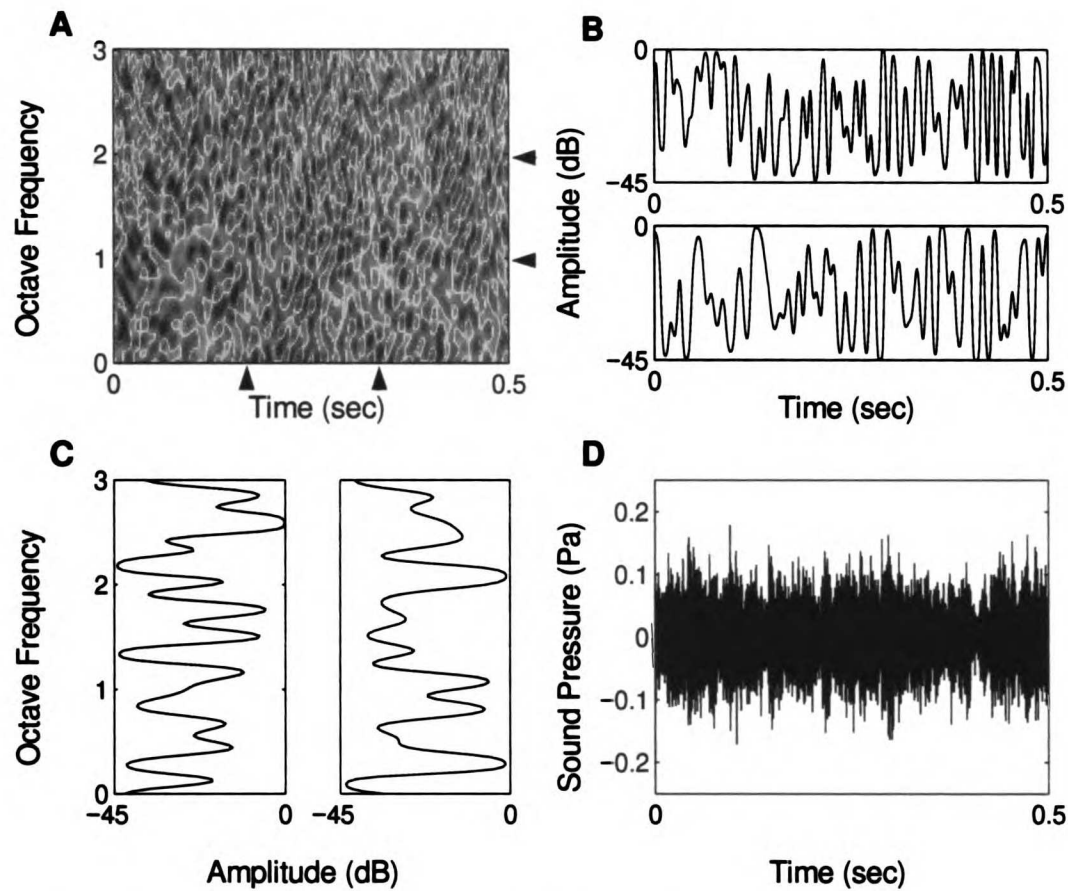


Figure 3: (A) The ripple noise spectro-temporal resembles a bandlimited spectro-temporal white noise signal with weak local correlations. Temporal (B) and spectral (C) cross-sections have a noisy character and are statistically uncorrelated. (D) Acoustic sound pressure waveform resembles a white noise signal. Shown for $\Omega_{Max}=4$ cycles/octave and $F_{Max}=70$ Hz.

In addition to preserving some spectro-temporal features that are common to **distinct** classes of natural sounds, the dynamic ripple and ripple noise stimulus are **designed** to retain the basic properties of white noise that are necessary for obtaining **reverse** correlation measurements: *i*) Flat power spectrum and impulsive autocorrelation

function, $r(\tau)$, in the vicinity of $\tau=0$, *ii*) Flat envelope spectrum and impulsive spectro-temporal envelope autocorrelation functions. We distinguish these two constraints and note that property (*i*) is imposed on the signal carriers requiring that they have a white noise character. To account for the fact that the auditory sensory epithelium is arranged logarithmically in frequency along the basilar membrane (Lieberman 1982; Greenwood 1990), the acoustic stimulus is designed so that requirement (*i*) is satisfied on a octave frequency axis. Constraint (*ii*), on the other hand, is imposed on the spectro-temporal envelope, a second-order property of the stimulus (Cohen 1995; Hermes *et al.* 1981; Klein *et al.* 2000). It is required that the stimulus envelope be globally unbiased, so that all spectral envelope and temporal modulation frequencies are equally represented within the physiologically relevant range. In addition, we also required that the stimulus be *globally* uncorrelated along these two dimensions, allowing us to perform reverse correlation measurements with respect to the stimulus spectro-temporal envelope. Despite this *global* correlation property, the dynamic ripple stimulus is *locally* correlated at any time-frequency instant (as is the case with many natural stimuli). It will be shown in subsequent sections that the local correlation structure of the dynamic ripple stimulus changes dynamically and is determined by the ripple density, $\Omega(t)$, and temporal modulation rate parameters, $F_m(t)$, at a given time instant.

2.8 Ripple Stimulus Design

The ripple noise and moving ripple stimuli are generated using a bank of $N=230$ **sinusoid** components of increasing frequency, f_k . Each sinusoid component is individually **amplitude** modulated by the linear amplitude spectro-temporal envelope $S_{Lin}(t, X_k)$

and summed to produce the time waveform $s(t)$. The noise signal is represented by

$$s(t) = \sum_{k=1}^N S_{L/n}(t, X_k) \sin(2\pi f_k t + \phi_k) \quad (2.15)$$

where ϕ_k is an uniformly distributed random phase in the interval $[0, 2\pi]$ which gives $s(t)$ noise like properties. The variable X_k represents an octave frequency axis and is related to the individual carrier frequencies, f_k , by $X_k = \log_2(f_k/f_1)$. Here $f_1=500$ Hz corresponds to the lower spectrum frequency and $f_N=20$ kHz is the maximum frequency of the ripple signal. The octave defined carrier components, X_k , are equally spaced on an octave axis and span a range of 5.32 octaves. This guarantees that the primary sensory epithelium is uniformly excited and equal energy is provided per unit octave (adhering to criterion (i)). To satisfy this property the carrier frequencies, f_k , must be geometrically spaced. The k^{th} carrier is related to adjacent frequency components by $f_k = \alpha f_{k-1}$, where $\alpha=1.01617$ is a constant strictly greater than unity, and related to the first carrier component, f_1 , by $f_k = \alpha^{k-1} f_1$. This general form for f_k results in linearly spaced octave elements where $X_k = (k-1)\Delta X$ and $\Delta X = \log_2(\alpha)$ is the spectral separation between adjacent components. For the chosen α , 43 carrier components are summed per unit octave at a spectral resolution of $\Delta X=0.0231$ octaves.

2.9 Design of the Dynamic Moving Ripple Envelope

The dynamic moving ripple envelope is an extension of the moving ripple used by Kawolski *et al.* (1996a and 1996b). The time-varying spectro-temporal envelope for this stimulus is shown in Fig. 2a. Temporal and spectral cross sections are shown in Fig. 2b and 2c. The spectral cross section, which is taken at fixed time instant, has a sinusoidal shape on an octave frequency and logarithmic amplitude axis (units of dB). Since the ripple spectrum noise excites the primary sensory epithelium with a sinusoidal energy distribution, it is therefore analogous to visual spatial gratings commonly used to study visual neurons which likewise excite the sensory epithelium in the retina with a sinusoidal energy distribution. The temporal cross section is time-varying and designed to probe different temporal periodicities.

The decibel amplitude spectro-temporal envelope is expressed as

$$S(t, X_k) = \frac{M}{2} \sin(2\pi \Omega(t) X_k + \Phi(t)) - \frac{M}{2} \quad (2.16)$$

where M is the modulation depth in units of decibels, and $\Omega(t)$ (units of cycles/octave) is the time-varying ripple density (i.e. the number of resonances per octave). The ripple phase, $\Phi(t)$, is time-varying and determines the instantaneous phase of the spectral envelope relative to the first component X_1 . This parameter additionally determines the *instantaneous modulation rate*, $F_m(t)$, (units of Hz) and the *frequency modulation sweep rate* of the spectro-temporal envelope. The parameters, $\Omega(t)$ and $F_m(t)$, vary randomly in *time* (Fig. 4), are statistically independent and unbiased within a chosen parameter range

(Fig. 5). The resulting spectral profile is therefore dynamic (as is the case with natural signals), globally spectro-temporally uncorrelated, and statistically unbiased and therefore adheres to criterion (ii). These stimulus characteristics are described in detail in sections 2.12–2.23.

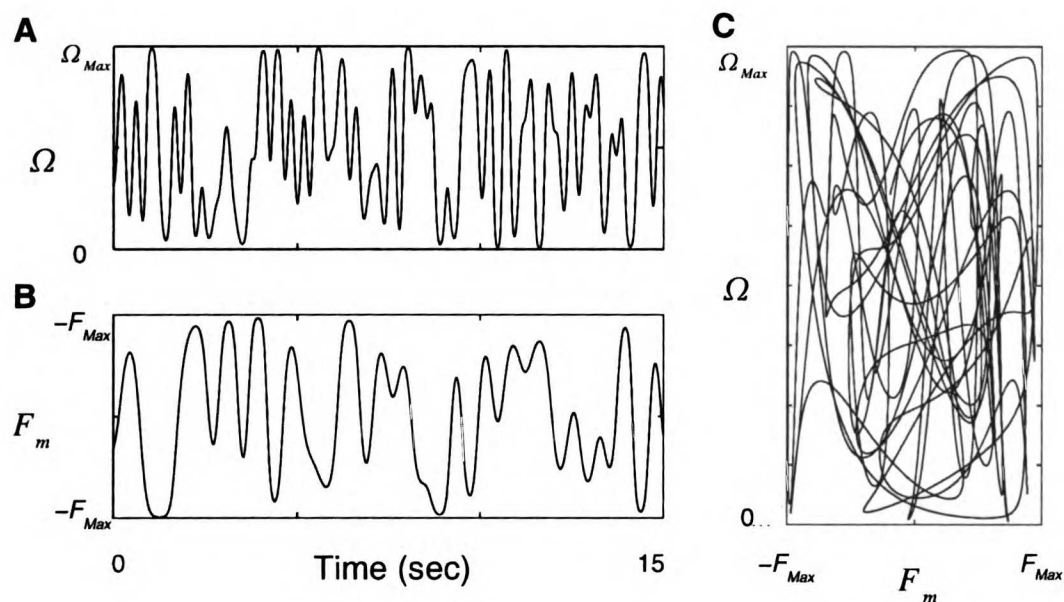


Figure 4: Time trajectory for the ripple frequency (A) and modulation rate (B) parameters vary randomly in time. Note that the parameter signals are confined to the range $[0, \Omega_{Max}]$ and $[-F_{Max}, F_{Max}]$ respectively and are bandlimited (3 Hz and 1.5 Hz respectively). The ripple density and modulation rate parameters (C) (shown as a space trajectory for the corresponding time trajectories of A and B) simultaneously probe the stimulus spectral and temporal acoustic space.

Because Eq. (2.16) is written in units of decibels (see Fig. 2) it is clear that the **moving** ripple envelope probes logarithmic amplitude variations. Most acoustic stimuli **used** in auditory experiments, however, probe spectral and temporal preferences using

linearly distributed amplitude gradations (e.g. Kowalski, Depireux, and Shamma 1996a). We choose logarithmic amplitude scale (e.g. SPL measured in decibels) for the relevant stimulus dimension since natural sounds have relative intensity gradations that cover a decibel space (Attias and Schreiner 1998a; chapter 1) and since neuronal response areas generally span several orders of magnitude (Ehret and Merzenich 1988; Eggermont 1989). Recent evidence additionally suggests that auditory (Attias and Schreiner 1998a) neurons are adapted and respond efficiently to such spectro-temporal gradations.

Although the moving ripple envelope defined in Eq. (2.16), $S(t, X_k)$, spans a decibel amplitude axis, Eq. (2.15) requires a linearly defined spectro-temporal envelope, $S_{Lin}(t, X_k)$. We must therefore transform and relate the envelope defined in Eq. (2.16), which is given in units of decibels, to a linear amplitude signal. The two signal descriptions are related by $S(t, X_k) = 20 \log_{10}[S_{Lin}(t, X_k)]$ where the reference amplitude used to define the decibel quantity is unity. Hence we only consider amplitude variations relative to a maximum amplitude of unity. During experiment sessions this unity reference point is chosen as the stimulus maximum sound pressure level, SPL_{Max} . Note that the minimum and maximum relative decibel intensities of the spectro-temporal envelope of Eq. (2.16) are $-M$ and 0. In absolute units these will be $SPL_{Max} - M$ and SPL_{Max} . The linear spectro-temporal envelope is therefore bounded between $10^{-M/20}$ (near zero) and unity satisfying the general conventions used to define amplitude modulation signals which limit the maximum and minimum signal amplitudes to range 0 and 1 (Cohen 1995). The linear spectro-temporal envelope used in Eq. (2.15) is obtained from Eq. (2.16). Taking the inverse logarithm of $S(t, X_k)$ results in the desired quantity:

$$S_{Lin}(t, X_k) = 10^{S(t, X_k)/20} = 10^{\frac{M}{40} \sin(2\pi \Omega(t) X_k + \Phi(t)) - \frac{M}{40}} \quad (2.17)$$

2.10 Parameter Design for the Dynamic Ripple

Since the spectro-temporal envelope varies along the temporal and spectral dimensions of the stimulus, we would like designate a parametric description of the stimulus a priori. Secondly we would like to derive the relationships of the time-varying parameters $\Omega(t)$, $F_m(t)$, and $\Phi(t)$ from first principles. Conceptually the dynamic moving ripple envelope corresponds to a moving wavefront, along the spectral axis X_k , with time-varying velocity and wavelengths. The rate of change of the spectral envelope,

$\Omega(t)$, is obtained by differentiating the argument of Eq. (2.16) with respect to X_k and dividing by 2π (Cohen 1995). Doing so it is easy to verify that

$$\Omega(t) = \frac{1}{2\pi} \frac{d}{dX_k} (2\pi \Omega(t) X_k + \Phi(t)) \quad (2.18)$$

The instantaneous ripple density, $\Omega(t)$, therefore determines the number of sinusoidal peaks per octave along the spectral axis, X_k , which exist at a given time instant. Since this parameter is time-varying, it allows one to dynamically probe neuronal responses to numerous spectral resolution.

Having derived the spectral properties of the envelope it is of equal interest to determine its temporal properties. The instantaneous rates of change of the temporal

envelope is similarly obtained by differentiating the argument of Eq. (2.16) now with respect to time and dividing by 2π . The instantaneous temporal modulation rate, $F_m(t)$, is therefore given by

$$F_m(t) = \frac{1}{2\pi} \frac{d}{dt} (2\pi \Omega(t) X_k + \Phi(t)) = \Omega'(t) X_k + \frac{1}{2\pi} \Phi'(t) \quad (2.19)$$

Ideally, the two parameters $\Omega(t)$ and $F_m(t)$ are chosen a priori so that the spectral and temporal properties of the envelope are statistically independent and unbiased. Clearly $F_m(t)$ is not statistically independent of $\Omega(t)$ since from it is a function of $\Omega'(t) X_k$ (Eq. (2.19)). To overcome this, we allow

$$\frac{1}{2\pi} \Phi'(t) \gg \Omega'(t) X_k, \quad (2.20)$$

so that $F_m(t)$ has its largest contribution from $\Phi(t)$. The ripple phase is obtained by solving Eq. (2.19) for $\Phi(t)$ and allowing $\Omega'(t) X_k \rightarrow 0$. Using this approximation, the instantaneous phase signal is designated as

$$\Phi(t) = 2\pi \int_0^t F_m^d(\tau) d\tau \quad (2.21)$$

where $F_m^d(t)$ is the desired temporal modulation rate profile. Hence as previously

mentioned the phase signal has a dual role. First it controls the temporal modulation rate of the dynamic ripple envelope by Eq. (2.19). Secondly, it serves to randomize the relative phase of the spectral envelope. Although the absolute phase is given by Eq. (2.21), the distribution of phases is best expressed as a relative quantity since the position of the spectral envelope is circularly symmetric (repeats every multiple integer of 2π). The phase distribution is obtained by considering the modulus phase,

$$\Phi_m(t) = \text{mod}(\Phi(t), 2\pi), \text{ where } \text{mod}(\cdot, a) \text{ designates the modulus operator base } a.$$

Using this quantity the phase signal, $\Phi_m(t)$, is confined to the interval $[0, 2\pi]$ and the ripple phase probability distribution is expressed as $p(\Phi_m)$. A segment of the phase trajectory and its distribution are shown in Fig. 5. Note that as for the ripple density and the temporal modulation rate parameters, the ripple phase is likewise uniformly distributed (statistically unbiased) so that it probes all spectral phases.

The actual modulation rate profile for the dynamic ripple is expressed as

$$F_m(t) = F_m^d(t) + \Omega'(t)X_k \quad (2.22)$$

where the error or bias between the actual and desired modulation profiles is

$$\Delta F_m(t) = F_m(t) - F_m^d(t) = \Omega'(t)X_k \quad (2.23)$$

Note that $\Delta F_m(t)$ is a function of the spectral location, X_k . This bias can be minimized

to any degree of accuracy by simply reducing the bandwidth of the signal $\Omega(t)$, thereby limiting its rate of change. If $\Omega(t)$ is chosen as a constant for all time (bandwidth of zero) then $\Omega'(t)=0$ and $F_m(t)=F_m^d(t)$. Although we can in theory minimize the stimulus bias in this manner, this is not strictly desired since choosing a signal for $\Omega(t)$ that is too slow will compromise the stimulus dynamics. A tradeoff must therefore be established between making the envelope dynamic, so that it preserves properties of natural sounds, and making the stimulus unbiased over the chosen range of temporal preferences.

As a basis for preserving stimulus dynamics in the range of natural sounds, the parameter bandwidths are chosen so that they overlap the word rates, syllable rates, and stressed syllable rates of speech, all of which fall in the range of 1–8 Hz (Plomp 1983; Greenberg 1998). The parameters $F_m(t)$ and $\Omega(t)$ vary randomly and independently, where $F_m(t)$ takes uniformly distributed values in the interval $[-350,350]$ Hz (negative modulation rates indicate that the ripples move from low to high frequencies producing upward FM sweeps) and $\Omega(t)$ takes uniformly distributed values in the interval $[0,4]$

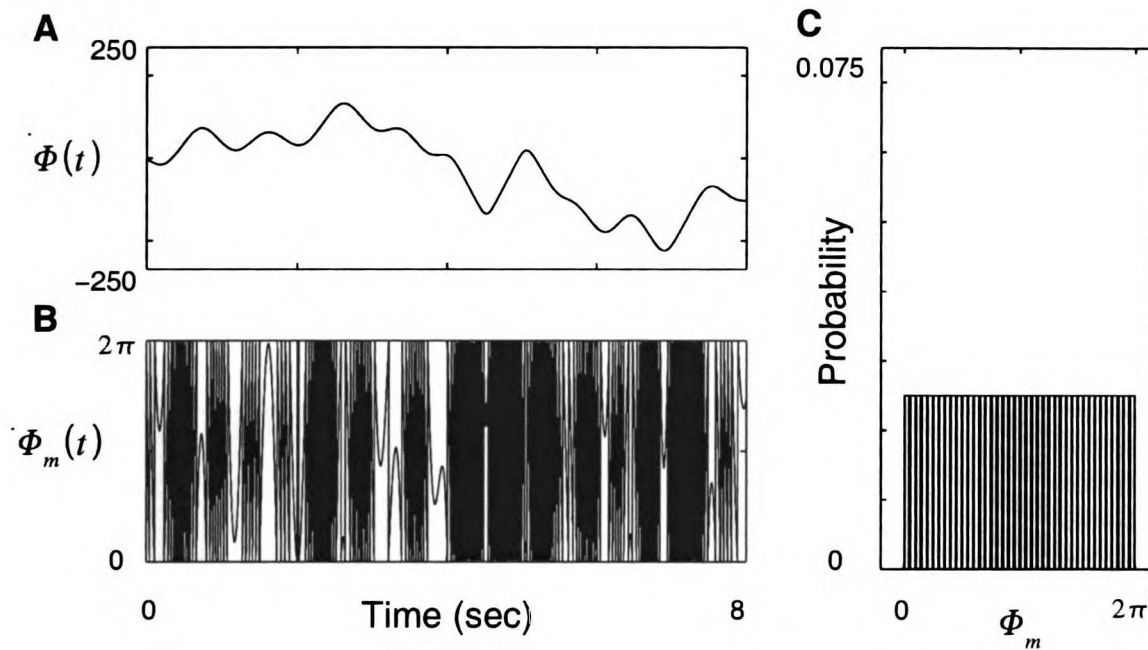


Figure 5: Time trajectories for the ripple phase, $\phi(t)$, (A) and the relative modulus phase, $\phi_m = \text{mod}(\phi(t), 2\pi)$, (B). As for the ripple density and modulation rate parameters (Fig. 4 and 6), the relative ripple phase, ϕ_m , follows an uniform distribution and is therefore statistically unbiased.

cycles per octave. To optimally excite auditory neurons in the range characteristic for speech we designed these parameters so that they continuously vary in time at a nominal rate of 1.5 Hz and 3 Hz, respectively. Note that the stimulus spectro-temporal correlation function likewise varies at this rate and the sound is therefore non-stationary (Hayes 1996; Marmarelis and Marmarelis 1978). Fig. 4 shows the parameter signal trajectories for a short time segment. The corresponding parameter distributions and their power spectra are shown in Fig. 6. Using a bandwidth of 1.5 Hz for $F_m(t)$ and 3 Hz for $\Omega(t)$, satisfies our dual motivation to preserve the stimulus dynamics in the range for speech

while minimizing the parameter error for $F_m(t)$. For these values, Eq. (2.20) is approximately satisfied and the parameter $F_m(t)$ has a mean RMS error of 5 %.

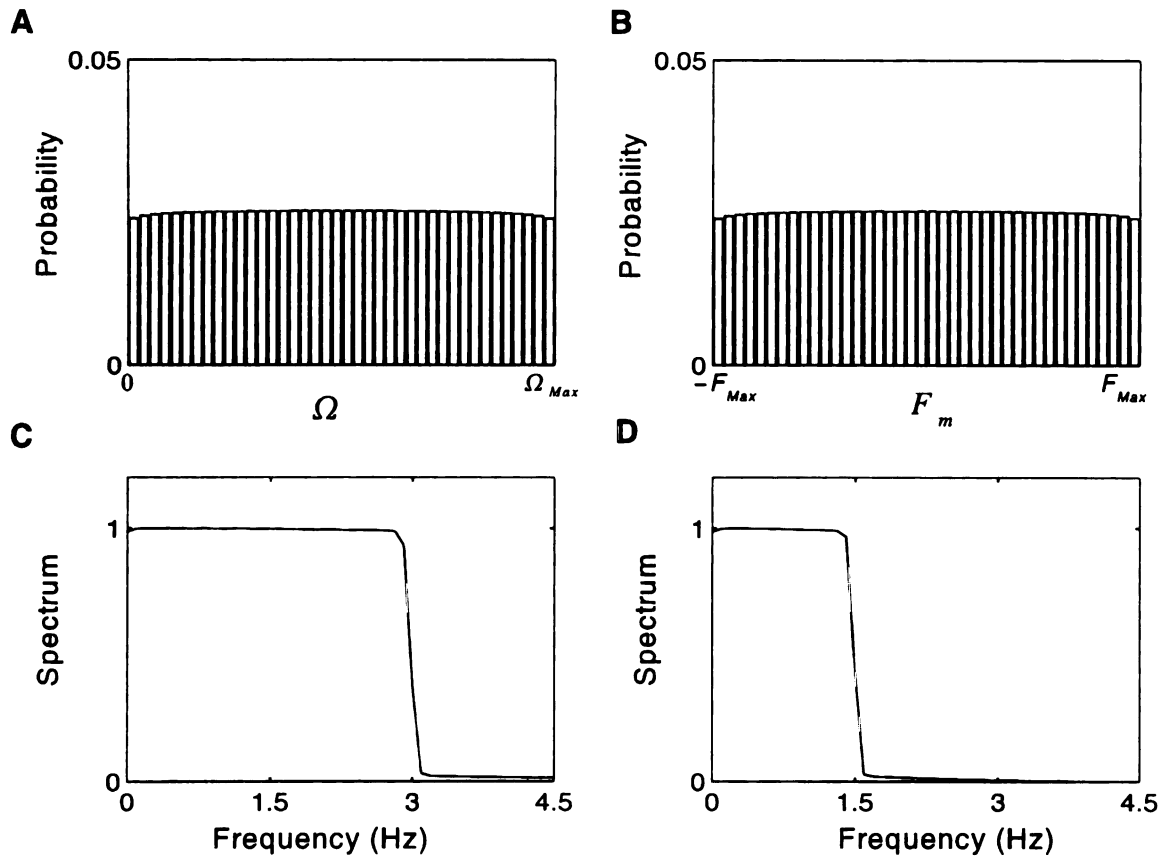


Figure 6: The parameter distributions for the ripple density (A) and modulation rate (B) parameters shown in Fig. 4. Both parameters are uniformly distributed and statistically unbiased over their defined range. Power spectrum of the ripple density parameter (C),

$\Omega(t)$, and modulation rate parameter (D), $F_m(t)$, signals. The time-varying ripple density parameter has a cutoff frequency of 3 Hz while the modulation rate parameter has a maximum turnover rate of 1.5 Hz.

2.11 Design of the Ripple Noise Envelope

The ripple noise spectro-temporal envelope is generated as a superposition of $L=16$ independent dynamic ripple envelopes. Each dynamic ripple is constructed from L statistically independent ripple density, $\Omega_{(k)}(t)$, and modulation rate, $\Phi_{(k)}(t)$, parameter signals. For any integer valued $k \neq l$, it is therefore required that the parameter autocorrelation functions satisfy $E[(\Omega_{(k)}(t) - \mu_{\Omega})(\Omega_{(l)}(t + \tau) - \mu_{\Omega})] = 0$ and $E[F_{m(k)}(t)F_{m(l)}(t + \tau)] = 0$ for all τ . Here $E[\cdot] = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \cdot dt$ is the expectation or time average operator. Consequently, the k^{th} dynamic ripple envelope is constructed so that it is statistically independent of the l^{th} dynamic ripple envelope (i.e.

$E[\bar{S}_{(k)}(t, X)\bar{S}_{(l)}(t + \tau, X + \zeta)] = 0$ for all τ and ζ where $E[\cdot]$ is now a spectro-temporal average). The signal

$$\bar{S}_k(t, X) = \frac{M}{2} \sin(2\pi \Omega_{(k)}(t)X + \Phi_{(k)}(t)) \quad (2.24)$$

is the zero-mean dynamic ripple envelope for the k^{th} ripple component.

Formally the ripple noise envelope is expressed as

$$Y(t, X_k) = f \left[\frac{1}{\sqrt{L}} \sum_{l=1}^L \bar{S}_l(t, X_k) \right] \quad (2.25)$$

where

$$f(x) = \frac{M}{2} \operatorname{erf}\left[\frac{x}{\sigma_{DR}}\right] - \frac{M}{2} \quad (2.26)$$

is a contrast transformation, $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\tau^2} d\tau$ is the error function, and

$\sigma_{DR} = M/\sqrt{8}$ is the dynamic ripple standard deviation. The function $f(x)$ transforms the amplitude distribution of the ripple noise from a normal distribution with standard deviation σ_{DR} to an uniform amplitude distribution in the interval $[-M, 0]$ (Fig. 7).

This transformation is performed for two reasons: first $f(x)$ matches the range of stimulus intensities for the ripple noise signal to those of the dynamic ripple envelope so that the two signals have almost identical amplitude distributions (Fig. 7). Both stimuli therefore probe the same amplitude operating range and have identical contrast statistics. Secondly, and more importantly, there is a potential problem that arise when characterizing neuronal responses, since central auditory neurons can have a strong non-linear dependency with the stimulus intensity (Ehret and Merzenich 1988; Eggermont 1989). Since the goal of this study is to characterize spectro-temporal nonlinearities at a fixed operating point, this normalization helps prevent simultaneous activation of other nonlinearities which arise from independent response components (i.e. intensity, contrast, etc.). Without this transformation, the long tails of the normally distributed ripple noise stimulus could possibly excite intensity nonlinearities that would not be excited by the

dynamic ripple stimulus.

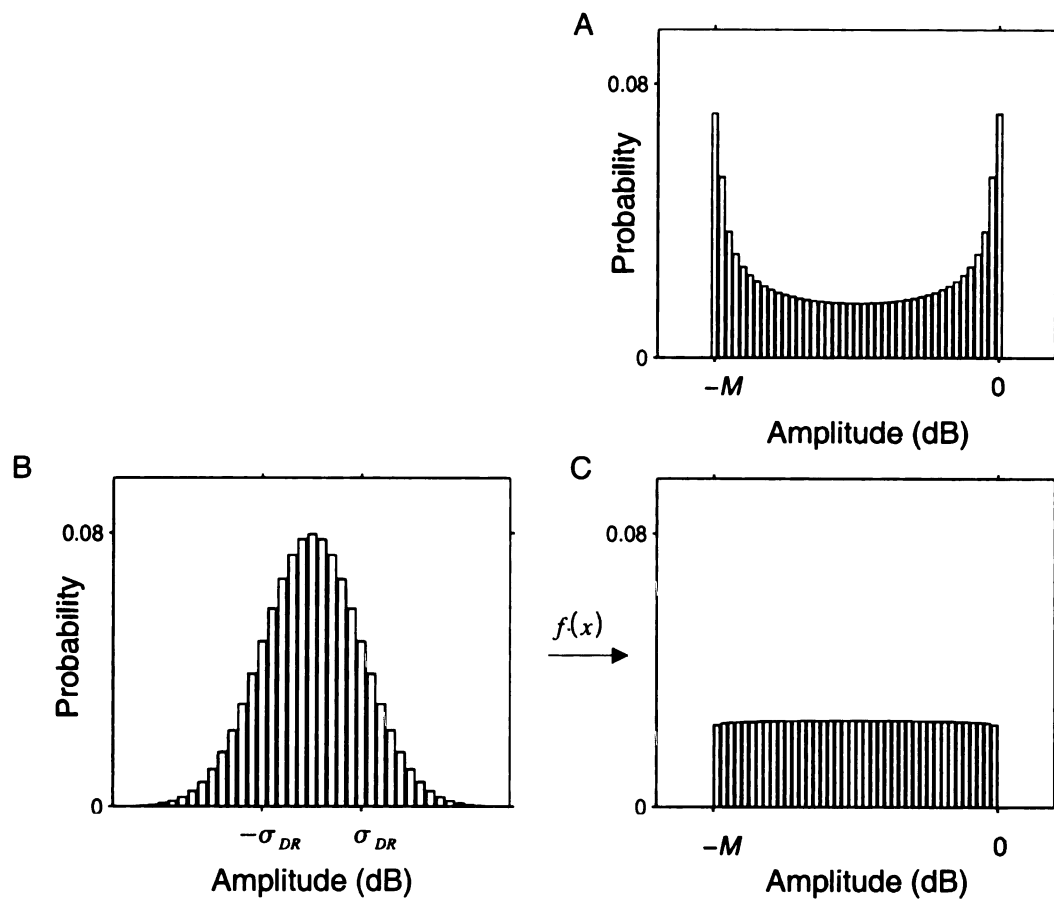


Figure 7: Amplitude distributions for the dynamic ripple envelope (A), uncompressed ripple noise (B), and for the compressed ripple noise envelope (C). After compression, the dynamic ripple and ripple noise cover identical ranges $[-M, 0]$. These distributions differ only at the extremities 0 and $-M$. Without compression, the ripple noise envelope (B) has a normally distributed amplitude with standard deviation

$$\sigma_{DR} = M/\sqrt{8} .$$

In the limiting case where $L \rightarrow \infty$ the ripple noise spectro-temporal envelope

approaches a band-limited spectro-temporal white noise envelope. Since the component dynamic ripple envelope components, $S_l(t, X_k)$, and their corresponding parameters, $\Omega_{(l)}(t)$ and $F_{m(l)}(t)$ are statistically independent stochastic processes, the central limit theorem guarantees that the sum inside $f(x)$ converges to a normal distribution of variance σ_{DR}^2 . The transformation, $f(x)$, serves only to alter the amplitude distribution and does not alter the spectro-temporal content. In a subsequent section it is shown that this transformation does not alter the shape of the stimulus autocorrelation function. Hence, similar to the dynamic ripple stimulus, the ripple noise stimulus is in principle well suited for reverse correlation procedures.

2.12 Dynamic Ripple and Ripple Noise Spectro-Temporal Correlation Statistics

By design, the dynamic ripple and ripple noise stimuli have a general appeal for studying nonlinear auditory processing and for studying central auditory representations. In particular, their suitability for reverse correlation combined with the numerous ethologic considerations (i.e. spectro-temporal characteristics, temporal and spectral envelope frequency ranges, stimulus dynamics, and logarithmic contrast) make them ideal for studying various aspects of central auditory processing. Of interest is the fact that by using such sounds one can study processing of spectral and temporal stimulus features simultaneously. In the remainder of this section we focus on thoroughly characterizing the local correlation properties of the spectro-temporal envelope as well as the dynamic properties of the two stimuli.

Although the stimulus design, chosen set of parameters, and stimulus dynamics for these two stimuli are ethologically motivated, it needs to be determined whether the resulting spectro-temporal statistics (of the resulting spectro-temporal envelopes) are biologically reasonable and well suited for reverse correlation. These concepts are highlighted in the following section. It is shown that both stimuli have identical second-order statistics and differ only in their higher-order moments and stimulus dynamics. Since the identification of linear system is dependent only on the second-order stimulus features and independent of the stimulus dynamics, we would expect identical systems characterizations for both sounds. This argument was proved analytically in section 2.4. Hence we can use this knowledge about the stimulus to learn about nonlinear auditory processing and response dynamics to such sounds.

2.13 Dynamic Ripple Local Approximation

The dynamic ripple spectro-temporal envelope, Eq. (2.16), can be expressed as

$$S(t, X) = \frac{M}{2} \sin(\text{Arg}(t)) - \frac{M}{2} \quad (2.27)$$

where the argument inside the sine function is

$$\text{Arg}(t) = 2\pi \Omega(t)X + \Phi(t) \quad (2.28)$$

Note that this argument is time-varying as a consequence of the slowly varying stimulus

parameters which control the spectral and temporal envelope properties. Since the temporal modulations of the dynamic ripple envelope are exceedingly fast (up to 350 Hz) in comparison to the stimulus parameters time rate of change (1.5 Hz for $F_m(t)$ and 3 Hz for $\Omega(t)$), one can easily model a local fit to the spectro-temporal envelope. To do so $Arg(t)$ is locally expanded using a Taylor series approximation about $t=t_i$

$$Arg(t)|_{t=t_i} = \Phi(t_i) + 2\pi\Omega(t_i)X + \Phi'(t_i)(t-t_i) + 2\pi\Omega'(t_i)X(t-t_i) + \dots \quad (2.29)$$

Since the parameters $\Omega(t)$ and $F_m(t)$ are slowly varying in time, the higher-order terms in the expansion will tend to be small in the vicinity of $t=t_i$. One can therefore ignore all terms in the expansion of higher than 1st order. An additional simplifying assumption can be made by noting that $F_m(t_i)t \gg \Omega'(t_i)Xt = \Delta F_m(t_i)t$ (Eq. (2.20) and (2.23)). Hence the first-order term in the Taylor expansion containing $\Omega'(t_i)Xt$ is likewise ignored. After rearranging terms and simplifying, the inner argument of the dynamic ripple profile is approximated by

$$Arg(t)|_{t=t_i} \approx 2\pi\Omega(t_i)X + \Phi'(t_i)t + \Phi(t_i) - \Phi'(t_i)t_i - 2\pi\Omega'(t_i)Xt_i \quad (2.30)$$

Noting that $\Phi'(t_i) = 2\pi F_m(t_i)$ and that $\Phi(t_i) - \Phi'(t_i)t_i - 2\pi\Omega'(t_i)Xt_i$ are constant terms, $Arg(t)$ is then expressed in the general form

$$\text{Arg}(t)|_{t=t_i} \approx 2\pi\Omega_i X + 2\pi F_{m,i} t + \Phi_i \quad (2.31)$$

where $\Phi_i = \Phi(t_i) - \Phi'(t_i)t_i - 2\pi\Omega'(t_i)Xt_i$, $\Omega_i = \Omega(t_i)$, and $F_{m,i} = F_m(t_i) = \frac{1}{2\pi}\Phi'(t_i)$ are the instantaneous stimulus parameters. The dynamic ripple spectro-temporal envelope,

$S(t, X)$, is locally approximated by

$$S(t, X)|_{t=t_i} \approx S(t, X|t_i) \quad (2.32)$$

where

$$S(t, X|t_i) = \frac{M}{2} \sin(2\pi\Omega_i X + 2\pi F_{m,i} t + \Phi_i) - \frac{M}{2} \quad (2.33)$$

is a static moving ripple envelope of constant ripple density, Ω_i , temporal modulation rate, $F_{m,i}$, and spectral phase, Φ_i .

The dynamic ripple envelope is therefore a generalization of the static moving ripple gratings used by Kowalski *et al.* (1996a 1996b). These stimuli are of interest since they form a joint basis set for spectral and temporal acoustic stimulus features. Although individual ripple gratings of the static moving ripple stimulus can move, following an upward or downward trajectory, they are nonetheless referred to as static since the

stimulus parameters are constant throughout the stimulus presentation. The dynamic ripple differs principally in that the parameters are chosen as time-varying stochastic processes. The resulting spectro-temporal grating is consequently dynamic, having structural components and combinations which are not probed by the static moving ripple gratings. This is especially true at points in time where the stimulus parameters are rapidly changing and, consequently, the spectro-temporal content transitions from one parameter regime to another in a continuous manner. Given that the dynamic ripple is locally well approximated by a static moving ripple of constant parameters, it can therefore be thought of as a local basis set decomposition which spans the ripple density and temporal modulation rate parameter space in a dynamic fashion.

2.14 Dynamic Ripple Local Autocorrelation Function

As demonstrated in Fig. 2a and Fig. 4, the dynamic ripple envelope is locally structured and has complex spectro-temporal dynamics that are determined by the ripple density and temporal modulation rate parameters. To further understand such characteristics of the dynamic ripple envelope which make it ideal for characterizing auditory neurons, its dynamic auto-correlation properties are further investigated (Cohen 1995). This theoretical framework will serve as a foundation for determining its suitability for general systems identification and for studying auditory system function with this sound.

Given that the dynamic ripple envelope is instantaneously approximated by Eq. (2.33), one can approximate the instantaneous spectro-temporal autocorrelation (Cohen 1995) function about $t=t_i$ by

$$R_{ss}(\tau, \zeta | t_i) = E \left[\bar{S}(t+\tau, X+\zeta | t_i) w_i(t+\tau, X+\zeta) \bar{S}(t, X | t_i) w_i(t, X) \right] \quad (2.34)$$

where the spectro-temporal average operator, $E[\cdot] = \lim_{T, Z \rightarrow \infty} \frac{1}{4TZ} \int_{-T}^T \int_{-Z}^Z \cdot dt dX$, is taken

with respect to t and X . The signal

$$\bar{S}(t, X | t_i) = S(t, X | t_i) + M/2 = \frac{M}{2} \sin(2\pi \Omega_i X + 2\pi F_{m,i} t + \Phi_i) \quad (2.35)$$

is the instantaneous zero-mean spectro-temporal envelope approximation, and

$w_i(t, X)$ is an unity energy two dimensional real valued window function centered about $t=t_i$ and $X=X_i$. This spectro-temporal window serves a similar purpose as for a spectrogram representation (Cohen 1995), where the local signal is restricted in time using a tapered window function. For practical reasons we will consider 2-D Gaussian window of the general form (Cohen 1995)

$$w_i(t, X) = \frac{1}{\sqrt{\pi \sigma_t \sigma_x}} \exp \left[-(t-t_i)^2 / 2\sigma_t^2 - (X-X_i)^2 / 2\sigma_x^2 \right], \quad (2.36)$$

although other 2-D rational window functions can be used. The variables σ_t and σ_x correspond to the standard deviations of the window along the temporal and spectral dimensions respectively. This window is chosen primarily because it facilitates much of

the analysis that will follow. Since the chosen window has a finite energy of unity (i.e.

$$\lim_{T,Z \rightarrow \infty} \int_{-T}^T \int_{-Z}^Z w_i^2(t,X) dt dX = 1) \text{ the two dimensional expectation operator can be}$$

expressed as $E[\cdot] = \lim_{T,Z \rightarrow \infty} \int_{-T}^T \int_{-Z}^Z \cdot dt dX$. Throughout we primarily consider the

autocorrelation function (Eq. (2.34)), which in all instances has a squared window term

embedded inside the expectation (e.g. $w_i^2(t, X)$). Hence the energy normalized

expectation will ease the general interpretations because the resultant window

components will have maximum amplitude of unity.

To facilitate the subsequent derivations an analytic signal representation (Cohen 1995) of the spectro-temporal envelope, Eq. (2.35), is used. The spectro-temporal envelope is expressed as

$$\bar{S}(t,X|t_i) = \text{Im} [A_i(t,X)] \quad (2.37)$$

where

$$A_i(t,X) = \frac{M}{2} \exp \left[j \left(2\pi \Omega_i X + 2\pi F_{m,i} t + \Phi_i \right) \right] \quad (2.38)$$

is the analytic signal version of $\bar{S}(t,X|t_i)$. Note that $\bar{S}(t, X|t_i)$ can be recovered

directly from $A_i(t,X)$ via Eq. (2.37) since $z = e^{j\vartheta} = \cos(\vartheta) + j \sin(\vartheta)$ and

$\text{Im}[z] = \sin(\vartheta)$ for any complex variable or function ϑ . Using this analytic signal

representation, the spectro-temporal autocorrelation function is given by

$$R_{,,}(\tau, \zeta | t_i) = E \left[\text{Im} \left[A_i(t+\tau, X+\zeta) \right] \text{Im} \left[A_i(t, X) \right] w_i(t+\tau, X+\zeta) w_i(t, X) \right] \quad (2.39)$$

Expanding Eq. (2.39) using the identity

$$\text{Im} \left[z_1 \right] \text{Im} \left[z_2 \right] = \frac{1}{2} \text{Re} \left[z_1 z_2^* \right] - \frac{1}{2} \text{Re} \left[z_1^* z_2 \right] \quad (2.40)$$

where z_1 and z_2 are imaginary numbers results in

$$R_{,,}(\tau, \zeta | t_i) = \frac{1}{2} E \left[\text{Re} \left[A_i(t+\tau, X+\zeta) A_i(t, X)^* \right] w_i(t+\tau, X+\zeta) w_i(t, X) \right] - \quad (2.41)$$

$$\frac{1}{2} E \left[\text{Re} \left[A_i(t+\tau, X+\zeta) A_i(t, X) \right] w_i(t+\tau, X+\zeta) w_i(t, X) \right] .$$

Before proceeding it is necessary to expand and simplify the terms

$$\text{Re} \left[A_i(t+\tau, X+\zeta) A_i(t, X)^* \right] \quad \text{and} \quad \text{Re} \left[A_i(t+\tau, X+\zeta) A_i(t, X) \right] \quad \text{inside Eq. (2.41).}$$

Substituting Eq. (2.38) and combining terms gives

$$A_i(t+\tau, X+\zeta) A_i(t, X)^* = \frac{M^2}{4} \exp \left[j \left(2\pi \Omega_i \zeta + 2\pi F_{m_i} \tau \right) \right] \quad (2.42)$$

and

$$A_i(t+\tau, X+\zeta)A_i(t, X) = \frac{M^2}{4} \exp\left[j\left(4\pi\Omega_i X + 4\pi F_{m,i} t + 2\pi\Omega_i \zeta + 2\pi F_{m,i} \tau + 2\Phi_i\right)\right] . \quad (2.43)$$

Taking the real part of Eqs. (2.42) and (2.43) results in

$$\operatorname{Re}\left[A_i(t+\tau, X+\zeta)A_i(t, X)^*\right] = \frac{M^2}{4} \cos\left(2\pi\Omega_i \zeta + 2\pi F_{m,i} \tau\right) \quad (2.44)$$

and

$$\operatorname{Re}\left[A_i(t+\tau, X+\zeta)A_i(t, X)\right] = \frac{M^2}{4} \cos\left(4\pi\Omega_i X + 4\pi F_{m,i} t + 2\pi\Omega_i \zeta + 2\pi F_{m,i} \tau + 2\Phi_i\right) . \quad (2.45)$$

These identities can now be substituted into Eq. (2.41).

Finally the instantaneous spectro-temporal autocorrelation function is expressed

as

$$R_{ss}(\tau, \zeta | t_i) = \frac{M^2}{8} E \left[\cos(2\pi \Omega_i \zeta + 2\pi F_{m,i} \tau) w_i(t+\tau, X+\zeta) w_i(t, X) \right] - \quad (2.46)$$

$$\frac{M^2}{8} E \left[\cos(2\pi [2\Omega_i X + 2F_{m,i} t + \Omega_i \zeta + F_{m,i} \tau] + 2\Phi_i) w_i(t+\tau, X+\zeta) w_i(t, X) \right]$$

where the second term in the sum cancels because it has a mean value of zero (i.e. the mean of $\cos(\cdot)$ multiplied by a positive valued rational window of the form of Eq. (2.36) is approximately zero). Note that the first term in the sum does not cancel since the arguments of the cosine term, τ and ζ , are independent of the integration variables, t and X . After simplifying and combining all terms the dynamic ripple instantaneous autocorrelation function is expressed as

$$R_{ss}(\tau, \zeta | t_i) = \frac{M^2}{8} \cos(2\pi \Omega_i \zeta + 2\pi F_{m,i} \tau) R_{ww}(\tau, \zeta) \quad (2.47)$$

where

$$R_{ww}(\tau, \zeta) = E \left[w_i(t+\tau, X+\zeta) w_i(t, X) \right] = \exp \left[- \left(\frac{\tau^2}{4\sigma_t^2} + \frac{\zeta^2}{4\sigma_x^2} \right) \right] \quad (2.48)$$

is the 2-D autocorrelation of the Gaussian window (Cohen 1995) of Eq. (2.36). Note that Eq. (2.48) is exactly a Gaussian window of unity amplitude (since the window energy

was constrained as $R_{ww}(0,0) = \int \int w_i(t, X)^2 dt dX = 1$) and standard deviations $\sqrt{2}\sigma_t$ and $\sqrt{2}\sigma_x$.

Examples of the instantaneous autocorrelation function are shown in Fig. 8. As for the dynamic ripple spectro-temporal envelope, the spectro-temporal autocorrelation function is time-varying. Note that the instantaneous correlation distance (i.e. the distance between peaks in the autocorrelation function) is inversely proportional to the instantaneous spectro-temporal parameters, Ω_i and $F_{m,i}$, which are themselves time-varying. This fundamental property of the dynamic ripple signal is captured by Eq. (2.47) since the instantaneous correlation function is determined by the instantaneous stimulus parameters. By design, the dynamic ripple signal therefore captures an essential property which is common to natural signals over short time scales: non-stationarity. By definition non-stationarity entails that the autocorrelation function is time dependent (Hayes 1996; Marmarelis and Marmarelis 1978) as is the case for the dynamic ripple envelope. Natural sounds, such as speech and animal vocalizations, are clearly time-varying since the instantaneous properties of the signal (e.g., formant locations and temporal modulations) vary from one time instant to another.

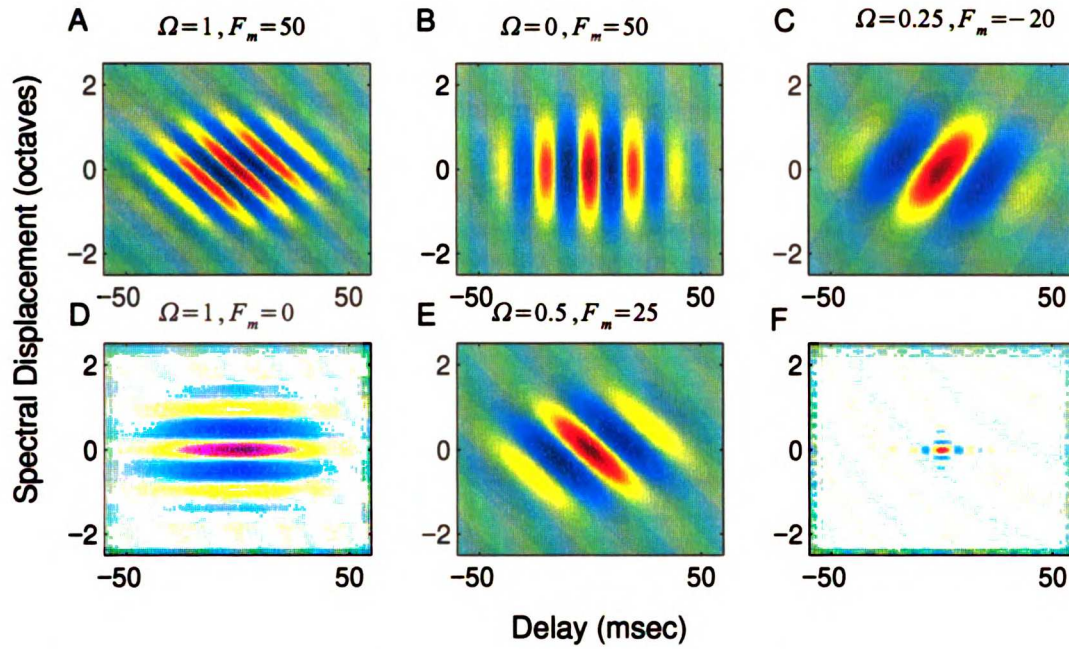


Figure 8: (A–E) Examples of the instantaneous spectro–temporal autocorrelation function for the dynamic ripple envelope. Shown for various parameter combinations at distinct time instants. (F) Global autocorrelation for the dynamic ripple envelope.

Shown for $F_{Max}=100$ Hz and $\Omega_{Max}=4$ cycles/octave.

2.15 Dynamic Ripple Global Autocorrelation Function

From the analysis of section 2.14, it is clear that the instantaneous spectro–temporal envelope of the dynamic ripple is locally biased since its local spectro–temporal autocorrelation function oscillates at a characteristic spectral and temporal frequency as shown in Fig. 8. A key characteristic of this signal, however, is that its parameters are continuously changing and, consequently, its local correlation changes in a time dependent manner. For example, at one time instant the dynamic ripple signal is largely determined by its two parameters, say $\Omega=2$ and $F_m=125$, whereas at a later time

(e.g. fractions of a second) its parameters have changed to say $\Omega=0.5$ and

$F_m = -250$. Given that these local stimulus characteristics are continuously changing, it is of interest to understand what happens in the long run. Thus, we need to examine the global statistics of the dynamic ripple spectro-temporal envelope. This characterization is crucial to determine the suitability of the dynamic ripple signal for use with reverse correlation procedure. The necessary constraints and stimulus statistics which make the dynamic ripple theoretically sound and statistically unbiased for such procedures are outlined here.

The global autocorrelation function of the dynamic ripple is obtained by averaging its instantaneous correlation function over all time. Formally this is expressed as

$$R(\tau, \zeta) = E[R_{ss}(\tau, \zeta | t_i)] = \frac{M^2}{8} E[\cos(2\pi\Omega_i + 2\pi F_{m,i}\tau)] R_{ww}(\tau, \zeta) \quad (2.49)$$

where the time average expectation, $E[\cdot]$, is taken with respect to the instantaneous

time, t_i . Since the only parameters that depend on t_i are $\Omega_i = \Omega(t_i)$,

$F_{m,i} = F_m(t_i)$, and Φ_i , this is equivalent to performing an ensemble average

(denoted by $\langle \cdot \rangle$) over the stimulus parameter space. This transformation can be

performed since the parameters signals are defined by a stationary and ergodic stochastic process.

This observations seems to be at odds with the described non-stationarity

properties of the dynamic ripple envelope (section 2.14). The apparent conflict is resolved by noting that stationarity is a time–scale dependent property of a signal. As mentioned in the previous section, the spectro–temporal envelope parameters are defined by an uniformly distributed band–limited stochastic process. The parameter signals have an upper cutoff frequencies of 1.5 Hz for $F_m(t)$ and 3 Hz for $\Omega(t)$ and, therefore, vary slowly in time. Over long time scale (i.e. presumably tens of seconds) the stimulus parameters are characterized by an ergodic and recurrent process with stationary statistics (Hayes 1996; Marmarelis and Marmarelis 1978). Over short time scales (i.e. those over which neuronal integration occurs, in the order of tens to hundreds of milliseconds), however, the parameter trajectories change dynamically and consequently the dynamic ripple envelope is locally biased and non–stationary.

I proceed by averaging Eq. (2.47) over the stimulus parameter space. As for section 2.14, the analytic signal representation of Eq. (2.38) is employed. The time average expectation of Eq. (2.49) is expressed as an ensemble average

$$E\left[\cos\left(2\pi\Omega_i\zeta+2\pi F_{m,i}\tau\right)\right]=\left\langle\cos\left(2\pi\Omega_i\zeta+2\pi F_{m,i}\tau\right)\right\rangle= \quad (2.50)$$

$$\operatorname{Re}\left\{\left\langle\exp\left[j\left(2\pi\Omega_i\zeta+2\pi F_{m,i}\tau\right)\right]\right\rangle\right\}.$$

Since the stimulus parameters are statistically independent, the expectation is separable

$$\left\langle\exp\left[j\left(2\pi\Omega_i+2\pi F_{m,i}\tau\right)\right]\right\rangle=\left\langle\exp\left(j2\pi\Omega_i\zeta\right)\right\rangle\cdot\left\langle\exp\left(j2\pi F_{m,i}\tau\right)\right\rangle \quad (2.51)$$

and the temporal and spectral autocorrelation functions are independent. This simplification allows one to compute the expected correlation function independently for both stimulus dimension.

Since the local autocorrelation function is given by the ensemble average of Eq. (2.51), the global autocorrelation function will be determined by the relative occurrence of each parameter. As previously mentioned, the probability distribution function for the ripple density and modulation rate parameters are uniform and are given by

$$p(F_m) = \begin{cases} 1/2 F_{Max} & -F_{Max} \leq F_m \leq F_{Max} \\ 0 & \textit{otherwise} \end{cases} \quad (2.52)$$

$$p(\Omega) = \begin{cases} 1/2 \Omega_{Max} & -\Omega_{Max} \leq \Omega \leq \Omega_{Max} \\ 0 & \textit{otherwise} \end{cases} \quad (2.53)$$

where F_{Max} and Ω_{Max} are the upper cutoff parameters for each distribution. The spectro-temporal autocorrelation is obtained by substituting the parameter distributions into Eq. (2.51). Using the spectro-temporal distributions, the autocorrelation functions are given by

$$\langle \exp(j 2 \pi \Omega_i \zeta) \rangle = \int p(\Omega) e^{j 2 \pi \Omega \zeta} d \Omega = \textit{sinc}(2 \Omega_{Max} \tau) \quad (2.54)$$

$$\langle \exp(j 2 \pi F_{m,i} \tau) \rangle = \int p(F_m) e^{j 2 \pi F_m \tau} d F_m = \textit{sinc}(2 F_{Max} \tau) \quad (2.55)$$

where $\text{sinc}(x) = \sin(\pi x)/\pi x$. Note that Eq. (2.54) and Eq. (2.55) are equivalent to the impulse response of an ideal lowpass filter function of bandwidth $[-F_{Max}, F_{Max}]$ and $[-\Omega_{Max}, \Omega_{Max}]$ and gain $1/2F_{Max}$ and $1/2\Omega_{Max}$ respectively. Hence the ripple density and temporal modulation rate parameter distributions determine a priori the shape of the stimulus spectro-temporal power spectrum (otherwise known as the characteristic function, Cohen 1995). The dynamic ripple envelope therefore achieves a flat power spectrum over the predefined range of ripple densities and temporal modulation rates.

After combining Eqs. (2.49), (2.54), and (2.55), the spectro-temporal autocorrelation function is given by

$$R_{ss}(\tau, \zeta) = \frac{M^2}{8} \text{sinc}(2F_{Max}\tau) \text{sinc}(2\Omega_{Max}\zeta) R_{ww}(\tau, \zeta) \quad (2.56)$$

where the constant $M^2/8$ is exactly the variance (i.e. RMS value) of a sinusoid signal of amplitude $M/2$. The window autocorrelation function, $R_{ww}(\tau, \zeta)$, is a residual term from performing the instantaneous analysis of section 2.14. If one chooses a window of infinite σ_t and σ_x the window term drops out. As for Eq. (2.54) and (2.55), note that Eq. (2.56) is effectively the impulse response of an ideal 2-D lowpass filter (if one ignores the residual window term) with the described cutoff frequencies and gain $M/(\sqrt{8} \cdot \Omega_{Max} \cdot F_{Max})$ (Fig. 8 F). The dynamic ripple envelope therefore probes in an unbiased manner the chosen spectral and temporal envelope fluctuations as described by the parameter distributions.

2.16 Ripple Noise Local Autocorrelation

As for the dynamic ripple envelope we would like to estimate the instantaneous and global statistics of the ripple noise envelope. We consider the ideal zero-mean ripple noise signal

$$\bar{Y}(t, X) = \lim_{L \rightarrow \infty} \frac{1}{\sqrt{L}} \sum_{k=1}^L \bar{S}_k(t, X) \quad (2.57)$$

where the contrast transformation, $f(x)$, and finite number of dynamic ripple envelopes, $L=16$, are ignored for simplicity of the analysis. These effects are examined in detail in section 2.17.

We proceed as for section 2.14, by estimating the local autocorrelation function for the ripple noise envelope. The ripple noise local autocorrelation is given by

$$R_{yy}(\tau, \zeta | t_i) = E \left[\bar{Y}(t, X) w_i(t, X) \bar{Y}(t+\tau, X+\zeta) w_i(t+\tau, X+\zeta) \right] \quad (2.58)$$

where the rational window function, $w_i(t, X)$, of Eq. (2.36) is used to extract the local ripple signal. Substituting Eq. (2.57) into Eq. (2.58) results in

$$R_{yy}(\tau, \zeta | t_i) =$$

$$\lim_{L \rightarrow \infty} \frac{1}{L} E \left[\left(\sum_{k=1}^L \bar{S}_k(t, X) \sum_{l=1}^L \bar{S}_l(t+\tau, X+\zeta) \right) w_i(t, X) w_i(t+\tau, X+\zeta) \right] . \quad (2.59)$$

Expanding the inner argument of the expectation

$$R_{yy}(\tau, \zeta | t_i) = \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{k=1}^L E \left[\bar{S}_k(t, X) \bar{S}_k(t+\tau, X+\zeta) w_i(t, X) w_i(t+\tau, X+\zeta) \right] + \quad (2.60)$$

$$\frac{1}{L} \sum_{k=1}^L \sum_{l \neq k} E \left[\bar{S}_k(t, X) \bar{S}_l(t+\tau, X+\zeta) w_i(t, X) w_i(t+\tau, X+\zeta) \right] .$$

Note that the second sum drops out when using a window, $w_i(t, X)$, of infinite extent

since \bar{S}_k and \bar{S}_l are statistically uncorrelated by definition. This is valid only for sufficiently large windows, however, where presumably a large number of cycles of

\bar{S}_k and \bar{S}_l are averaged. For the window of finite extent used here this does not strictly hold and the second term must be considered. Since the individual elements,

\bar{S}_k and \bar{S}_l , are statistically independent by design the sum of the cross products of these terms will resemble (second term in Eq. (2.60)) a bandlimited spectro-temporal noise signal which we refer to as the error term, $e(t, X)$. For now we only consider the first term of Eq. (2.60). The effects of the error term, $e(t, X)$, which arise from using a window of finite extent are outlined in section 2.17.

Following the procedure for section 2.13, each individual dynamic ripple envelope is approximate using Eq. (2.32). The k^{th} dynamic ripple envelope is therefore

given by $\bar{S}_k(t, X|t_i) \approx \bar{S}_k(t, X)|_{t=t_i}$. Proceeding as for section 2.14 the ripple noise local autocorrelation function is expressed as

$$R_{yy}(\tau, \zeta) = \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{k=1}^L E[\bar{S}_k(t, X|t_i) \bar{S}_k(t+\tau, X+\zeta|t_i) w_i(t, X) w_i(t+\tau, X+\zeta)] + e(t, X) \quad (2.61)$$

The expectation inside the sum is the autocorrelation function for the k^{th} dynamic ripple and is therefore identical to Eq. (2.47). Substituting Eq. (2.47) the ripple noise local autocorrelation is

$$R_i(\tau, \zeta) = \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{k=1}^L R_{S_i S_k}(\tau, \zeta|t_i) + e(t, X) = \lim_{L \rightarrow \infty} \frac{1}{L} \sum_{k=1}^L \frac{M^2}{8} \cos(2\pi \Omega_{i(k)} \zeta + 2\pi F_{m,i(k)} \tau) R_{ww}(\tau, \zeta) + e(t, X) \quad (2.62)$$

where $R_{S_i S_k}(t, X|t_i)$ is the local autocorrelation function for the k^{th} dynamic ripple envelope and $\Omega_{i(k)}$ and $F_{m,i(k)}$ are its instantaneous parameters.

Finally we note that the parameters $\Omega_{i(k)}$ and $F_{m,i(k)}$ are random variables with distributions defined by Eq. (2.52) and (2.53). The sum of Eq. (2.62) therefore approaches the ensemble average operator, $\langle \cdot \rangle$, as $L \rightarrow \infty$ and the first term of Eq. (2.62) is effectively identical to Eq. (2.49). Using the result from section 2.15 we arrive

1. **Introduction**
2. **Background**
3. **Methodology**
4. **Results**
5. **Discussion**
6. **Conclusion**
7. **References**
8. **Appendix**
9. **Index**
10. **Table of Contents**

11. **Index**

12. **Index**
13. **Index**
14. **Index**
15. **Index**
16. **Index**
17. **Index**
18. **Index**

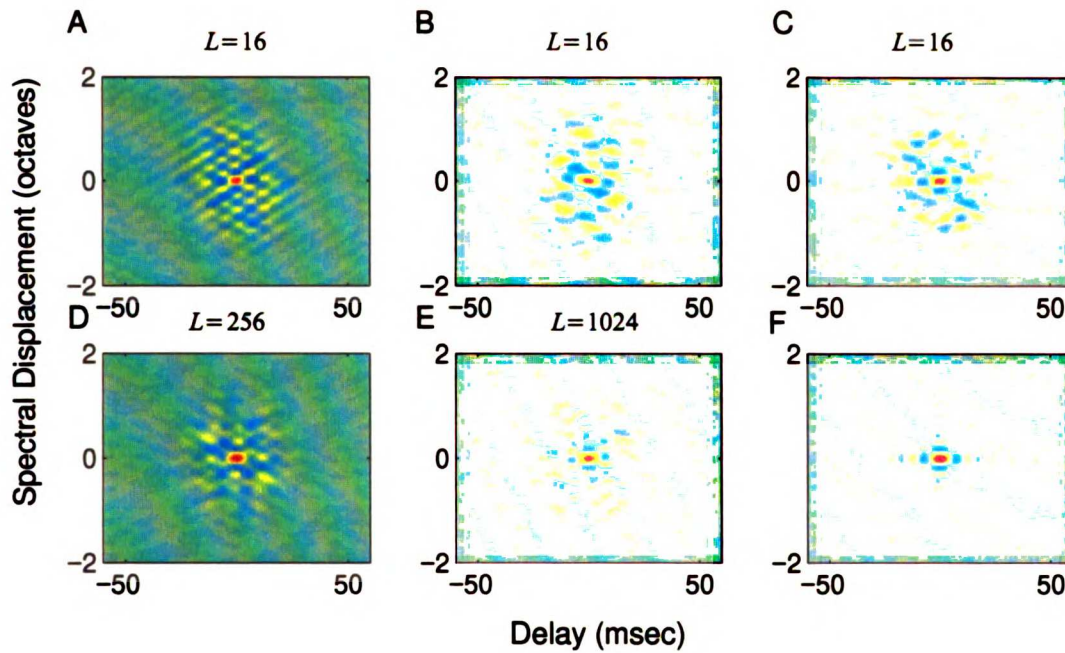


Figure 9: (A–E) Examples of the instantaneous spectro–temporal autocorrelation

function for the ripple noise envelope shown for $F_{Max} = 100$ Hz, $\Omega_{Max} = 4$

cycles/octave, $\sigma_t = 20$ ms and $\sigma_x = 0.5$ octaves. Local autocorrelation for $L=16$

(A–C), $L=256$ (D), and $L=1024$ (E) at distinct time instants. (F) The ripple noise global

autocorrelation function is impulsive and identical to the dynamic ripple global

autocorrelation (Fig. 8 F).

at the final result

$$R_{yy}(\tau, \zeta | t_i) = \frac{M^2}{8} \text{sinc}(2F_{Max}\tau) \text{sinc}(2\Omega_{Max}\zeta) R_{ww}(\tau, \zeta) + e(t, X) \quad (2.63)$$

The local autocorrelation of the ripple noise signal is therefore a "noisy" version

of the dynamic ripple global autocorrelation function. Examples of the instantaneous ripple noise autocorrelation function are shown for different time instants in Fig. 9. Analogous to the dynamic ripple global autocorrelation, the instantaneous autocorrelations have a central peak but now the surround is corrupted by noise. Hence, the overall effect of summing L independent dynamic ripple envelopes is that it generates a signal which is locally weakly correlated. The locally strong spectro-temporal correlations which are prevalent for the dynamic ripple (Fig. 8), speech, and numerous other vocalization sounds, are therefore absent in this sound.

2.17 Ripple Noise Local Autocorrelation: Effects of Finite Number of Dynamic Ripples ($L=16$) and Finite Window Size (σ_x and σ_t)

Here we consider the effects of summing a finite number of dynamic ripple envelopes to generate the ripple noise envelope as well as using a window of finite extent to estimate the local correlation statistics. It is desired to construct a ripple noise signal that closely matches the statistical properties described in the previous section ($L \rightarrow \infty$). The large computational demands required to generate this signal, however, prohibits us from using very large values of L . As an example, the twenty minute segment of the ripple noise signal used in this series of experiments took roughly 7 days to generate on a DEC Alpha series 500 (500 MHz CPU) workstation using MATLAB 5.1 (©, Mathworks Inc.).

Fig. 9 compares the target (Eq. 2.56) and the actual ripple noise local autocorrelation functions obtained using $L=16$, $L=256$, $L=1024$. Note that for all cases

the local autocorrelation varies on a trial to trial basis. This variability depends on the signal length used to estimate the local autocorrelation function (i.e. the window size σ_t and σ_x) and the number of averages, L , performed to construct the ripple noise signal. The autocorrelation estimates for very short segments are clearly much more variable than for long segments where the error is no longer reduced. Since auditory neurons integrate stimulus information over a restricted range of spectral and temporal scales, it is of interest to choose values of σ_t and σ_x which are physiologically plausible.

The key feature of the ripple noise autocorrelation which is relevant for characterizing auditory neurons is its impulse like properties. It is desired that the ripple noise local autocorrelation function approximate this property as closely as possible. From Fig. 9, it is clear that the central peak for the different values of L are qualitatively similar. The surround, however, is slightly more variable for lower values of L . We can quantify this behavior by measuring the peak-to-surround ratio, $\eta = \delta/\sigma$, which characterizes the amplitude of the central peak, $\delta = M^2/8$, relative to the surround standard deviation, σ .

To do so, we need to derive the variance of the ripple noise autocorrelation function, $R_{yy}(\tau, \zeta | t_i)$, for finite L and finite window dimensions σ_t and σ_x . We turn to Eq. (2.62) and note that both terms contribute to the autocorrelation error. We can express Eq. (2.62) as

$$R_{yy(L)}(\tau, \zeta | t_i) = \frac{1}{L} \sum_{k=1}^L R_{s_i s_k}(\tau, \zeta | t_i) + \frac{1}{L} \sum_{k=1}^L \sum_{l \neq k} R_{s_i s_l}(\tau, \zeta | t_i) \quad (2.64)$$

where $R_{s_i s_l}(\tau, \zeta | t_i)$ is the crosscorrelation between the k^{th} and the l^{th} local dynamic ripple signals, and L is now a finite quantity. The second quantity of Eq. (2.64) is exactly the error term, $e(t, X)$, described in section 2.16. The autocorrelation variance is expressed as

$$\sigma_L^2 = \text{var}[R_{yy(L)}] = \frac{1}{E[R_{ww}^2]} \left\langle E \left[\left(\frac{1}{L} \sum_{k=1}^L R_{s_i s_k} + \frac{1}{L} \sum_{k=1}^L \sum_{l \neq k} R_{s_i s_l} - R_{yy} \right)^2 \right] \right\rangle \quad (2.65)$$

where the shorthand notation $R_{yy(L)} = R_{yy(L)}(\tau, \zeta | t_i)$, $R_{yy} = R_{yy}(\tau, \zeta | t_i)$

$R_{s_i s_k} = R_{s_i s_k}(\tau, \zeta | t_i)$, $R_{s_i s_l} = R_{s_i s_l}(\tau, \zeta | t_i)$, and $R_{ww} = R_{ww}(\tau, \zeta)$ is used. Note that

we normalize by the energy of the window autocorrelation function,

$E[R_{ww}^2] = 2\pi\sigma_t\sigma_x$, since unlike $w_i(t, X)$, which has unit energy, the window

autocorrelation, R_{ww} , has a finite energy which biases the variance estimate. The trial

to trial variance is estimated by performing a spectro-temporal average with respect to

τ and ζ and subsequently an ensemble average with respect to the L element

dynamic ripple ensemble. The spectro-temporal average computes the variance from the

residual noise signal for a single trial. Note that the local ripple noise signal consists of

the sum of L statistically independent moving ripple envelopes which are randomly

chosen for each trial. Each trial therefore produces a local ripple noise autocorrelation

function (and consequently an error term) which varies from trial to trial. Here we are interested in determining this average trial to trial variability by computing the ensemble average for this error term.

We continue by dropping cross product terms. Since these are statistically independent and small we have

$$\sigma_L^2 \approx \frac{1}{E[R_{ww}^2]} \left\langle E \left[\left(\frac{1}{L} \sum_{k=1}^L R_{s_i s_i} - R_{yy} \right)^2 \right] \right\rangle + \frac{1}{L^2 E[R_{ww}^2]} \left\langle E \left[\left(\sum_{k=1}^L \sum_{l \neq k} R_{s_i s_l} \right)^2 \right] \right\rangle . \quad (2.66)$$

The first term (Term 1) of Eq. (2.66) contains the dominant source of error arising from the finite number of dynamic ripples which are summed to create the ripple noise signal.

The second term (Term 2), corresponds to the variance of the error signal, $e(t, X)$. This term contains a finite combination of cross product terms, and hence will be the dominant source of error from choosing a finite window size. Note that if the window is made infinite in its extent, the expectation $E[R_{s_i s_l}] = 0$ since the k^{th} and l^{th} envelope are statistically independent. Here we proceed by evaluating Term 1 and subsequently Term 2.

Term 1: Note that R_{yy} can be expanded as in Eq. (2.62). The first term can therefore be expanded as

$$\sigma_1^2 = \frac{1}{E[R_{ww}^2]} \left\langle E \left[\left(\frac{1}{L} \sum_{k=1}^L R_{S_i S_i} - \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_{S_i S_i} \right)^2 \right] \right\rangle \approx \quad (2.67)$$

$$\begin{aligned} & \frac{1}{E[R_{ww}^2]} \left[\frac{1}{L^2} \sum_{k=1}^L \langle E[R_{S_i S_i}^2] \rangle + \lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{k=1}^N \langle E[R_{S_i S_i}^2] \rangle - \lim_{N \rightarrow \infty} \frac{2}{NL} \langle E \left[\sum_{k=1}^L R_{S_i S_i} \sum_{n=1}^N R_{S_i S_i} \right] \rangle \right] \approx \\ & \frac{1}{E[R_{ww}^2]} \left[\frac{1}{L^2} \sum_{k=1}^L \langle E[R_{S_i S_i}^2] \rangle + \lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{k=1}^N \langle E[R_{S_i S_i}^2] \rangle - \lim_{N \rightarrow \infty} \frac{2}{NL} \sum_{k=1}^L \langle E[R_{S_i S_i}^2] \rangle \right] = \\ & \lim_{N \rightarrow \infty} \frac{\langle E[R_{S_i S_i}^2] \rangle}{E[R_{ww}^2]} \left[\frac{1}{L} + \frac{1}{N} - \frac{2}{N} \right] = \frac{\langle E[R_{S_i S_i}^2] \rangle}{E[R_{ww}^2] \cdot L} \end{aligned}$$

where the cross product terms and all terms from N to ∞ drop out. We need to evaluate

$$\langle E[R_{S_i S_i}^2] \rangle. \text{ Proposition: } \langle E[R_{S_i S_i}^2] \rangle = E[R_{ww}^2] \cdot M^4 / 128.$$

Proof: Consider a two dimensional Gabor function of general form $s(x_1, x_2)w(x_1, x_2)$

where $s(x_1, x_2) = A \cos(2\pi f_1 x_1 + 2i f_2 x_2) = \text{Re} \left[A \exp \left[j(2\pi f_1 x_1 + 2i f_2 x_2) \right] \right]$ and

$w(x_1, x_2) = \exp(-x_1^2/4/\sigma_1^2 - x_2^2/4/\sigma_2^2)$. Using the identity

$$\text{Re}[z_1] \text{Re}[z_2] = \frac{1}{2} \text{Re}[z_1 z_2^*] + \frac{1}{2} \text{Re}[z_1 z_2] \quad (2.68)$$

we can expand $s(x_1, x_2)^2 w(x_1, x_2)^2$ as

$$\frac{A^2}{2} w(x_1, x_2)^2 + \frac{A^2}{2} w(x_1, x_2)^2 \operatorname{Re} \left[\exp \left[j(4\pi f_1 x_1 + 4\pi f_2 x_2) \right] \right] . \quad (2.69)$$

Since the second term of Eq. (2.69) has zero-mean and $w(x_1, x_2)$ is independent of the parameter ensemble, the expectation gives

$$\langle E[s(x_1, x_2)^2 w(x_1, x_2)^2] \rangle = \frac{A^2}{2} E[w(x_1, x_2)^2] . \quad (2.70)$$

where $E[w(x_1, x_2)^2] = 2\pi\sigma_1\sigma_2$ for the given window. Extending these results for the more general case we note that $R_{s_i s_i}$ is a Gabor function of amplitude $A = M^2/8$ so that $\langle E[R_{s_i s_i}] \rangle = E[R_{ww}^2] \cdot M^4/128$.

The final result is obtained by substituting this equality into Eq. (2.67). This results in

$$\sigma_1^2 \approx \frac{M^4}{128L} , \quad (2.71)$$

the error variance of the first term of Eq. (2.66).

Term 2: The variance of term 2 is approximated as

$$\sigma_2^2 = \frac{1}{L^2 E[R_{ww}^2]} \left\langle E \left[\left(\sum_{k=1}^L \sum_{l \neq k} R_{S_k S_l}^2 \right)^2 \right] \right\rangle \approx \quad (2.72)$$

$$\frac{1}{L^2 \cdot E[R_{ww}^2]} \sum_{k=1}^L \sum_{l \neq k} \langle E[R_{S_k S_l}^2] \rangle = \frac{L-1}{L \cdot E[R_{ww}^2]} \cdot \langle E[R_{S_k S_l}^2] \rangle ,$$

where the cross terms drop out and the expectation is taken with respect to τ , ζ , and then with respect to the parameter ensemble. Although $E[R_{S_k S_l}]$ is strictly zero for infinite size windows ($\sigma_t \rightarrow \infty$ and $\sigma_x \rightarrow \infty$) this is not true for the finite duration window used here.

Combining Eq. (2.71) and (2.72), the overall error variance is expressed as

$$\sigma_L^2 \approx \frac{M^4}{128 \cdot L} + \frac{(L-1) \cdot \langle E[R_{S_k S_l}^2] \rangle}{L \cdot E[R_{ww}^2]} = \frac{M^4}{L} \left[\frac{1}{128} + \frac{(L-1) \cdot \langle E[R_{S_k S_l}^2] \rangle}{E[R_{ww}^2]} \right] \quad (2.73)$$

where $R_{S_k S_l}$ is the autocorrelation function for a ripple noise signal with $M=1$ and M^4 was factored out using the identity $R_{S_k S_l} = M^2 \cdot R_{S_k S_l}$. Note that the error variance has two components, one which is strongly dependent on $1/L$ and another which has only a weak dependence on L ($(L-1)/L$). As previously stated, the error variance from the first component arises from choosing a finite value of L and is therefore the dominant source of error from adding a finite number of dynamic ripples to generate the ripple noise signal. The second source of error, however, is largely dependent on the integration

window used to derive the local autocorrelation function.

Although we do not solve for Eq. (2.73) analytically, we numerically estimated the error, σ_L^2 , as a function of the integration window size (σ_x and σ_t) and as a function of the envelope bandwidth (F_{Max} and Ω_{Max}). In all instances it was noted that the average error was invariant of the window size and of the envelope bandwidth when measured as a function of the products $N_t = \sigma_t \cdot F_{Max}$ and $N_x = \sigma_x \cdot \Omega_{Max}$. These unit-less quantities are proportional to the maximum number of spectral and temporal cycles that fit in a window of standard deviation σ_x and σ_t , respectively.

The overall errors, σ_L^2 , are shown in Fig. 10 for various values of the N_t and N_x . The initial decrease in the curves corresponds exactly to the errors that arise from choosing a finite number (L) of dynamic ripple envelopes. Upon reaching the critical value $L_c = 1 + E[R_{ww}^2] / (128 \cdot \langle E[R_{s_i s_i}^2] \rangle)$ the curve quickly saturates. At this point, the errors associated with choosing a window of finite extent dominate and the curve becomes independent of L .

We again consider, the peak-to-surround ratio, η , which is a direct measure of the SNR for the ripple noise autocorrelation obtained for a fixed window size and finite L . Using the error variance of Eq. (2.73) we instantly obtain the peak-to-surround ratio as a function of L and window size. After combining terms

$$\eta_L = \frac{\delta}{\sqrt{\sigma_1^2 + \sigma_2^2}} \approx \sqrt{\frac{2L}{1 + 128 \cdot (L-1) \cdot \langle E[R_{s_i s_i}^2] \rangle / E[R_{ww}^2]}} \quad (2.74)$$

where the limiting values are $\eta \approx \sqrt{E[R_{ww}^2]/(64 \cdot \langle E[R_{s_i s_i}] \rangle)}$ for $L \geq L_c$ and $\eta \approx \sqrt{2L}$

for $L \leq L_c$. Of interest is the fact that the bound initially increases proportional to

\sqrt{L} but upon reaching the critical value, $L_c = 1 + E[R_{ww}^2]/(128 \cdot \langle E[R_{s_i s_i}] \rangle)$, is

subsequently independent of L and only determined by $\langle E[R_{s_i s_i}] \rangle$ and $E[R_{ww}^2]$. At this point, the peak-to-surround ratio is strictly a function of the integration window size used to estimate the autocorrelation.

Several implications follow: First we consider the problem of how many dynamic ripple envelopes are necessary to generate a ripple noise signal which closely approximate the ideal ripple noise envelope ($L \rightarrow \infty$). Since auditory neurons integrate stimulus information over a restricted spectral bandwidth and temporal extent, it makes sense to consider an idealized linear neuron of finite memory say $2\sigma_t$ and finite spectral integration bandwidth $2\sigma_x$. We consider the physiologically relevant times scales of $\sigma_t = 50$ ms for cortical neurons and $\sigma_t = 10$ ms for subcortical neurons. In either case an integration bandwidth of $\sigma_x = 0.5$ octave is used.

It is of interest to determine whether such a neuron could distinguish two arbitrary ripple noise signals, $S_1(t, X)$ and $S_2(t, X)$, which are composed from L_1 and L_2 dynamic ripple envelopes respectively, by performing a local analysis of the sound. Since a linear neuron can at most detect local 2nd order statistics of a stimulus,

1. The first part of the document is a list of names and titles, including the names of the authors and the titles of their works. This list is organized in a structured manner, likely serving as a table of contents or a list of references.

2. The second part of the document contains a series of numbered entries, possibly representing a list of items or a sequence of events. These entries are organized in a structured manner, likely serving as a table of contents or a list of references.

Eq. (2.73) tells us that such a neuron can in theory detect differences between $S_1(t, X)$ and $S_2(t, X)$ as long as the SNR, η_{L_1} and η_{L_2} , are different for the two envelopes. Note that upon reaching the critical value, $L=L_c$, the signal to noise ratio is fixed for all $L \geq L_c$ and is independent of L . The idealized neuron can therefore only distinguish the two ripple envelopes as long as one of the two envelopes has a signal to noise ratio which falls on the increasing portion of the peak-to-surround ration curve.

As an example, consider a ripple noise signal which is designed to excite subcortical neurons with parameters $F_{Max}=350$ Hz, $\Omega_{Max}=4$ cycles/octave, and $L=16$. For a linear neuron which has an integration bandwidth of $2\sigma_x=1$ octaves and temporal memory of $2\sigma_t=15$ ms, the corresponding scale invariant parameters are $N_x=4$ and $N_t=5.25$. From Fig. 10, note that for the chosen parameters, the peak-to-surround ratio η_{16} falls on the increasing portion of the curve and is therefore statistically different than for $L=\infty$. Hence such a neuron can in principle detect difference between the ideal and $L=16$ ripple noise envelopes since $\eta_\infty \approx 3 \cdot \eta_{16}$. In addition, note that the local statistics of the $L=16$ ripple noise envelope are likewise significantly different than for the dynamic ripple case, $L=1$, where $\eta_{16} = 3.7 \cdot \eta_1$ for this set of parameters.

Hence when designing ripple noise envelope to approximate bandlimited spectro-temporal white noise it is necessary to jointly consider the relevant parameter (i.e.

F_{Max} and Ω_{Max}) scales as well as the neuronal integration scales (σ_t and σ_x) since these set the limits for choosing reasonable value of L for a given application. Although we can in principle choose very large values of L to guarantee that we achieve statistical similarity to the ideal spectro-temporal ripple noise, such a procedure is not practical because of the large computational demands necessary to generate the ripple noise stimulus.

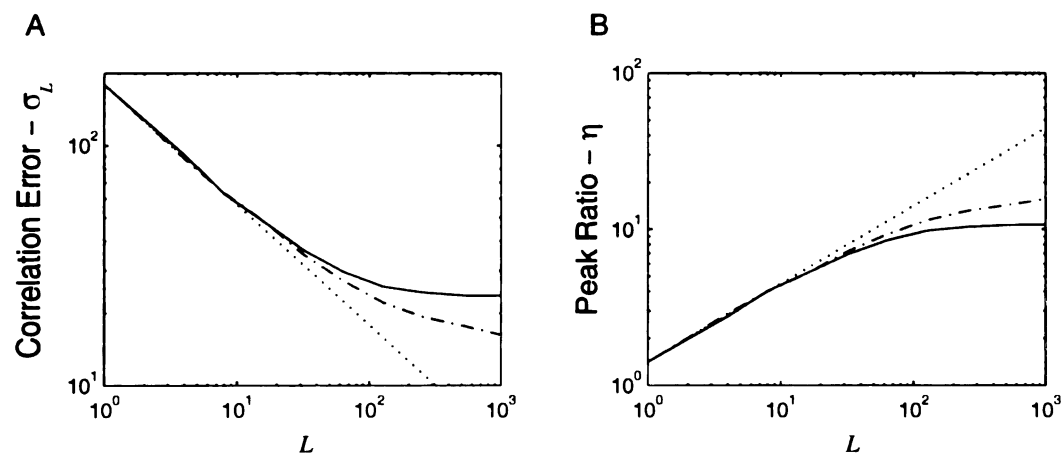


Figure 10: Ripple noise instantaneous-correlation error (A) and peak to surround noise ratio (B). Correlation error decreases monotonically with increasing L . Upon reaching a critical value of L_c the correlation error plateaus due to the finite integration window size, σ_t and σ_x . Shown for $2\sigma_t = 15$ msec (continuous), $2\sigma_t = 30$ msec (dashed-dotted), and $2\sigma_t = \infty$ msec (dotted) for a spectral integration window o

$2\sigma_x = 1$ octave and ripple parameters $F_{Max} = 350$ Hz and $\Omega_{Max} = 4$ cycles/octave. The correlation peak to surround noise ratio (shown for identical conditions, B) increases monotonically and flattens upon reaching L_c .

2.18 Ripple Noise Global Autocorrelation

As for the dynamic ripple, the ripple noise global autocorrelation function is obtained by averaging the instantaneous autocorrelation function, Eq. (2.63), over all time. Explicitly, this is expressed as

$$R_{yy}(\tau, \zeta) = E[R_{yy}(\tau, \zeta | t_i)] = \frac{M^2}{8} \text{sinc}(2F_{Max}\tau) \text{sinc}(2\Omega_{Max}\zeta) R_{ww}(\tau, \zeta) + E[e(t, X)] \quad (2.75)$$

where the first term is independent of the time expectation and the error signal has zero mean. Consequently the second term of Eq. (2.75) drops out and the ripple noise global autocorrelation is identical to the dynamic ripple global autocorrelation function, Eq. (2.56).

2.19 Compressed Ripple Noise Autocorrelation Function (Effects of $f(x)$)

The compressing nonlinearity, $f(x)$, was applied to the ideal ripple noise stimulus, Eq. (2.57), for experimental and practical considerations. This transformation serves as a contrast transformation which converts the amplitude distribution of the ripple noise from an normally distributed amplitude of variance $M^2/8$ to an uniformly distributed amplitude confined to an overall range $[0, -M]$ (in a decibel amplitude space). The amplitude distribution for this ripple envelope therefore has a slightly smaller variance of

$M^2/12$. Two arguments are presented for performing such a transformation: first, this transformation is performed so that one can consider the spectro-temporal processing capabilities of individual auditory neurons under identical intensity and contrast operating conditions. Secondly this nonlinearity serves to undermine any possible effects and nonlinearities which may arise from sound level dependencies in the neuronal responses (e.g. monotonic and nonmonotonic rate level curves) (Ehret and Merzenich 1982; Eggermont 1989).

Despite these arguments and experimental considerations, it is possible that performing such a transformation on a signal can have deleterious effects by significantly altering its spectro-temporal correlation characteristics. Although this transformation may be experimentally reasonable, it is possible that our experimental paradigm (in which the dynamic ripple and ripple noise stimuli have identical autocorrelation functions) is compromised. Here we present an analytically derived proof which shows that this is not so and that the compressed ripple noise autocorrelation function is essentially unaltered.

Consider a one dimensional ripple noise like signal, in this case a normally distributed bandlimited noise signal with unity standard deviation (Fig. 11). A one dimensional counterpart of the ripple noise envelope is used in order to facilitate the following derivation and since the following results can be directly extended to the more general two dimensional signal case (because the ripple noise signal has identical amplitude distribution). We start by considering the contrast transformation of Eq. (2.25) which can be expanded into a Taylor series. Upon substituting the Taylor series expansion (Gieck 1974)

$$e^{-\tau^2} = \sum_{n=0}^{\infty} \frac{(-1)^n \tau^{2n}}{n!} \quad (2.76)$$

into Eq. (2.26), the error function is expressed as a power series

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \int_0^x \tau^{2n} d\tau = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{x^{2n+1}}{2n+1} \quad (2.77)$$

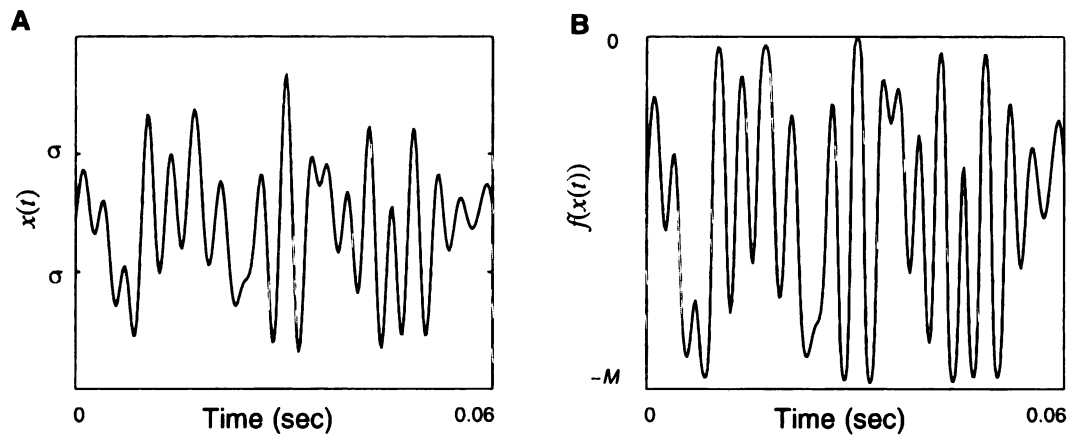


Figure 11: (A) Normally distributed bandlimited noise signal, $x(t)$. Compressed uniformly distributed signal, $f(x(t))$. Both signals have a bandwidth of 350 Hz.

As for the correlation analysis of sections 2.14–2.18, the quantity of interest is the autocorrelation function

$$r_{\bar{f}(x)\bar{f}(x)}(\tau) = E[\bar{f}(x(t))\bar{f}(x(t+\tau))] = \frac{M^2}{4} E[\text{erf}(x(t))\text{erf}(x(t+\tau))] \quad (2.78)$$

where $\bar{f}(x) = f(x) + M/2$ has zero mean, and $x(t)$ is a bandlimited normally distributed stochastic process with zero mean and unity standard deviation. Substituting Eq. (2.77) into Eq. (2.78) and expanding gives

$$r_{f(x)f(x)}(\tau) = \frac{M^2}{2\pi} \sum_{n=0}^{\infty} \frac{1}{(n!)^2 (2n+1)^2} \cdot E[x(t)^{2n+1} x(t+\tau)^{2n+1}] + \quad (2.79)$$

$$\frac{M^2}{2\pi} \sum_{n=0}^{\infty} \sum_{k \neq n} \frac{1}{n! k! (2n+1)(2k+1)} \cdot E[x(t)^{2n+1} x(t+\tau)^{2k+1}] .$$

Several points are immediately of interest. First the autocorrelation function of the transformed signal $\bar{f}(x(t))$ is a sum of the autocorrelation functions of $x(t)^{2n+1}$ and a sum of the crosscorrelation between $x(t)^{2n+1}$ and $x(t)^{2k+1}$ for $n \neq k$. Since $2n+1$ and $2k+1$ are odd, we are therefore dealing with odd powers of $x(t)$. Note that the correlation between two odd order powers of $x(t)$ is an even order moment of the signal and Eq. (2.79) is a sum of even order moments of $x(t)$ which are always positive non zero valued (Marmarelis and Marmarelis 1978). Consider the identity (Laning and Battin, 1956)

$$E[x_1 x_2 \cdots x_N] = \sum \prod E[x_i x_j] \quad (2.80)$$

where N is an even number, x_1, x_2, \dots, x_N are normally distributed random numbers with unity standard deviation, and the operator $\sum \prod$ corresponds to the sum of all

possible products of $E[x_i x_j]$ (a total of $(2N)!/N!2^N$ distinct combinations). This identity tells us that the expectations in Eq. (2.79) can be expanded as a sum of products of all the possible permutations of $E[x_i x_j]$ where x_i and x_j take the values of $x(t)$ or $x(t+\tau)$. With this in mind Eq. (2.79) is conceptually a composition of the following elementary building blocks

$$\begin{aligned}
 E[x(t)x(t)] &= 1 \\
 E[x(t+\tau)x(t+\tau)] &= 1 \\
 E[x(t)x(t+\tau)] &= r_{xx}(\tau) .
 \end{aligned}
 \tag{2.81}$$

It follows that the autocorrelation for $f(x(t))$, Eq. (2.78), can be expressed as a power series of general form

$$r_{f(x)f(x)}(\tau) = \sum_{n=1}^{\infty} \alpha_n r_{xx}(\tau)^n .
 \tag{2.82}$$

Although no attempts are made to evaluate the series coefficients, α_n , it is worthwhile to point out some observations. First the series coefficients are monotonically decreasing since the arguments in the sum of Eq. (2.79) have factorial terms which predominate in the denominator. Secondly, the elementary components which make up the power series are the autocorrelation function of the normally distributed signal, $x(t)$. Since the dominant terms are of low order (since the coefficients are rapidly decreasing with

1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100

increasing n), it is expected that the autocorrelation of the contrast transformed ripple noise does not differ much from that of the ideal normally distributed ripple noise.

Fig. 12 shows the experimentally derived autocorrelation functions for $x(t)$ and $f(x(t))$. As expected, the two curves are in close agreement. Likewise the power spectrum of $x(t)$ and $f(x(t))$ are also in close agreement. This result also generalizes for the two dimensional ripple noise case which is shown in Fig. 13. The contrast transformation therefore does not significantly alter the shape of the ripple noise autocorrelation function and its power spectrum. It can therefore be used without any deleterious effects.

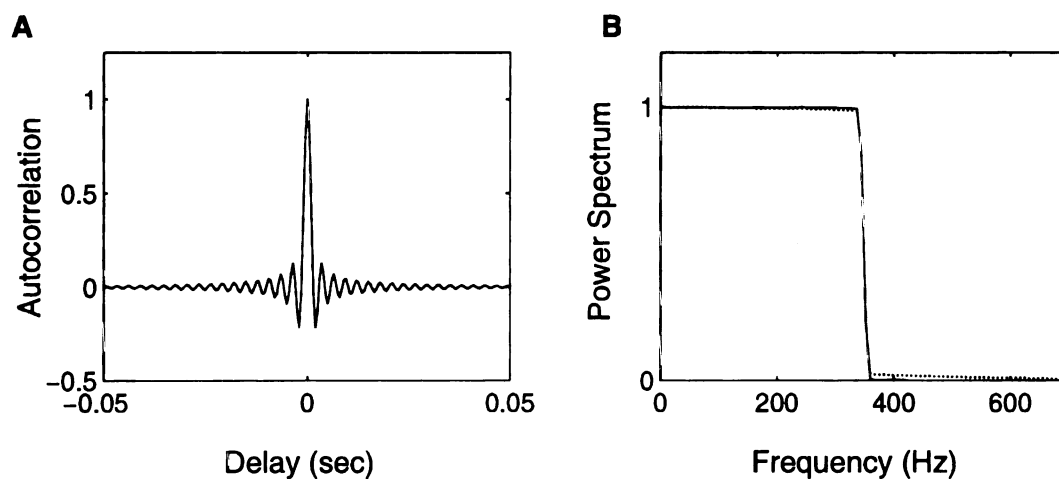


Figure 12: (A) Autocorrelation function for the uniformly distributed signal (continuous), $x(t)$, and for the compressed uniformly distributed signal, $f(x(t))$, of Fig. 11. The autocorrelation are in close agreement and are virtually indistinguishable for the two signals. (B) The corresponding power spectra are likewise in close agreement.

11/11/11
11/11/11
11/11/11

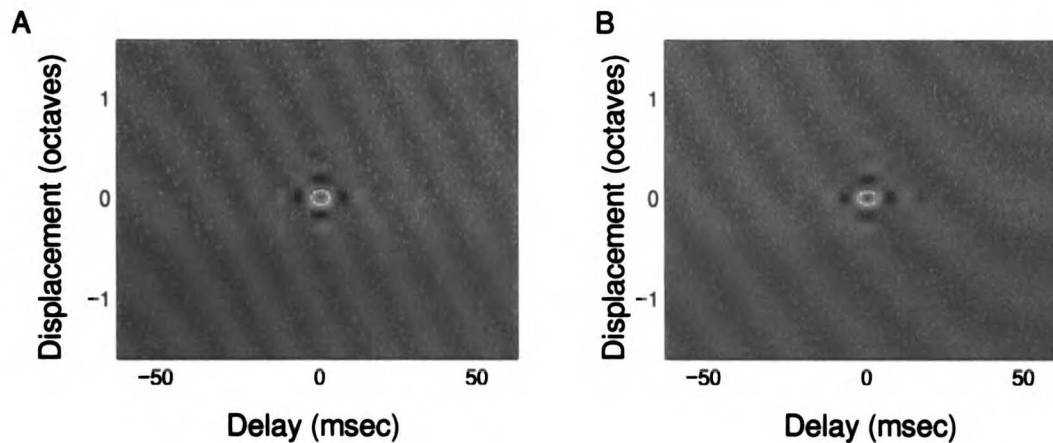


Figure 13: Theoretical spectro-temporal ripple noise autocorrelation function (A) and experimentally derived autocorrelation function for the compressed ripple noise (B).

The autocorrelations are in close agreement and have identical spectro-temporal patterns.

2.20 Dynamic Ripple Cross-Channel Correlations

The dynamic moving ripple and ripple noise stimuli have an appeal for studying the auditory system because they thoroughly probe a large range of spectro-temporal envelope correlations and stimulus dynamics. Although this is a necessary stimulus requirement for thoroughly testing out the responses preferences of a system, additional requirements are also necessary when one wishes to use a stimulus to derive spectro-temporal receptive fields (*STRF*) via reverse correlation procedures. In particular, it is prerequisite that distinct frequency channels, of the system under study, are independently activated by the driving stimulus. Here we derive the necessary conditions which show that the dynamic ripple is well suited for deriving *STRFs* via reverse correlation.

From Eq. (2.33), the dynamic ripple is locally expressed by a static spectro-

temporal sinusoid of spectral and temporal frequencies Ω_i and $F_{m,i}$. Here we are interested in computing the instantaneous cross-channel correlation function

$$R_{kl}(\tau|t_i) = E \left[\bar{S}(t+\tau, X_k|t_i) w_i(t+\tau) \bar{S}(t, X_l|t_i) w_i(t) \right] \quad (2.83)$$

where k and l are the carrier channel indices, $w_i(t)$ is a temporal Gaussian window (1-D version of Eq. (2.36)) of unit energy, and $E[\cdot]$ is taken with respect to the time variable only (unlike sections 2.14–2.18 where the expectation is taken with respect to the temporal and spectral variables). Unlike the analysis of sections 2.14–2.18, where the spectro-temporal correlation was computed jointly, here we are interested in the temporal correlations that exist between two distinct channels ($k \neq l$). Proceeding by replacing $\bar{S}(t, X_k|t_i)$ with Eq. (2.83) we get

$$R_{kl}(\tau|t_i) = \frac{M^2}{4} E \left[\sin(\text{Arg}_k(t+\tau)) \sin(\text{Arg}_l(t)) w_i(t+\tau) w_i(t) \right] \quad (2.84)$$

where $\text{Arg}_k(t) = 2\pi\Omega_i X_k + 2\pi F_{m,i} t + \Phi_i$ is the instantaneous argument (Eq. (2.31)).

Following a procedure almost identical to section 2.14 reveals that the instantaneous cross-channel correlation function is given by

$$R_{kl}(\tau|t_i) = \frac{M^2}{8} \cos(2\pi\Omega_i(X_k - X_l) + 2\pi F_{m,i}\tau) R_{ww}(\tau) \quad (2.85)$$

where

$$R_{ww}(\tau) = E[w_i(t+\tau)w_i(t)] = \exp\left[-\frac{\tau^2}{4\sigma_i^2}\right] \quad (2.86)$$

is the unit energy Gaussian window temporal autocorrelation function.

It is noted that the local cross-channel temporal correlations of the ripple signal are strongly influenced by the spectral displacement, $X_k - X_l$, between different channels. As for the analysis of natural sounds (chapter 1), we are interested in the cross-channel local correlation coefficient which quantifies the similarity of the temporal modulations across spectral bands. Using Eq. (2.85) it is easy to show that local cross-channel correlation coefficient is given by (see chapter 1)

$$\rho_{kl} = \cos(2\pi\Omega_i(X_k - X_l)) \quad (2.87)$$

Hence the instantaneous temporal modulations of the dynamic ripple have a local correlation distance which is continuously changing and determined by Ω_i . Note that the correlation coefficient oscillates between -1 and 1 indicating that the dynamic ripple temporal modulations are continuously phase shifted from -180 to 180 degrees across distinct carrier channels. These cross-channel influences are reminiscent of those observed for vocalization sounds which can show strong across-channel interactions (see chapter 1).

100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200

str
-2
78
t
v
LX
inte
dista

Although the local cross-channel correlations are strictly dependent on the instantaneous ripple density parameter, the global cross-channel correlations do not obey this rule. Following a procedure analogous to section 2.15, the global cross-channel correlation function is expressed as

$$R_{kl}(\tau) = \frac{M^2}{8} \text{sinc}(2F_{Max}\tau) \text{sinc}(2\Omega_{Max}(X_k - X_l)) R_{ww}(\tau, \zeta) . \quad (2.88)$$

The cross-channel correlation coefficient is likewise obtained as

$$\rho_{kl}(\tau) = \text{sinc}(2\Omega_{Max}(X_k - X_l)) . \quad (2.89)$$

Although the instantaneous temporal modulations of distinct carrier channels are strongly correlated for the dynamic ripple as described by Eq. (2.87), Eq. (2.88) and (2.89) tell us that the dynamic ripple envelope nonetheless preserves the property by which the temporal modulations of distinct carrier channels are globally uncorrelated up to the limits set by the spectral bandwidth (Ω_{Max}) of the spectral gratings. The dynamic ripple envelope therefore satisfies the necessary conditions for deriving *STRF* which require that the temporal modulations of distinct channels are globally uncorrelated. Note that this is true assuming that the neurons under study have a spectral integration bandwidths of at least $1/\Omega_{Max}$ which corresponds to the spectral correlation distance of this stimulus. For our case, $\Omega_{Max} = 4$ cycles/octave, which corresponds to a

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

spectral resolution of 1/4 octaves.

2.21 Ripple Noise Cross-Channel Correlations

Similar to the dynamic ripple, it is likewise necessary to consider the local and the global cross-channel correlations for the ripple noise envelope in order to determine its suitability for estimating *STRFs*. Consider the instantaneous cross-channel correlation function

$$R_{kl}(\tau|t_i) = E \left[\bar{Y}(t+\tau, X_k|t_i) w_i(t+\tau) \bar{Y}(t, X_l|t_i) w_i(t) \right] \quad (2.90)$$

for the idealized ripple noise envelope, $\bar{Y}(t, X)$, of Eq. (2.57). Following a similar approach to sections 2.16, it is easy to show that

$$R_{kl}(\tau|t_i) = \frac{M^2}{8} \text{sinc}(2F_{Max}\tau) \text{sinc}(2\Omega_{Max}(X_k - X_l)) R_{ww}(\tau, \zeta) + e(\tau) \quad (2.91)$$

the instantaneous cross-channel correlation function is simply a "noisy" version of the dynamic ripple global cross-channel correlation function. By averaging over all time instants, t_i , it is clear the error term, $e(t)$, averages out and so the global cross-channel correlations are identical to the global cross-channel correlation for the dynamic ripple envelope, Eq. (2.88). Similar to the arguments presented in sections 2.14, 2.15, 2.16, and 2.18 the dynamic ripple and ripple noise envelopes differ only in their local

statistics and are identical in their global statistics. Both the ripple noise and dynamic ripple are therefore equally suited for estimating *STRFs*, assuming that the neuron under study responds in a linear or quasi-linear fashion.

2.22 Dynamic Ripple and Ripple Noise Higher-Order Correlations

As previously mentioned, the response properties of a nonlinear system can be strongly influenced by the properties of the driving stimulus. Response gain terms, response interaction terms, and response dynamics all interact in a complex manner which determines the mode of operation of a dynamic nonlinear system. Examples of such include nonlinear rate level dependencies of auditory neurons (Ehret and Merzenich 1982; Eggermont 1989), response adaptation (Smirnakis *et al.* 1997), feature selectivity (Suga, Simmons, and Jen 1975; Suga and Jen 1976; Margoliash 1983), and combination sensitivity (Margoliash and Fortune 1992; Olsen and Suga 1991a 1991b). Consequently, it is necessary to understand, at least conceptually, the higher-order characteristics of the driving stimulus since these are responsible for activating nonlinear response components and since these serve as constraints on the possible modes of operation of the system under study.

In sections 2.15 and 2.18, the spectro-temporal autocorrelation function was derived and shown to be identical (with the exception of a multiplicative constant of 1.5) for the ripple noise and the dynamic ripple envelopes. This stimulus characterization showed that both stimuli satisfy the necessary requirements which are prerequisite for systematically identifying and characterizing central auditory neurons via reverse correlation. Although the grand average autocorrelations are identical for these two

stimuli, it was noted that the stimulus dynamics are nonetheless vastly different since the dynamic ripple has a time-varying autocorrelation function with locally correlated statistics. For a system with nonlinear response dynamics, such dynamic stimulus attributes can alter the effective operating regime of the nonlinear system in a time dependent manner, therefore altering its response characteristics and efficiency.

A second source of motivation for understanding the higher-order stimulus characteristics arises from the fact that the estimated linear impulse responses, obtained using reverse correlation methods, are actually a joint characterization of the linear and nonlinear elements of the system (section 2.5). When estimating a systems "linear" impulse response via reverse correlation, higher-order response terms are projected onto the first-order Wiener kernel (see Fig. 1 and section 2.5; Marmarelis and Marmarelis 1978). Although the estimated linear kernel provides an efficient linear descriptor of the system under study, it is nonetheless corrupted by nonlinear response components. The computed Wiener kernel is actually a composition of the system's first-order Voltera kernel (the true linear part of the system) and all of the higher-order odd Voltera kernels of the system which are functionally driven by higher-order characteristics of the stimulus (Marmarelis and Marmarelis 1978). Although this property makes Wiener kernel a very efficient estimator (since linear and nonlinear characteristics are combined into one descriptor), it makes it difficult dissociate linear and nonlinear response mechanism. A thorough understanding of the higher-order stimulus properties facilitates this process by setting constraints on the types of responses that can be elicited.

For a given spectro-temporal envelope, $S(t, X_k)$, consider the n^{th} order autocorrelation function

$$R(\tau_1, \zeta_1, \dots, \tau_n, \zeta_n) = E[S(t, X)S(t+\tau_1, X+\zeta_1) \cdots S(t+\tau_n, X+\zeta_n)] \quad (2.92)$$

Although it is not within the scope of this manuscript to derive an analytic solution for the stimulus higher-order autocorrelation functions, it is nonetheless of interest to point out some general observations.

A clear distinction between the ripple noise and dynamic ripple stimulus is the presence of cross product terms in the stimulus higher-order autocorrelations. Since the ripple noise stimulus is a composition of L independently chosen dynamic ripple envelopes, a nonlinear system which is probed using the ripple noise is subjected to such interaction components. As an example consider a second-order nonlinear system exposed to a ripple noise stimulus with $L=2$. At a given time instant it is noted that the ripple noise is composed of two static moving ripple envelopes with temporal modulation rates of say $F_{m,1} = 183$ and $F_{m,2} = 23$ Hz and ripple densities of $\Omega_1 = 0.2$ and

$\Omega_2 = 2.3$ cycles/octave. Upon examining the response of such a system to the ripple noise ($L=2$), one can in principle find linear responses to the stimulus 1st order components, $F_{m,1}$ and $F_{m,2}$, as well as nonlinear responses to the stimulus higher-order components which include DC terms, frequency doubling terms, $2F_{m,1}$ and

$2F_{m,2}$, and cross product terms, $F_{m,1} - F_{m,2}$ and $F_{m,1} + F_{m,2}$. The same ideas hold for the possible types of responses and interaction terms along the spectral axis if the system has a similar nonlinearity along that dimension. In general the ripple noise

stimulus preserves all possible parameter interaction terms of up to L^{th} order.

By comparison, the dynamic ripple envelope represents a subset of the ripple noise with $L=1$. In accordance with the above observations, the dynamic ripple stimulus does not have any interaction terms in its higher-order autocorrelation function. This does not indicate that the dynamic ripple envelope lacks all of the higher-order autocorrelation functions. On the contrary, the dynamic ripple has well defined higher-order autocorrelations functions that are simply missing the interaction terms which are prevalent in the ripple noise stimulus. Following the example for the ripple noise envelope, we note that for a general nonlinear system of n^{th} order which is exposed to the dynamic ripple with driving frequency $F_{m,1}$ ($L=1$) at a fixed time instant, one can in theory observe DC response components as well as response components at integer multiples of the driving frequency $kF_{m,1}$ for $k=1, \dots, n$. With the exception of the 1st order component, $F_{m,1}$, which arises at the output partly from the systems linear elements and partly from the odd order nonlinear elements, all other response components arise solely from the systems nonlinearities (Marmarelis and Marmarelis 1978).

In addition to considering the possible interaction products for the ripple noise and dynamic ripple stimuli, we likewise need to consider the subtle differences in the amplitude distribution for the two envelopes. It is conceptually clear, for example, that increasing the contrast or modulation depth of a signal increases the effective power and, hence, the effective driving force. Note, however, that in addition to the low-order statistics of the amplitude distribution (e.g. the variance) higher-order moments of the

envelope can increase or decrease the effective nonlinear driving force by activating nonlinear contrast dependencies, all of which can change the effective operating point of the system. Hence it is necessary to devise a measure of the effective nonlinear driving force which the stimulus provides. We do so in the next section by considering the higher-order moments of the ripple noise and dynamic ripple stimuli.

2.23 Dynamic Ripple and Ripple Noise Higher-Order Correlation Strength

As a measure of the strength of correlation we consider the higher-order moments of the spectro-temporal envelope

$$E[S(t, X)^{n+1}] . \quad (2.93)$$

The n^{th} order moment is derived from Eq. (2.92) by evaluating the n^{th} order spectro-temporal autocorrelation function at zero delay, $\tau_i = 0$, and at zero spectral displacement, $\zeta_i = 0$, for $i = 1, \dots, n$. Since the n^{th} order moment of a two-dimensional signal is independent of the signal's spectro-temporal correlations (independent of τ_i and ζ_i), and is only a function of the amplitude statistics (e.g. contrast statistics), we can evaluate Eq. (2.93) by considering the ensemble average of a random variable with identical amplitude distributions as for the signals of interest. For the dynamic ripple envelope one can therefore consider a random variable

$$y = \frac{M}{2} \sin(2\pi x) \quad (2.94)$$

where x is an uniformly distributed random variable in the interval $[0,1]$. Since the amplitude distribution for this random variable is identical to the amplitude distribution of the zero mean dynamic ripple envelope (Fig. 7) one can therefore use it to evaluate its higher-order moments. Evaluating the ensemble average gives

$$\begin{aligned} \langle y^{n+1} \rangle &= \frac{M^{n+1}}{2^{n+1}} \langle \sin^{n+1}(2\pi x) \rangle = \\ &= \frac{M^{n+1}}{2^{n+1}} \int_0^1 \sin^{n+1}(2\pi x) p(x) dx = \frac{M^{n+1}}{2^{n+1}} \int_0^1 \sin^{n+1}(2\pi x) dx \end{aligned} \quad (2.95)$$

Since y has a zero-mean symmetric amplitude distribution, all odd order moments (n even) are zero valued. We can therefore evaluate (2.95) for odd n only.

For the ripple noise envelope, consider a random variable

$$z = \frac{M}{2} x \quad (2.96)$$

where x is uniformly distributed in the interval $[-1,1]$, and z is uniformly distributed in the interval $[-M/2, M/2]$. The amplitude distribution of z is identical to that of the zero mean ripple noise envelope. The n^{th} order moment is given by

$$\langle z^{n+1} \rangle = \frac{M^{n+1}}{2^{n+1}} \int x^{n+1} p(x) dx = \frac{M^{n+1}}{2^{n+1}} \int_{-1}^1 x^{n+1} / 2 dx = \frac{M^{n+1}}{2^{n+1}(n+2)} \quad (2.97)$$

for odd n . Similar to y , z also has a symmetric zero mean amplitude distribution and consequently $\langle z^{n+1} \rangle = 0$ for even n .

Having derived the n^{th} moments for the dynamic ripple and the ripple noise signals, it is useful to compare the relative magnitude of these quantities since these partially determine the amount of nonlinear driving force. To do so we consider the ratio of the n^{th} moments

$$\eta_n = \frac{\langle y^{n+1} \rangle}{\langle z^{n+1} \rangle} = (n+2) \int_0^1 \sin^{n+1}(2\pi x) dx \quad (2.98)$$

for odd n . The moments are available for odd n only since $\langle y^{n+1} \rangle$ and $\langle z^{n+1} \rangle$ are zero valued for even n . Fig. 14 shows a plot of the correlation strength ratio, η_n , for odd values of n . Of interest is the fact that η_n is a monotonically increasing function of n . This indicates that the dynamic ripple envelope has considerably more drive force to activate nonlinear elements of a system. This source of functional drive arises solely from the amplitude (contrast) statistics of the signal and is independent of the spectro-temporal content.

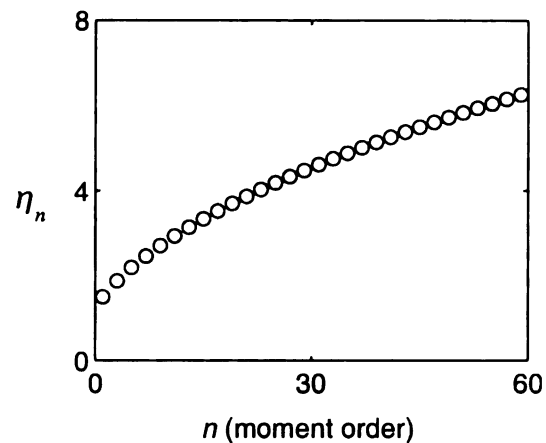


Figure 14: Ratio of the n^{th} -order moment for the dynamic ripple relative to the ripple noise, η_n . Shown for even values of n only. For all values of n the dynamic ripple n^{th} -order moment is larger than for the ripple noise and the correlation strength ratio is therefore above unity.

2.24 Ripple Noise and Dynamic Ripple Stimulus Alterations

Although a number of structural components and parameters distributions of the dynamic ripple and ripple noise stimuli are ethologically derived and are well suited for studying numerous aspects of auditory processing, they nonetheless do not possess all of the structural characteristics present in natural sounds. Of interest is the fact that certain high-order spectro-temporal statistics of these sounds are not representative of those found in natural sound environments. These include harmonicity, comodulation, and $1/f$ modulation spectrum, to name a few. Clearly, these sounds can not be used to study auditory processing for these acoustic parameters. In principle one can study auditory processing to such structural sound components independently of those features found in the dynamic ripple and ripple noise signals by using simple sounds which incorporate

these structural components. However, since environmental sounds generally consist of a mixture of many structural components, all of which can in principle modify the neuronal response characteristic of central auditory neurons to other sounds components (because of nonlinear interactions), it is necessary to consider sounds which can simultaneously probe numerous stimulus dimensions in a theoretically sound manner. Here we consider several of numerous sound alterations which can be performed on the dynamic ripple and ripple noise stimuli to accommodate a more general analysis of auditory processing. Such sound generation considerations are necessary for understand the general operating principles which the auditory system utilizes for complex sound encoding.

2.25 Harmonic Ripple

In both instances the dynamic ripple and ripple noise envelope were derived using logarithmically spaced carrier components which were individually amplitude modulated by the sounds spectro-temporal envelope. This carrier distribution was principally chosen so that the stimuli excite the primary sensory epithelium with an uniform energy distribution. In many natural signals, such as voiced speech, animal vocalizations, and other environmental sound sources, such carrier distributions are not observed. Instead, such sounds generally have harmonically spaced carrier components which arise from vocal fold vibrations in the larynx and other vibrating media. Such harmonically spaced carrier elements convey important acoustic cues which determine the quality (e.g. pitch) and other perceptual properties of the sound (Plomp 1967).

A simple alternative to using octave spaced carriers is derived by incorporating

harmonically spaced carrier elements. To do this, one can use Eq. (2.15) but now the carrier elements are chosen according to the rule: $f_k = k \cdot f_0$, where f_0 is the fundamental frequency of the harmonic stack, $k = 1, 2, \dots, N$, and $N \cdot f_0$ is the maximum frequency of the signal. Despite the fact that the carrier elements are now equally spaced on a linear frequency axis, the spectro-temporal envelope is still described on an octave frequency axis. The overall properties of the spectro-temporal envelope are not altered and all of the derivations for the autocorrelation and *STRF* measurements still hold since the carrier components do not affect these.

2.26 Comodulated Ripple

Comodulation is a common source of functional driving force present in numerous environmental sound sources (Nelken, Rotman, and Yosef 1999) by which distinct spectral channels are coherently turned on or off. Psychophysical studies (Hall, Haggard, and Fernandes 1984) have demonstrated that across channel comparisons of such characteristic features can increase detection thresholds. Clearly, since the auditory system makes use of such structural sound features it makes sense to incorporate these if one is interested studying neuronal responses to such.

We considering a comodulation function, $C(t)$, which is constructed so that it is independent of the dynamic ripple spectro-temporal envelope, $S(t, X)$. In particular, we require that the temporal crosscorrelation function of these two signals satisfy

$$r_{CS}(\tau) = E[C(t+\tau)S(t, X)] = 0 \text{ for all } \tau, \text{ so that it does not interfere with the}$$

correlation statistics of the dynamic ripple envelope. The comodulated dynamic ripple

envelope is expressed as $S_c(t,X)=C(t)\cdot S(t,X)$.

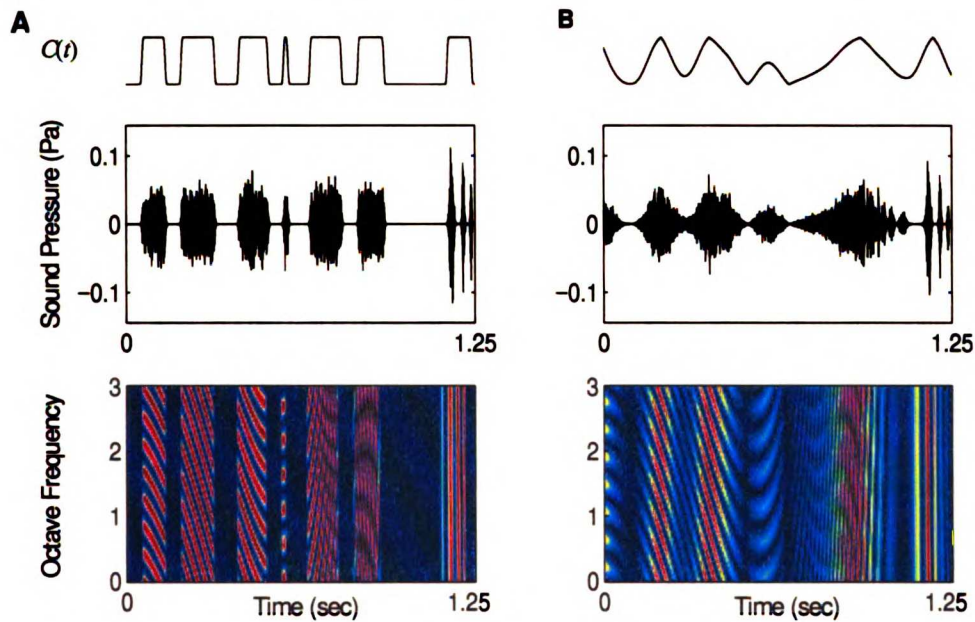


Figure 15: Comodulated dynamic ripple stimulus. The comodulation function, $C(t)$, is used to turn the dynamic ripple stimulus on and off. Shown for two comodulation sequences: m-sequence comodulation signal (A, top) and a bandlimited uniformly distributed signal (B, top). The comodulation function turns the sound pressure stimulus waveform, $s(t)$, on and off in a random fashion (A and B, middle). The comodulated spectro-temporal envelope (A and B, bottom) shown for the comodulation sequence of A and B respectively.

In adhering to our general design formulation, we consider several functional forms of $C(t)$ which are compatible with reverse correlation procedures. It is required that, in addition to the independence criterion $r_{CS}(\tau)=0$, the temporal autocorrelation

function of $C(t)$, $r_{CC}(\tau) = E[C(t+\tau)C(t)]$, have impulse like properties.

Examples of such are shown in Fig. 15 for an m-sequence and for a bandlimited uniformly distributed noise sequence. Clearly, the overall effect of such a modulation component is to turn the spectro-temporal envelope on and off in a coherent fashion, independently of the spectro-temporal envelope.

2.27 1/f Envelope Spectrum Ripple

As previously mentioned natural sounds often have modulation spectrum which show a strongly biased 1/f dependency (Attias and Schreiner, 1998a). This is unlike the flat modulation spectrum of the dynamic ripple and ripple noise, which have equal energy distribution for all modulation frequencies. To design ripple stimulus with a biased 1/f energy distribution we simply need to consider the probability distribution of the stimulus parameters. Recall that the shape of the modulation rate parameter distribution, Eq. (2.52), determines a priori the shape of the modulation spectrum for the ripple envelope. We can therefore consider a dynamic ripple and ripple noise envelope with a 1/f modulation rate parameter distribution

$$p(F_m) = \begin{cases} C \cdot 1/F_m^{-\alpha} & F_{Min} \leq |F_m| \leq F_{Max} \\ 0 & \text{otherwise} \end{cases} \quad (2.99)$$

where α is the spectral exponent which determines the slope of the power spectrum (on

a doubly logarithmic plot) and the constant $C = 2 / [(1 - \alpha) \cdot (F_{Max}^{1-\alpha} - F_{Min}^{1-\alpha})]$ is chosen so that $\int p(F_m) dF_m = 1$.

An example of the $1/f$ ripple envelope is shown in Fig. 16 for $F_{Min} = 1$ and $F_{Max} = 50$ Hz and $\alpha = 2$. Unlike the flat modulation spectrum ripple noise of Fig. 3, the $1/f$ modulation spectrum ripple envelope has higher energy for the low frequency modulations.

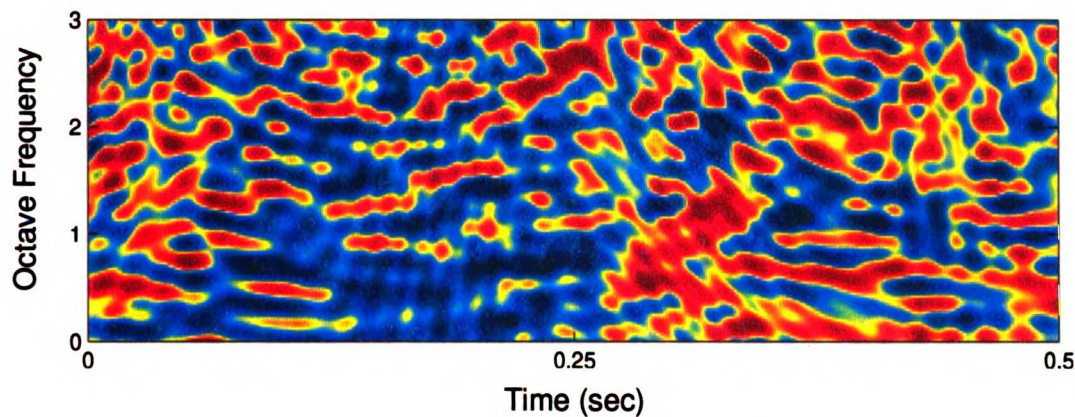


Figure 16: $1/f$ temporal modulation spectrum ripple noise signal. Low frequency temporal modulations are more prevalent than for the flat modulation spectrum ripple noise of Fig. 3.

2.28 Conclusion

Although the reverse correlation procedure has been readily used in a number of neuronal systems (including visual and auditory), little attention has been given to its practical limitations and its general compatibility with neuronal systems. Recently the *STRF* procedure has gained overwhelming popularity in audition because of its overall

simplicity and because it provides more complete estimate of the neuronal transformations performed by central auditory neurons (when compared to the conventional tuning curve). Because of the increasing complexity of the neuronal circuits in central stations and the subsequent increase in complexity of the neuronal processing it is possible that standard reverse correlation procedures may fail when performed with conventional stimulus (e.g. white noise, m-sequence etc.) (Theunissen *et al.* 2000). Despite this possible limitation, little attention has been given to proper stimulus design. Here we address both ecological and theoretical issues related to the compatibility of this procedure for identifying general mechanisms for complex sound processing.

To overcome possible limitations we designed two complex acoustic stimuli with spectro-temporal correlation statistics that mimic those observed in natural sounds (see chapter 1). These stimuli circumvent many of the theoretical and practical limitations of conventional reverse correlation stimuli. Although they preserve some of the basic correlation characteristics observed in natural sounds, they are nonetheless significantly simpler to quantify. Because of this, they allow for direct experimental control over a number of stimulus parameters while allowing us to determine which spectro-temporal parameters are of relevance to the auditory system. Realistically, the dynamic ripple and the ripple noise sounds can be used to test the hypothesis that instantaneous spectro-temporal correlations are of significant importance to the auditory system (and possibly to all sensory systems). A number of higher-order stimulus correlations and stimulus dynamics were independently evaluated for the ripple noise and the dynamic ripple sounds. These included global versus instantaneous correlations as well as low-order and high-order correlation statistics. These sounds are functionally distinct insofar as their

stimulus dynamics and instantaneous correlations are concerned. Given that both sounds have identical global and low-order statistics, the overall effects of stimulus dynamics and higher-order correlations can therefore be evaluated experimentally (see chapter 3).

2.29 References

- A. M. H. J. Aersten, J. H. J. Olders, and P. I. M. Johannesma. Spectro-temporal receptive fields in auditory neurons in the grass frog: analysis of the stimulus-event relation for tonal stimulus. *Biological Cybernetics* **38**, 235–248, 1980.
- A. M. H. J. Aersten, J.H.J. Olders, and P. I. M. Johannesma. Spectro-temporal receptive fields in auditory neurons in the grass frog: analysis of the stimulus-event relation for natural stimulus. *Biological Cybernetics* **30**, 195–209, 1981.
- A. Anzai, I. Ohzawa, R.D. Freeman. Neural mechanisms for processing binocular information: I. Simple cells. *J. Neurophysiol.* **82** (2), 891–908, 1999.
- H. Attias, and C.E. Schreiner. Temporal Low Order Statistics of Natural Sounds. *Advances in Neural Information Processing Systems* **10**, 27–33 (1998a).
- H. Attias, and C.E. Schreiner. Coding of Naturalistic Stimuli by Auditory Neurons. *Advances in Neural Information Processing Systems* **10**, 103–109 (1998b).
- J.S. Bendat. *Nonlinear systems analysis an identification from random data*. John Wiley and Sons, New York, 1990.
- V. Bringuier, F. Chavane, L. Glaesr, Y. Frégnac. Horizontal propagation of visual activity in the synaptic integration field of are 17 neurons. *Science* **283**, 695–699, 1999.
- B.M. Calhoun, and C.E. Schreiner. Spectral Envelope Coding in Cat Primary Auditory Cortex: Linear and non-linear effects of stimulus characteristics. *European J. of Neurosci.* **10**, 926–940 (1998).
- J.H. Casseday, D. Ehrlich, and E. Covey. Neural Tuning for Sound Duration. Role of Inhibitory Mechanisms in the Inferior Colliculus. *Science*, Vol. 264, 847–850 (1994).
- L. Cohen. *Time Frequency Analysis*. Prentice Hall, New Jersey, 1995.
- R.C. deCharms, D.T. Blake, and M.M. Merzenich. Optimizing Sound Features for Cortical Neurons. *Science* **280**, 1439–1443 (1998).
- G.C. Deangelis, I. Ohzawa, R.D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex: II. Linearity of temporal and spatial summation. *J. Neurophysiol.* **69** (4), 1118–1135 (1993).
- A.J. Doupe. Song- and Order-Selective Neurons in the Songbird Anterior Forbrain and their Emergence during Vocal Development. *J. of Neuroscience* **17** (3), 1147–1167 (1997).

- J.J. Eggermont. Coding of free field intensity in the auditory midbrain of the leopard frog. I. Results for tonal stimuli. *Hearing Research* **40**, 147–166 (1989).
- J.J. Eggermont. The magnitude and phase of temporal modulation transfer functions in cat auditory cortex. *J. Neurosci.* **19** (7), 2780–8 (1999).
- G. Ehret and M.M. Merzenich. Neural Discharge Rate is Unsuitable for Encoding Sound Intensity at the Inferior Colliculus Level. *Hearing Research* **35**, 1–8 (1988).
- G. Ehret, and A.J.M. Moffat. Inferior colliculus of the mouse. III. Responses probabilities and threshold of single units to synthesized mouse calls compared to tone and noise bursts. *J. Comp. Physiol. A* **156**, 637–644 (1985b).
- R.R. Fay and A.N. Popper (Eds). *The Mammalian Auditory Pathway: Neurophysiology*. Springer Verlag, New York, (1992).
- O. Ghitza and J.L. Goldstein. JNDs for the spectral envelope parameters in natural speech. *Hearing Physiological Bases and Psychophysics* (R. Klinke and R. Hartmann Eds.), Springer Verlag, New York, 352–358 (1983).
- K. Gieck. *Engineering Formulas*. McGraw–Hill (1974).
- S.V. Girman, Y. Sauve, and R.D. Lund. Receptive field properties of single neurons in rat primary visual cortex. *J. Neurophysiol.* **82** (1), 301–311, 1999.
- I. Glass and Z. Wollberg. Auditory cortex response to sequences of normal and reversed squirrel monkey vocalizations. *Brain. Behav. Evol.* **22**, 13–21 (1983).
- S. Greenberg. Speaking in Shorthand A Syllabic–Centric Perspective for Understanding Pronunciation Variation. *Proc. Of the ESCA Workshop, Kexrade*, 47–56 (1998).
- D.D. Greenwood. A cochlear frequency position function for several species – 29 years later. *J. Acoust. Soc. Am.* **87**, 2592–2605 (1990).
- J.W. Hall, M.P. Haggard, and M.A. Fernandes. Detection in noise by spectro–temporal pattern analysis. *J. Acoust. Soc. Am.* **76**, 50–56 (1984).
- M.H. Hayes. *Statistical Digital Signal Processing and Modeling*. Wiley & Sons (1996).
- D. J. Hermes, A.M.H.J. Aertsen, P.I.M. Johannesma and J.J. Eggermont. Spectro–temporal characteristics of single units in the auditory midbrain of the lightly anesthetized grass frog (*Rana temporaria* L.) investigated with noise stimuli. *Hearing Research, Elsevier Biomedical Press* **5**, 147–148 (1981).
- T. Houtgast. Auditory–filter characteristics derived from direct masking data and

pulsation threshold data using a rippled noise masker. *J. Acoust. Soc. Am.* **62**, 409–415 (1977).

R.L. Jenison, S. Greenberg, K.R. Klunder, W.S. Rhode. A composite model of the auditory periphery for the processing of speech based filter response functions of single auditory–nerve fibers. *J. Acoust. Soc. Am.* **90** (1), 773–786 (1991).

J.P. Jones and L.A. Palmer. The two dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1187–1211, 1987.

D.J. Klein, D.A. Depireux, J.Z. Simon, and S.A. Shamma. Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. *J. Comp. Neurosci.* **9**, 85–111 (2000).

N. Kowalski, D.A. Depireux, and S.A. Shamma. Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra. *J. Neurophysiol. (Bethesda)* **76** (5), 3524–3534 (1996a).

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

S. Kuwada, R. Batra, T.C.T. Yin, D.L. Oliver, L.B. Haberly, and T.R. Stanford. Intracellular Recordings in Response to Monaural and Binaural Stimulation of Neurons in the Inferior Colliculus of the Cat. *J. Neurosci.* **17** (19), 1565–7581 (1997).

G. Langner and C.E. Schreiner. Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J. Neurophysiol* **60**, 1799–1822 (1988).

J.H. Laning, and R.H. Batin. *Random Process in Automatic Control*. McGraw Hill, New York (1956).

C.E. Liberman. The cochlear frequency map of the cat: Labeling auditory–nerve fibers of unknown characteristic frequency. *J. Acoust. Soc. Am.* **72**, 1441–1449 (1982).

L. Ljung. *System Identification: Theory for the User*. Prentice Hall, New Jersey (1987).

D. Margoliash. Acoustic parameters underlying the response of song–specific neurons in the white–crowned sparrow. *J. Neurosci.* **3**, 1039–1057 (1983).

D. Margoliash, and E.S. Fortune. Temporal and harmonic combination–sensitive neurons in the zebra finch’s Hvc. *J. Neurosci.* **12**, 4309–4326 (1992).

P.Z. Marmarelis and V.Z. Marmarelis. *Analysis of Physiological Systems Modeling. The White Noise Approach*, Plenum Press, New York, 1978.

- P.Z. Marmarelis and K.I. Naka. Identification of Multi-Input Biological Systems. *IEEE Transactions on Biomedical Engineering*, **21** (2), 88–101 (1974).
- J.R. Mendelson, C.E. Schreiner, M.L. Sutter, K.L. Grasse. Functional topography of cat primary auditory cortex: Responses to frequency-modulated sweeps. *Experimental Brain Research* **94** (1), 65–87 (1993.).
- I. Nelken, P.J. Kim, E.D. Young. Linear and nonlinear spectral integration in type IV neurons in the dorsal cochlear nucleus. II. Predicting responses with the use of nonlinear models. *J. Neurophysiol.* **78**, 800–811 (1997).
- I. Nelken, Y. Rotman, and O.B. Yosef. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* **37**, 154–157 (1999).
- K.K. Ohlemiller, J.S. Kanwal, N. Suga. Facilitative responses to species-specific calls in the cortical FM-FM neurons of the mustache bat. *NuroReport.* **7**, 1749–1755 (1996).
- J.F. Olsen and N. Suga. Combination sensitive neurons in the medial geniculate body of the mustache bat: encoding of target range information. *J. Neurophysiol.* **65**, 1254–1274 (1991a).
- J.F. Olsen and N. Suga. Combination sensitive neurons in the medial geniculate body of the mustache bat: encoding of relative velocity information. *J. Neurophysiol.* **65**, 1275–1296 (1991b).
- R. Plomp. Pitch of complex tones. *J. Acoust. Soc. Am.* **41**, 1526–1533 (1967).
- R. Plomp. Timbre as a multidimensional attribute of complex tones. *Frequency analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G.F. Smoorenburg, Sijthoff Linden (1970).
- R. Plomp. The Role of Modulations in Hearing. *Hearing Physiological Bases and Psychophysics* (R. Klinke and R. Hartmann Eds.), Springer Verlag, New York, 270–276 (1983).
- D. Ploog. Neurobiology of Primate Audio-Vocal Behavior. *Brain Research Reviews* **3**, 35–61 (1981).
- L.C.W. Pols, L.J.T. Kamp, and R. Plomp. Perceptual and physical space of vowel sounds. *J. Acoust. Soc. Am.*, **46**, 458–467 (1969).
- K.A. Razak, Z.M. Fuzessery, T.D. Lohuis. Single cortical neurons serve both echolocation and passive sound localization. *J. Neurophysiol.* **81**, 1438–1442 (1999).

- A. Rees and A.R. Moller. Response of neurons in the inferior colliculus of the rat to AM and FM tones. *Hear. Res.* **10**, 301–330 (1983).
- D.S. Reich F. Mechler, K.P. Purpura, J.D. Victor. Interspike intervals, receptive fields, and information encoding in primary visual cortex. *J. Neurosci.* **20** (5), 1964–1974 (2000).
- F. Rieke, D.A. Bodnar, and W. Bialek. Naturalistic Stimuli Increase the Rate and Efficiency of Information Transmission by Primary Auditory Fibers. *Proc. R. Soc. Lond.* **262**, 259–265 (1995).
- C.E. Schreiner, J.V. Urbas, S. Mehrgardt. Temporal resolution of amplitude modulation and complex signals in the auditory cortex of the cat. *Hearing—Physiological bases and Psychophysics*, R. Klinke and R. Hartmann (eds.). New York, Springer-Verlag, 169–175 (1983).
- C.E. Schreiner, and B.M. Calhoun. Spectral envelope coding in the cat primary auditory cortex: Properties of ripple transfer function. *Auditory Neurosci.* **1**, 39–61, 1994.
- M.R. Schroder, D. Gottlob, and K.F. Siebrasse. Comparative study of European concert halls: correlation of subjective preferences with geometric and acoustic parameters. *J. Acoust. Soc. Am.* **56**, 1192–1201 (1974).
- S.M. Smirnakis, M.J. Berry, D.K. Warland, W. Bialek, M. Meister. Adaptation of retinal processing to image contrast and spatial scale. *Nature* **386** (6620), 69–73 (1997).
- J.W.T. Smolders, A.M.H.J. Aertsten, and P.I.M. Johannesma. Neural representation of the acoustic biotope: A comparison of the response of auditory neurons to tonal and natural stimuli in the cat. *Biological Cybernetics* **35**, 11–20 (1979).
- J.E. Spiro, M.B. Dalva, and R. Mooney. Long-Range Inhibition Within the Zebra Finch Song Nucleus RA Can Coordinate the Firing of Multiple Projection Neurons. *J. Neurophysiol. (Bethesda)* **81** (6), 3007–3020 (1999).
- N. Suga, W.E. O'neil, and T. Manabe. Cortical neurons sensitive to particular combinations of information bearing elements of bio-sonar signals in the mustache bat. *Science* **200**, 778–781 (1978).
- N. Suga, J.A. Simmons, and P. Jen. Peripheral specialization for fine frequency analysis of Doppler shifted echoes in the CF-FM bats, *Pteronotous parnellii*. *J. Exp. Biol.* **63**, 161–192 (1975).
- N. Suga and P. Jen. Disproportionate tonotopic representation for processing of specific CF-FM sonar signal in the mustache bat auditory cortex. *Science* **194**, 542–544 (1976).

Theunissen, F.E., Sen, K. & Doupe, A.J. Spectro-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. of Neurosci.* **20** (6), 2–17 (2000).

T.M. Van Veen and T. Houtgast. On the Perception of the Spectral Envelope. *Hearing Physiological Bases and Psychophysics* (R. Klinke and R. Hartmann Eds.), Springer Verlag, New York, 277–281 (1983).

T.M. Van Veen and T. Houtgast. Spectral Sharpness and Vowel Dissimilarity. *J. Acoust. Soc. Am.* **77** (2), 628–634 (1985).

J.D. Victor and K.P. Purpura. Spatial phase and the temporal structure of the response to gratings in V1. *J. Neurophysiol.* **80** (2), 554–571, (1998).

R.V. Voss and J. Clarke. 1/f noise in music and speech. *Nature* **258**, 317–318 (1975).

X.Wang, M. Merzenich, R. Beitel, and C.E. Schreiner. Representation of Species-Specific Vocalization in the Primary Auditory Cortex of the Common Marmoset: Temporal and Spectral Characteristics. *J. Neurophysiol.* **24**, 6, 2685–2706 (1995).

P. Winter and H.H. Funkenstein. The effects of species-specific vocalizations on the discharge of auditory cortical cells in the awake squirrel monkey (*Saimiri sciureus*). *Exp. Brain. Res.* **18**, 489–504 (1973).

Y. Yeshurun, Z. Wollberg, and N. Dyn. Identification of MGB cells by Volterra kernels. II Towards a functional classification of cells., *Biol. Cybern.* **56**, 203–208 (1987).

T.C.T. Yin, J.C.K. Chan, and D.R.F. Irvine. Effects of interaural time delays of noise stimuli on low-frequency cells in the cat inferior colliculus. I. Responses to wideband noise. *J. Neurophysiol.* **55**, 280–300 (1986).

E.D. Young, and W.F. Brownell. Responses to tones and noise of single cells in the dorsal cochlear nucleus of unanesthetized cats. *J. Neurophysiol.* **39**, 282–300 (1976).

E.D. Young. What's the Best Sound. *Science* **280**, 1402–1403 (1998).

Nonlinear Spectro–Temporal Processing and Feature Selectivity

Abstract

Neurons in the central auditory system of acoustically specialized animals, such as bats and songbirds, are often highly nonlinear and specialized for processing specific aspects of behaviorally important sounds (Suga and Jen 1976; Suga, O'neil, Manabe 1978; Margoliash 1983; Margoliash and Fortune 1992; Olsen and Suga 1993a 1993b; Casseday, Ehrlich, and Covey 1994; Doupe 1997). In most mammals, including cats and monkeys, a similar link between acoustic sound structure, neuronal processing, and behavior has not been established. Consequently ecological paradigms have not been exploited to study their auditory system. Alternative methods and conventional stimuli have not revealed equivalent higher-order neuronal processes for sound encoding in such mammals. Using synthetic acoustic stimuli, that incorporate some basic features common to a wide range of natural sounds, we demonstrate that reverse correlation methods can reveal nonlinear neuronal response classes that can not be identified with conventional reverse correlation stimuli. Neuronal recordings in the Inferior Colliculus of cats reveals three distinct response classes. One class of neurons showed nearly linear response characteristics resembling simple cells in the visual cortex. Two additional response classes were distinctly nonlinear: one type of neuron showed selective responses that are not time-locked to the stimulus spectro-temporal envelope, resembling visual complex cells. The other neuronal class responded exclusively to specific spectro-temporal sound components, resembling feature sensitive cells of the bat auditory systems. These findings indicate that the mammalian auditory midbrain, in general, contains neuronal specializations for higher-order sound processing that have previously been considered to emerge at the cortical level and only in acoustically specialized animals.

3.1 Introduction

A fundamental requirement of auditory processing is the extraction and decomposition of spectral and temporal information. Identification of a complex acoustic signal is highly dependent on the analysis of the time-varying spectrum. Natural sounds, such as human speech and animal vocalizations, are characterized by a structurally rich and complex time-varying spectrum. This is most evident in the spectrographic decomposition of natural sounds that is performed by the cochlea (Sachs and Young 1979; Delgutte and Kiang 1984; Shamma 1985; Carney and Geisler 1986; Geisler and Gamble 1989) and, consequently, serves as inputs to higher-order processing stations in the brain. Furthermore, natural sounds are characterized by a number of spectro-temporal correlations (Voss and Clarke 1975; Attias and Schreiner 1999a; Nelken, Rotman, and Yosef 1999; chapter 1) all of which play important roles in perception (Plomp 1967, 1970, 1983; Pols *et al.* 1969) and neuronal encoding (Plomp 1983; Attias and Schreiner 1999b; Nelken, Rotman, and Yosef 1999). These take the form of spectral resonances, temporal modulations, comodulation, and time-varying frequency conjunctions.

Despite the structural complexity of natural sounds and its presumed importance for natural sound processing, much of central auditory neuroscience has proceeded by using stimuli that are structurally simple, lacking many of the structural features characteristic of natural sounds. Furthermore, the most widely used stimuli consist of single frequency component which excite a small portion of the auditory neuronal network, a scenario not common for natural sounds. Due to the general nonlinear nature of the auditory system it has become apparent that structurally simple stimuli can not be

used directly for identifying neuronal mechanisms which the auditory system uses for natural sound analysis (Nelken and Yosef 1998; Nelken *et al.* 1999; Theunissen *et al.* 2000).

Recent methodological advances have demonstrated that complex natural sounds can be used directly to identify nonlinear auditory neurons in the avian forebrain using spectro-temporal receptive field (*STRF*) methods (Theunissen *et al.* 2000). The *STRF* has a long history in auditory neurophysiology (Aertsen *et al.* 1980 1981; Hermes *et al.* 1981; Yeshurun, Wollberg, and Dyn 1987; Nelken *et al.* 1997; deCharms *et al.* 1998; Theunissen *et al.* 2000; Klein *et al.* 2000) and has been shown to have numerous advantages over more conventional stimulus-response characterizations (Eggermont 1993; Klein *et al.* 2000). This method allows one to estimate the linear processing capabilities of auditory neurons for complex and spectro-temporally rich stimulus ensembles. Unlike conventional methods, which break up the stimulus-response function into a separable combination of spectral and temporal components (Schreiner, Urbas, and Mehrgardt 1983; Langner and Schreiner 1988; Schreiner and Langner 1988;), the *STRF* allows one to jointly estimate the spectro-temporal preferences without any a priori assumptions about separability and independence of the stimulus-response relationship. Despite the attractiveness of these methods, the *STRF* method by itself is largely limited in that it has been associated with quantifying only "linear" response characteristics (Young 1998). Although more elaborate methods can be used to identify the presence of response nonlinearities (Theunissen *et al.* 2000), it is not clear how such methods can be used for parsing out higher-order response attributes and identifying the nature of the underlying nonlinearities. Thus it still remains to be seen whether this approach can be

extended for systematically identifying and parsing out complex nonlinearities, such as those arising from higher-order stimulus attributes (e.g. contrast and high-order spectro-temporal correlations).

The most direct approach for parsing out response nonlinearities is to use simple stimuli while systematically changing some stimulus parameter (e.g. intensity, modulation depth, etc.) (Rees and Møller 1983; Møller and Rees 1986; Rees and Møller 1987; Rees and Palmer 1988; Rees and Palmer 1989; Krishna and Semple 2000). Although this approach is useful, it is essentially limited to low-order nonlinearities that are activated by low-order features of the sound waveform (e.g. intensity, modulation depth, carrier frequency etc.). Furthermore, given the complex structure of natural sounds and the general dependence of neuronal responses to a number of response parameters these results are not easily extended to more complex stimulus scenarios. To date, more complex and possibly relevant nonlinearities, related to higher-order spectro-temporal correlations and structural components of natural sound have not been described in detail (Nelken *et al.* 1999).

To overcome such limitations, we employ an alternate and more direct approach for analyzing spectrographic response nonlinearities. Motivated by understanding how arbitrary natural sounds and vocalizations are represented, we designed a set of structurally rich ripple noise sounds that emulate some basic statistics of natural sounds. The relative degree of instantaneous coherency of these sounds was systematically altered while controlling for the overall statistics of these sounds. Using spectro-temporal reverse correlation methods, it is shown that midbrain neurons are highly adapted to analyze specific aspects of the stimulus spectro-temporal envelope. We

identify a number of higher-order nonlinearities related to the degree of local coherency of the sound and use these to classify neuronal populations in the inferior colliculus of the cat. The presented findings demonstrate the presence of distinct spectro-temporal nonlinearities while identifying possible mechanisms used for complex sound analysis in the inferior colliculus.

3.2 Methods

Experimental Methods: Data was obtained from $n=84$ single units in the central nucleus of the inferior colliculus (ICC) of three ketamine (10 mg/kg) and diazepam (0.5 mg/kg) anesthetized cats. The ICC was exposed by removing the overlying cerebrum and part of the bony tentorium using a dorsal approach. Electrode penetration trajectories were at 45° relative to the sagittal plane and approximately orthogonal to the isofrequency band lamina. Spike trains were recorded using parylen coated tungsten electrodes (1–3 M Ω at 1kHz) onto a digital audio tape (Cygnus CDAT16) at a sampling rate of 24.0 kHz (41.7 μ sec resolution) for off line analysis.

Stimuli were presented binaurally with an independent sound sequence for each ear. This allowed us to compute independent *STRFs*, *RTFs*, and conditional-response histograms for the contralateral and ipsilateral ears. Single neurons and/or clusters of neurons were isolated audio-visually by presenting pure tones and/or white noise. The dynamic ripple stimulus was presented for a period of 20 minutes, followed by 18 minutes of the ripple noise at 30–70 dB above the neurons response threshold. Both sounds were presented at identical intensities. For 6 neurons that did not respond to the ripple noise stimulus, the dynamic ripple stimulus was again presented at the end of the

recording session to verify that the given neurons were still responsive and to verify the stability of the electrode placement. For 66% of the recording sites, a five- or seven-second segment (repeated 40 or 100 times respectively) of the dynamic ripple and ripple noise were also played at the end of the recording sessions. All experiments were conducted in an acoustically sealed sound chamber (IAC). All surgical methods and experimental procedures were approved by the committee on animal research, UCSF.

Off line analysis consisted of digital bandpass filtering (0.3–10 kHz) all spike trains and individually spike sorting the action potential traces using a Bayesian spike sorting algorithm (Lewicki 1994) before computing *STRFs*, *RTFs* and conditional-response histograms.

Acoustic Stimuli: Detailed description of the ripple noise and dynamic ripple acoustic stimuli are provided in chapter 2. A brief description is provided here for convenience. The acoustic stimulus time waveform, $s(t)$, is generated via a bank of 230 frequency carriers (linearly spaced on an octave frequency axis) which are individually amplitude modulated by the dynamic ripple or the ripple noise spectro-temporal envelopes (Figs. 2 and 3; chapter 2). Mathematically the acoustic waveform for this general class of signals is expressed as $s(t) = \sum_k S_{Lin}(t, f_k) \sin(2\pi f_k t + \phi_k)$ where ϕ_k is a randomly chosen phase ($0 - 2\pi$), f_k correspond to the frequency carrier elements (spacing resolution of 0.0231 octaves) and $S_{Lin}(t, f_k)$ corresponds to the linear amplitude spectro-temporal envelope of the dynamic ripple or the ripple noise stimulus.

The dynamic ripple spectro-temporal envelope is designed as a dynamic

sinusoidal grating on a octave frequency and decibel amplitude axis. It is expressed as

$S_{MR}(t, X_k) = M/2 \cdot \sin(2\pi \Omega(t) X_k + \Phi(t))$ where the decibel amplitude spectro-temporal envelope, $S_{MR}(t, X_k)$, is related to the linear amplitude spectro-temporal envelope by $S_{MR}(t, X_k) = 20 \log_{10}(S_{Lin}(t, X_k)) + M/2$ (note that S_{Lin} is bounded between zero and one). Here $X_k = \log_2(f_k/500)$ is an octave frequency axis (the frequency variable f_k doubles with every octave above 500 Hz), $M=30$ or 45 is the modulation depth of the envelope in decibels, $\Omega(t)$ is the time varying ripple density which determines the number of sinusoidal peaks per unit octave, and

$\Phi(t) = 2\pi \int_0^t F_m(\tau) d\tau$ controls the time varying temporal modulation rate, $F_m(t)$, of the envelope. Both parameters are independent, time varying (bandlimit 3 dB frequency of 1.5 Hz for F_m and 3.0 Hz for Ω) uniformly distributed stochastic processes ($\Omega=0$ to 4 cycles / octave and $F_m=-350$ to 350 Hz).

The ripple noise envelope is generated as a superposition of $L=16$ independently chosen dynamic ripple envelopes, $S_l(t, X_k)$,

$$S_{RN}(t, X_k) = f \left[\frac{1}{\sqrt{L}} \sum_{l=1}^L S_l(t, X_k) \right] \quad (3.1)$$

where the amplitude distribution is compressed (using a contrast transformation

$f(x) = M/2 \cdot \text{erf}(x/\sigma_{DR})$, where $\text{erf}(\cdot)$ is the error function) so that it covers a range of $M=30$ or 45 dB uniformly in a similar manner as the dynamic ripple envelope. Thus, both sounds probe the same intensity operating range, allowing one to isolate spectro-temporal nonlinearities from intensity or contrast dependent ones.

Spectro-Temporal Receptive Field (STRF): STRFs are computed by averaging the pre-event spectro-temporal envelope. For a sequence of N neural events at times, t_n (sampled at $41.7 \mu\text{sec}$ resolution), the STRF is obtained as

$$STRF(\tau, X_k) = 1/(\sigma_s^2 \cdot T) \sum_n S(t_n - \tau, X_k) \quad (3.2)$$

where T is the experimental recording time in seconds, τ is the temporal delay of the stimulus relative to the neural event time, and σ_s^2 is the variance of the decibel spectro-temporal envelope $S(t, X_k)$ for the moving ripple or the ripple noise stimulus. Both envelopes were sampled at a rate of 4.0 ksamples/sec (temporal) and 43 samples/octave (spectral). The STRF is formally given in units of spikes / second / decibel. We use a normalized version of the STRF, $\sigma_s \cdot STRF(\tau, X_k)$, which corresponds to the mean difference output produced at time zero, in units of spikes/second, for the average differential stimulus (units of dB) presented within the receptive field. The statistically significant portion of the STRF ($p < 0.002$) is obtained by keeping all values of the STRF which satisfy

$|\sigma_s^2 \cdot T \cdot STRF(\tau, X_k) / \sqrt{N}| > 3.09 \sigma_s$, and setting all other values to zero. No smoothing

was performed prior to or after thresholding. For further detail refer to section 3.4.

Ripple Transfer Function (RTF): The statistically significant ripple transfer function is obtained directly from the significant *STRF* ($p < 0.002$) by applying a two-dimensional Fourier transformation: $RTF(F_m, \Omega) = |\mathfrak{F}_2\{STRF(\tau, X_k)\}|$ where $\mathfrak{F}_2\{\cdot\}$ designates the two-dimensional Fourier transform and $|\cdot|$ is the magnitude. This transformation converts the *STRF* into a Fourier domain transfer function. Upon performing this transformation, the time axis of the *STRF* is transformed into a temporal modulation rate axis (F_m) and the frequency axis (given in units of octaves) to a ripple density axis (Ω). Refer to section 3.14 for further detail.

Conditional-Response Histogram: The conditional-response histogram provides a measure of the number of responses for each parameter combination (Ω and F_m) (dynamic ripple only). For a given neuronal response (spike) at time t_n , the response histogram is computed by determining the instantaneous dynamic ripple parameters $\Omega(t_n)$ and $F_m(t_n)$ at the time of the neuronal event. The bin of the response histogram corresponding to that range of parameters is then incremented by +1 (Fig. 19). The exact position used to estimate the parameters relative to the neuronal spike time, t_n , did not alter the resulting histogram (tested for a time lag of 0–50 ms) since the parameters vary in time at a slow rate (1.5 Hz and 3 Hz) compared to the integration time

of ICC neurons (tens of milliseconds). This measure differs from the *RTF* since it does not require that the neuronal response be precisely aligned with respect to the fast temporal modulations of the stimulus envelope (up to 350 Hz) (recall that the *RTF* is derived from the *STRF* which requires precise time-locking). The response histogram provides an estimate of the conditional distribution $p(F_m, \Omega | t_k)$ from the time varying dynamic ripple parameters $\Omega(t)$ and $F_m(t)$. This conditional distribution function describes the likelihood of a given set of parameters, given the occurrence of a single spike. See section 3.15.

Null hypothesis: The relative degree of nonlinearity is tested against the expected response characterizations for a linear neuron. Since the ripple noise and dynamic ripple both have an identical impulsive autocorrelation function (Chapter 2; Sections 2.15 and 2.18), a hypothetical linear neuron would produce identical *STRFs* and *RTFs* for these sounds (Chapter 2; Section 2.4; Eq. (2.5)). Lack of or statistically significant differences in these descriptors for either sound are indicative of response nonlinearities.

3.3 Binaural Receptive Fields

The standard procedure for evaluating spectro-temporal receptive fields was extended by using a two-input reverse correlation procedure that allows one to jointly estimate *STRFs* bilaterally. Using this procedure response sensitivities for the contralateral and ipsilateral ears could be estimated within a single trial of the experiment. The devised method consists of presenting independent spectro-temporal

sound sequences bilaterally. After recording neuronal responses, these two inputs are then used to construct *STRFs* for each ear independently using the standard reverse correlation procedure. A similar scheme was first described by Marmarelis (1974) for computing Wiener kernels (for the more general case of a multi-input multi-output system) and has since been employed for characterizing binocular sensitivities of visual cortex neurons (Anzai *et al.* 1999).

For all recording locations, sounds were presented ototically via speakers that were inserted into the acoustic meatus – each ear stimulated by an independent sound sequence. Across-channel independence was forced on the dynamic ripple and ripple noise stimuli by choosing statistically independent contralateral and ipsilateral sounds. This was achieved by choosing independent carrier phase components for Eq. (2.15) (chapter 2) and independent ripple density and modulation rate parameters between the two ears. The first order stimulus crosscorrelation and the spectro-temporal crosscorrelation between the contralateral and ipsilateral stimuli thus satisfy

$$r_{ci}(\tau) = E[s_c(t)s_i(t+\tau)] = 0 \quad (3.3)$$

$$R_{ci}(\tau, \zeta) = E[S_c(t, X)S_i(t+\tau, X+\zeta)] = 0$$

were the subscripts *c* and *i* designate the contralateral and ipsilateral ears respectively.

Given the independence of the contralateral and ipsilateral inputs, the standard procedure for computing *STRFs* can then be employed. Binaural *STRFs* were derived as

$$STRF_c(\tau, X_k) = 1/(\sigma_s^2 \cdot T) \sum_n S_c(t_n - \tau, X_k) \quad (3.4)$$

$$STRF_i(\tau, X_k) = 1/(\sigma_s^2 \cdot T) \sum_n S_i(t_n - \tau, X_k) .$$

Note that the contralateral and ipsilateral *STRFs* are obtained by crosscorrelating the output spike train with different input channels.

Example binaural *STRFs* are provided in Fig. 1. Binaural sensitivities were largely varied and showed spectro-temporal response patterns that are consistent with classical definitions of binaurality (EE, EI, IE, EO, OE etc.) (Goldberg and Brown 1969). These include neurons which show a strong excitatory component for one ear and inhibitory component for the adjacent ear (IE; Fig. 1 C/D) as well as neurons that have similar excitatory response patterns for both ears (EE; Fig. 1 A/B and E/F). Neurons can also show strong response components to one ear and none for the opposite ear (EO; Fig. G/H). A handful of neurons alternately showed response components that are highly dissimilar in shape across the contralateral and ipsilateral ears.

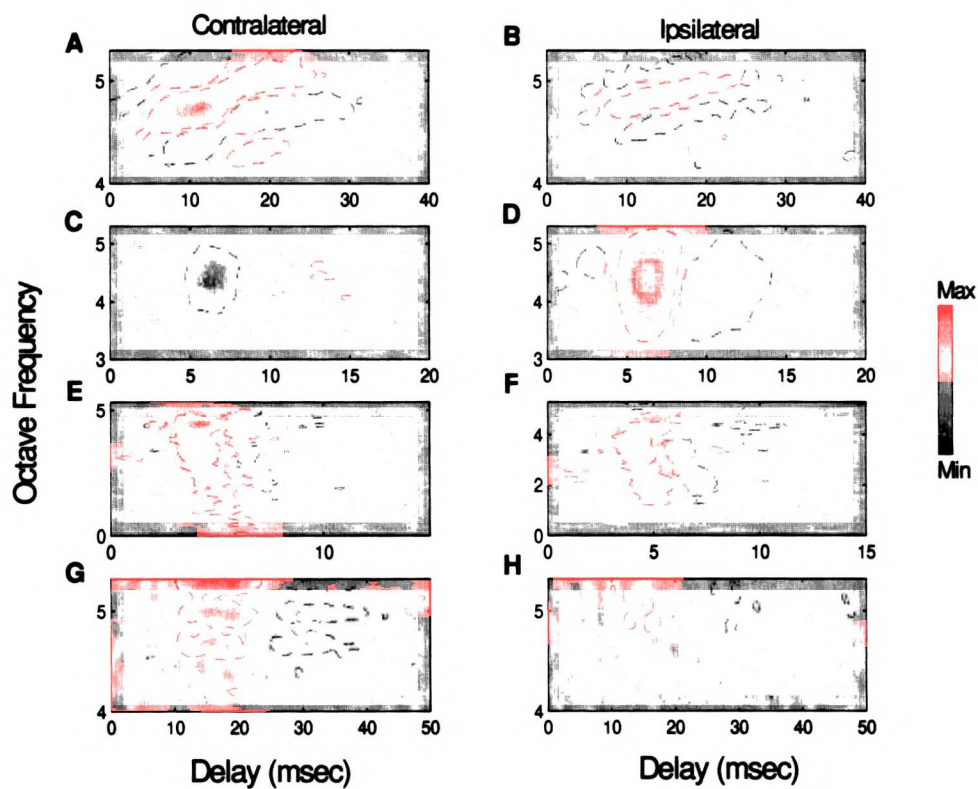


Figure 1: Example binaural receptive field patterns. The vast majority of neurons showed binaural response patterns with similar *STRFs* across both ears (A/B, E/F). Other neurons had binaural response patterns with a strong excitatory component in one ear and strong inhibitory component in the adjacent ear (C/D). Some neurons had no obvious binaural response, responding only to the contralateral or the ipsilateral ear. All neurons are shown using the same color scale for the contralateral and ipsilateral ears.

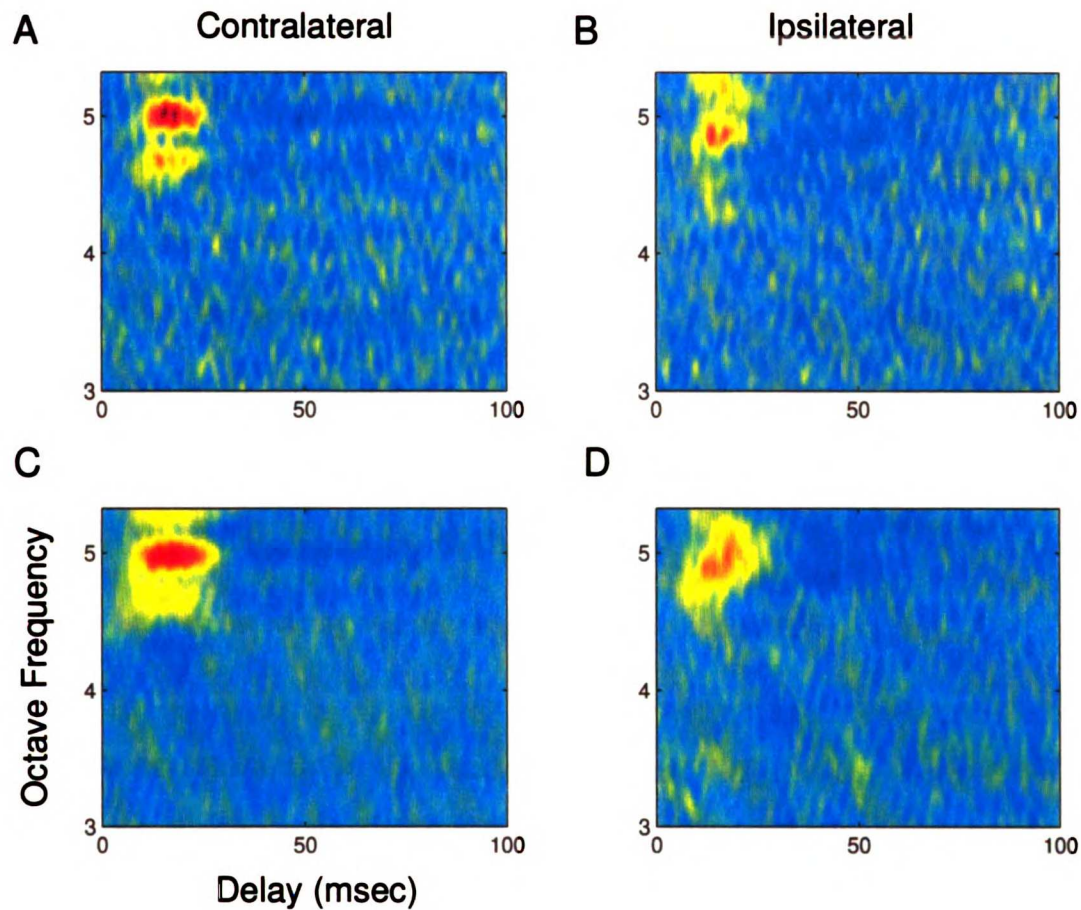


Figure 2: The contralateral (A) and ipsilateral (B) *STRFs* obtained during bilateral stimulation are effectively identical to the contralateral (C) and ipsilateral (D) *STRFs* obtained for monaural stimulation at identical intensity (80 dB SPL).

For six neurons, a series of controls were performed to assure that *STRFs* with bilateral stimulation were identical to those obtained with monaural stimulation, thus verifying that this procedure is experimentally and theoretically sound. Dynamic ripple and/or ripple noise stimuli were first presented bilaterally at 20–80 dB above neuron’s response threshold. The corresponding sound sequences was then presented monaurally, either for the contralateral ear or for the ipsilateral ear, at an identical intensity. For one

of the six neuron both the contralateral and ipsilateral ear were tested monaurally. Example result for these controls is shown for the neuron in Fig. 2. The obtained contralateral and/or ipsilateral *STRFs* for binaural and for monaural stimulation were generally identical in shape and differed only in magnitude, suggesting that the binaural stimulation procedure does not introduce any artifacts. Similar results were obtained for all neurons tested with this control.

3.4 Testing For Significance of the STRF

A procedure for testing significance of the computed *STRFs* was derived in closed form.

Given a neuronal response of N spikes and the measured *STRF*, the statistically significant portion of the *STRF* was derived by considering a null condition in which a set of N random spikes is put through Eq. (3.2). Since the input spectro-temporal envelope for the dynamic ripple and ripple noise have uniformly distributed amplitudes, the test for significance can be derived by randomly summing independent identically distributed (iid) random variables.

Consider the uniformly distributed iid random variable $X_n \in u[-M/2, M/2]$ where u designates an uniform distribution random variable in the chosen interval.

The amplitude distribution for a random spike train *STRF* is given by $p_{Y_N}(y)$ where

$$Y_N = 1/(\sigma_s^2 \cdot T) \sum_{n=1}^N X_n = \sum_{n=1}^N Z_n, \quad N \text{ is the number of spikes considered for the}$$

significance test, and $Z_n \in u[-M/(2\sigma_s^2 \cdot T), M/(2\sigma_s^2 \cdot T)]$ is an uniformly distributed

random variable. The distribution for Y_N is derived by convolving the uniform distribution of Z_n N times (Ross 1993). Recursively this is expressed as

$$p_{Y_t}(y) = \int p_{Y_{t-1}}(x-y) p_Z(x) dx \quad (3.5)$$

where $p_Z(x)$ is the amplitude distribution of Z_n . Upon modifying a formula provided by Chui (Chui 1992, pg. 84) the amplitude distribution for the null condition is given by

$$p_{Y_n}(y) = b_N\left(y\sigma_s^2 T/M + N/2\right) = \frac{1}{(N-1)!} \sum_{k=0}^N (-1)^k \binom{N}{k} \left[\left(\frac{y\sigma_s^2 T}{M} + \frac{N}{2} \right) - k \right]_+^{N-1} \quad (3.6)$$

where $b_N(x)$ is the N^{th} order B -spline function (Chui 1992; Roark and Escabí 1999) and $[x]_+ = \max(0, x)$. For a given significance probability, p , the significant portion of the $STRF$ is found by considering the $STRF$ values which exceed a the two tail significance test (Zar 1999). Thus we seek to find the threshold value, t , which satisfies

$$\int_{-t}^t p_{Y_n}(x) dx = 1 - p \quad (3.7)$$

The significant $STRF$ is immediately provided by finding all amplitude values which exceed this threshold value

$$|STRF(\tau, X)| > t \quad (3.8)$$

The search for t is simplified by noting that for sufficiently large values of N the amplitude distribution for Y_N approaches a normal distribution with mean zero and

$$\sigma_{Y_N} = \sqrt{N}/(T \cdot \sigma_s) \quad . \quad \text{Using this approximation the significant } STRF \text{ is obtained by}$$

finding all amplitude values which satisfy

$$|\sigma_s^2 \cdot T \cdot STRF(\tau, X_k) / \sqrt{N}| > 3.09 \sigma_s \quad (3.9)$$

for a p value of 0.002. Although most neurons had reasonably large values of N (usually thousands to tens of thousands), a significant number of neurons had low spike rates with spike counts of $N < 1000$ spikes. We need to determine if this approximation is sufficient for such low values of N . We verify this by plotting the normal approximation against the actual distribution for Y_N for a value of $N=50$. The two curves are visually indistinguishable (Fig. 3 A) and differ only at the extremities (Fig. 3 B) . For a chosen significance value of $p < 0.002$, this approximation yields an actual significance value of $p < 0.0019$ for $N=50$. Thus this approximation is valid for all neurons tested which in all instances had spike counts greater than 50 spikes.

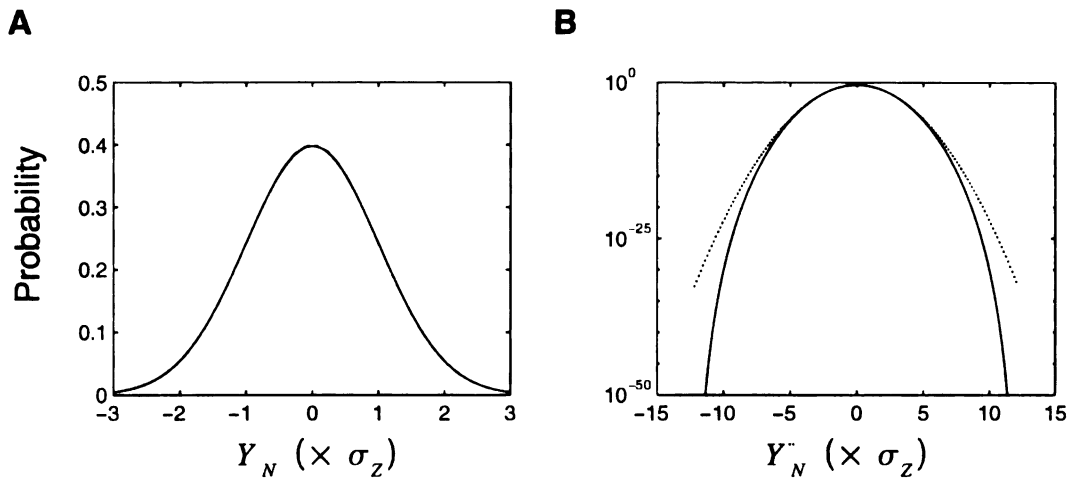


Figure 3: Testing for significance of the *STRF*. The statistically significant *STRF* is estimated by performing a two tail probability test against a null condition of randomly chosen spikes. The null *STRF* amplitude probability distribution function for $N=50$ spikes (dotted line) and for the ideal case ($N = \infty$ spikes, continuous line). On a linear probability axis (A) the two distributions are indistinguishable. The distribution for Y_{50} , however, has finite support and differs from $p_{Y_\infty}(y)$ at the extremities (shown on logarithmic probability axis, B).

3.5 STRF Comparisons – Moving Ripple versus Ripple Noise

As a means to identify nonlinear response components of ICC neurons, the spectro-temporal receptive field (*STRF*), ripple transfer function (*RTF*), and conditional response histogram were computed for the moving ripple and ripple noise stimulus. To avoid intensity dependent nonlinearities, both sounds were presented at identical intensities and contrast. The relative degree of nonlinearity was determined for all neurons by comparing responses to the structured dynamic ripple and the unstructured ripple noise stimuli. Since both of these stimuli have identical low-order statistics (e.g.,

intensity, global spectro-temporal autocorrelation function, and contrast) it is expected that a hypothetical linear neuron produce identical *STRFs* and transfer functions for these two sounds. Using this test as a null hypothesis, significant response differences between the two stimuli therefore result from nonlinear response components which are activated by the higher-order stimulus statistics (see chapter 2.4 and 2.5).

Example *STRFs* for the dynamic ripple and the ripple noise stimulus are shown in Figs. 4 and 5. By comparing responses to the dynamic ripple and ripple noise it was possible to identify two classes of spectro-temporal feature selectivity based on the relative strength of responses to either of these two sounds. The first class of neuron (52%), (s-cells) generally had high firing rates to both sounds (mean spike rate: 12.3 spikes/s for dynamic ripple and 11.4 spikes/s for ripple noise) and had quasi-linear response characteristics. That is, such neurons responded to the locally unstructured ripple noise much as they did to the structured dynamic ripple, thereby producing similar *STRFs*.

Examples for this type of neurons are shown in Fig. 4. In all instances, slight differences in mean spike rate and *STRF* energy were observed for these neurons. The shape of the *STRF* for either condition and its overall strength, however, were similar. As depicted in Fig. 4 all neurons showed similar spectro-temporal patterns. Neurons are shown on identical colorscales for the dynamic ripple and ripple noise stimulus (for ease of comparison). The presence of a well defined statistically significant *STRF* ($p < 0.002$) indicates that s-neurons respond in a phase-locked fashion to the stimulus spectro-temporal envelope. The fact that similar *STRFs* are produced by the ripple noise and dynamic ripple stimulus further demonstrates that these neurons behave quasi-linearly.

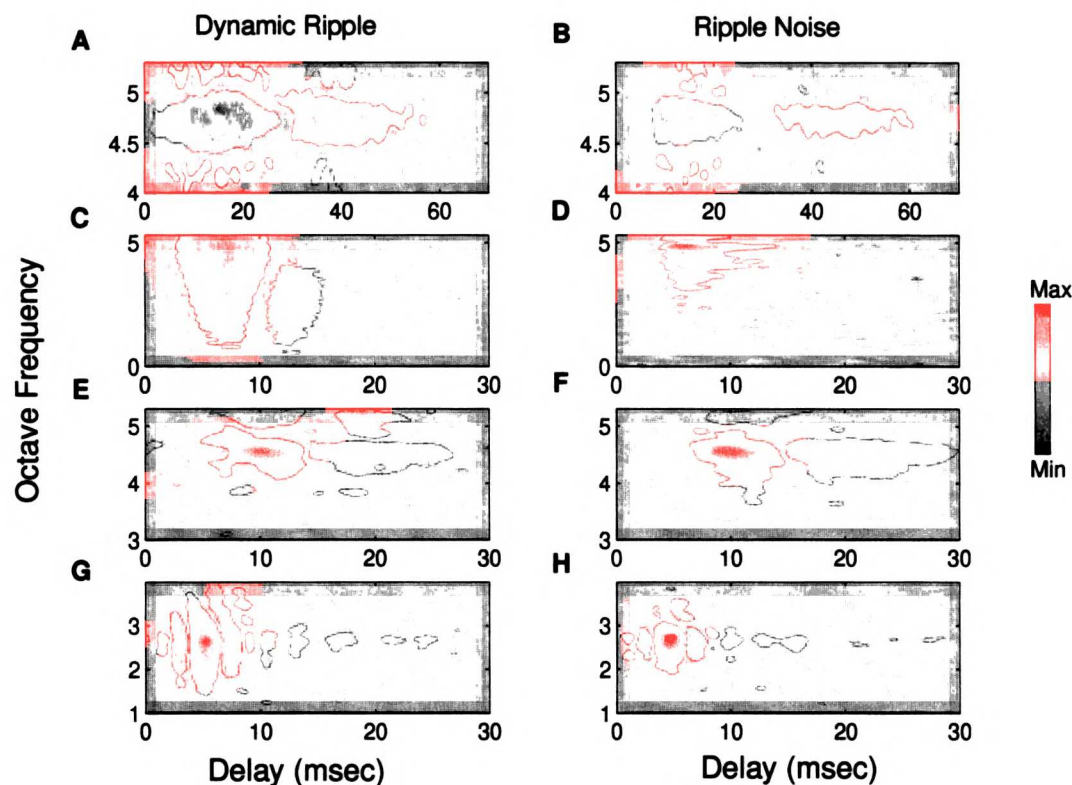


Figure 4: Spectro-temporal receptive fields for neurons that showed quasi linear response characteristics (s-neurons). All neurons were tested with the dynamic ripple (left column, A,C,E,G) and the ripple noise (right column, B,D,F,H) stimulus. Receptive fields have similar shapes and strength as determined by the color scale. For ease of comparison, all neurons are shown with the same color scale for the ripple noise and the dynamic ripple conditions. Significant patterns of the *STRF* are denoted by red contours ($p < 0.002$ contour). The neurons of A, E, G showed almost identical receptive field patterns for both conditions. The neuron of C has a common excitatory area for the dynamic ripple and ripple noise but is missing the post-excitatory suppression for the ripple noise stimulus.

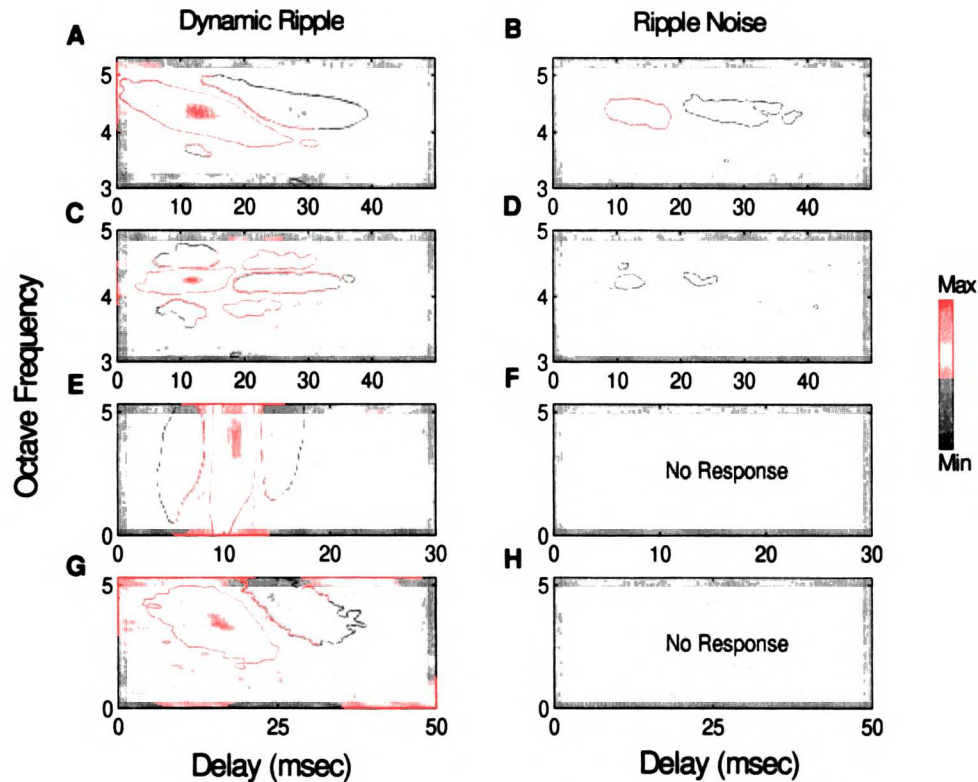


Figure 5: spectro-temporal receptive fields for neurons that responded specifically to the dynamic ripple sound (left column, A,C,E,G) but responded weakly or had no response to the ripple noise stimulus (right column, B,D,F,H). Significant *STRF* patterns are denoted by red contours. The neuron of A responded to both the moving ripple (Rate=1.1 spikes/sec) and the ripple noise stimulus (Rate=0.41 spikes/sec). The *STRF* for this neuron has similar shape for the dynamic ripple and ripple noise sounds ($SI_{DR,RN}=0.78$) but is significantly weaker in its magnitude for the ripple noise stimulus ($ASI=798\%$). The neuron of C responded to both sounds but its response to the dynamic ripple was significantly stronger (1.4 spikes/sec vs. 0.2 spikes/sec) and the ripple noise *STRF* does not show a significant spectro-temporal pattern ($SI_{DR,RN}=0.56$, $ASI=3 \times 10^5\%$). The neurons of E and G had significant responses to the dynamic ripple (0.45 spikes/sec and 0.11 spikes/sec respectively) but did not respond to the ripple noise (Rate=0 spikes/sec, $SI_{DR,RN}=0$, $ASI=+\infty\%$).

Since these response properties resemble those of simple cells in the visual cortex, in which neurons phase-lock to spatio-temporal gratings and respond to both structured gratings/bars (Victor and Purpura 1998; Girman, Sauve, and Lund 1999; DeAngelis, Ghose, and Ohzawa 1999) and to unstructured visual inputs, such as spatio-temporal m-sequences (Anzai *et al.* 1999; Reich *et al.* 2000), they are thus labeled as s-neurons.

In contrast to s-neurons, which responded strongly to both the dynamic ripple and ripple noise, a second class of neurons (18%) (Fig. 5) responded exclusively to the dynamic ripple stimulus. These cells generally had low firing rates to the dynamic ripple and little or no response to the ripple noise. The mean spike rate to the dynamic ripple was approximately 10% (1.4 spikes/sec) of that of the average s-neuron response. The mean response rate to the ripple noise, however, was approximately 2% (0.2 spikes/sec) of that of s-neurons. Despite their low spike rates *STRFs* to the dynamic ripple were highly significant ($p < 0.002$) despite the fact that only a few spikes were elicited over the experimental recording period. No significant *STRFs* were obtained for the ripple noise. Because of their apparent high feature sensitivity to structured sound inputs these neurons are designated as f-neurons.

F-neurons appear to be unresponsive when studied with unstructured noise sounds commonly used for reverse correlation (deCharms, Blake, and Merzenich 1998; Theunissen, Sen, and Doupe 2000). Accordingly the ability to characterize these neurons depends on the chosen stimulus and analysis method. The neuron of Fig. 5 A/B and C/D, responded to both the dynamic ripple (1.1 spikes/sec and 1.4 spikes/sec) and the ripple noise sound (0.4 spikes/sec and 0.2 spikes/sec) but their firing rate to the dynamic ripple

was significantly stronger. The neuron of Fig. 5 A/B had a significant *STRFs* ($p < 0.002$) to both sounds but the ripple noise *STRF* was significantly weaker. The neuron of C/D, however, had a significant *STRF* pattern only for the dynamic ripple sound indicating that it was not efficiently activated by the ripple noise. The neurons of Fig. 5 E–H had extremely low spike rates (0.45 spikes/sec and 0.11 spikes/sec respectively) to the dynamic ripple sound and no response to the dynamic ripple sound. The *STRFs* for these neurons were constructed using only 276 (Fig. 5 E) and 139 (Fig. 5 G) spikes for the dynamic ripple over a 10 and 20 minute recording period, respectively. Nevertheless, the *STRFs* obtained for these f–neurons are as noise free as those of s–neurons with much higher firing rates. This suggests that these neurons responded exclusively to a specific features of the dynamic ripple sound and consequently may require a high degree of local structure as found in FM sweeps and broadband clicks. Although such sound features are indeed present in the ripple noise stimulus (since this sound is constructed directly from the dynamic ripple) they generally occur at a low power levels and are not present in isolation. The low firing rate, high response specificity to the dynamic ripple, unresponsiveness to ripple noise demonstrate that these cells are extremely nonlinear and highly selective.

Such high specificity can, generally, not be detected with conventional methods unless there is a neuroethologic reason to use a specific stimulus for testing a neuron's response specificity. In such instances, sounds that are known to elicit a specific behavior, such as echolocation calls and species specific vocalizations, are used to quantify a neuron's response selectivity (Suga and Jen 1976; Suga, O'neil, Manabe 1978; Margoliash 1983; Margoliash and Fortune 1992; Olsen and Suga 1993a 1993b;

Casseday, Ehrlich, and Covey 1994; Ohlemiller, Kanwal, Suga 1996; Doupe 1997). Such procedures require spectral and/or temporal modification of the native sound: these include playing the normal sound followed by time-reversed, time-dilated (stretched), and/or frequency-shifted versions of the native sound. Selectivity is then determined by comparing the relative response strength for the normal and altered versions of the sound. Unlike such procedures, the described feature selectivity was determined with a general stimulus sequence with little a priori knowledge of the relevant stimulus features needed to efficiently activate the studied neurons. The apparent high selectivity of these neurons likely arise from intracellular thresholding, as has been demonstrated for somatosensory and visual cortex neurons (Moore and Nelson 1998; Binguier *et al.* 1999). Details for verifying and quantifying feature selectivity using the described sounds and reverse correlation procedures are outlined in detail sections 3.8–3.13.

3.6 STRF Similarity for the Moving Ripple and Ripple Noise Stimulus

To quantify the observed response differences between the ripple noise and dynamic ripple stimuli, we consider a procedure which quantifies the observed *STRF* differences. Given two experimental conditions (*A* and *B*) to be tested, we consider the vectorized *RFs* which consists of all sample values of $STRF_A$ and $STRF_B$ which exceed a significance test ($P < 0.002$) (see methods) for condition *A* or for condition *B*. The vectorized *RFs*, RF_A and RF_B , thus consists of the sample values of $STRF_A$ and $STRF_B$ for which either of the *STRFs* exceeded the significance test. To quantify the similarity of the *STRFs* we use the correlation coefficient or similarity index

(*SI*) (DeAngelis *et al.* 1999; Reich *et al.* 2000)

$$SI_{A,B} = \frac{\langle RF_A, RF_B \rangle}{\|RF_A\| \cdot \|RF_B\|} \quad (3.10)$$

where RF_A and RF_B are the significant *STRFs* for condition *A* and condition *B* respectively, $\langle \cdot, \cdot \rangle$ corresponds to the vector inner product, and $\|\cdot\|$ designates the vector norm operator. The similarity index quantifies the *STRF* shape differences or similarity independently of *STRF* amplitude. The *SI* assumes a numerical value normalized to the range -1 to 1 . Values near 1 indicate maximal shape similarity between $STRF_A$ and $STRF_B$, whereas values near 0 indicate that the *STRFs* have nothing in common and are thus orthogonal. *SI* values near -1 indicate that both *RFs* have similar spectro-temporal patterns but differ by a sign inversion.

For reference similarity index values are provided for the neurons depicted in Fig. 4 and 5. The *s*-neurons of Fig. 4 had relatively high *SI* values indicating that the shape of the *STRFs* for the ripple noise and dynamic ripple conditions were similar. The *SI* values for these responses pairs are: $A/B=0.77$, $C/D=0.78$, $E/F=0.94$, $G/H=0.7$. The high *SI* values indicates that these neurons responded to similar sound components for either condition. Two of the *f*-neuron response pairs of Fig. 5 likewise had high *SI* values ($A/B=0.78$, $C/D=0.56$) indicating that these neurons likewise responded to similar sound components for the ripple noise and dynamic ripple. Despite this *STRF* shape similarity, the response strength was significantly stronger for the dynamic ripple condition. Thus

these neurons show response sensitivity by virtue of their increased response strength for the dynamic ripple condition. The f–neuron response pairs of Fig. 5 E/F and G/H had *SI* values that were identically zero since no response was observed for the ripple noise stimulus.

Fig. 6 shows the similarity index distribution for all neurons which showed significant *STRFs* for the dynamic ripple and/or ripple noise stimulus conditions. The distribution of *SI* values is bimodally distributed. The vast majority of neurons ($n=49$) had large *SI* values, $SI > 0.5$. The mean *SI* value for this subset of neurons was relatively high, 0.75, indicating the shape of the obtained *STRFs* for the dynamic ripple and ripple noise sound bear a large degree of similarity. The remaining neurons ($n=13$) showed *SI* values less than 0.5. Of these, two neurons had values of *SI* that were nearly 0.5 (0.49 and 0.48), and six neurons had *SI* values that were identically zero. These neurons responded to the dynamic ripple sound and produced statistically significant *STRFs*, but did not respond to the ripple noise sound. Consequently these were classified as f–neurons. The *STRFs* for the subset of neurons with *SI* values less than 0.5 either showed no shape similarity ($SI=0$) or alternately showed a low degree of shape similarity.

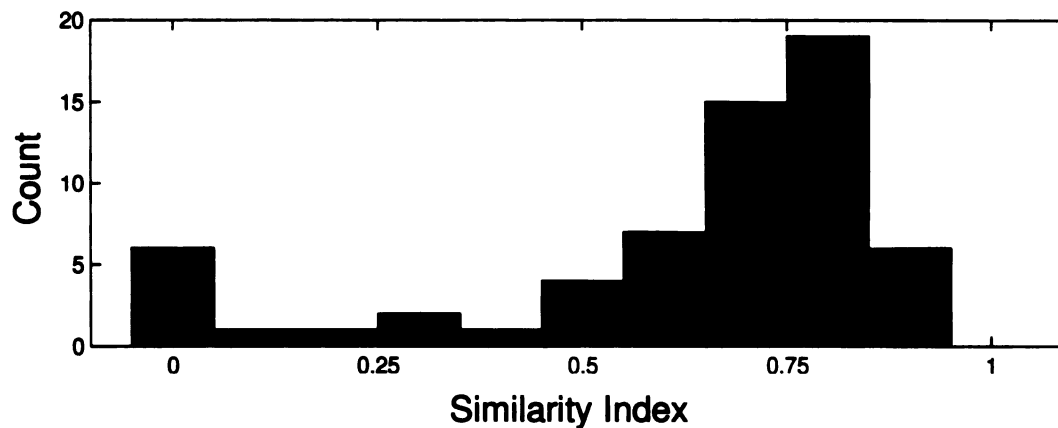


Figure 6: Similarity index distribution for all neurons that had spectro-temporal receptive fields for either the dynamic ripple *and/or* the ripple noise conditions. The distribution of similarity indices is bimodally distributed. Most neurons ($n=49$) had high *SI* values ($SI \geq 0.5$) indicating the ripple noise and dynamic ripple *STRFs* were similar in structure. Eleven neurons had low *SI* values of which 6 had values identically zero.

3.7 Response Strength – Moving Ripple versus Ripple Noise

In addition to considering the differences in *STRF* shape for the moving ripple and ripple noise stimulus conditions, it is likewise important to consider differences in the neuron's mean firing rate and the neuron's *STRF* amplitude for these two conditions. It is possible that a neuron produces *STRFs* with similar shape (high *SI* values) for the ripple noise and dynamic ripple sounds (Fig. 5 A/B, C/D) but the overall response strength and driving efficiency is significantly higher for one condition. Thus the similarity index is insufficient insofar that it can't tell us anything about the overall response strength for a given condition.

From our null hypothesis, recall that for an ideal linear neuron the shape of the

STRFs should be identical for the dynamic ripple and ripple noise since the shape of their autocorrelation functions are identical (see chapter 2; section 2.4; Eq. (2.)5). For a linear model neuron, the expected *STRF* amplitude should likewise be the same for the dynamic ripple and ripple noise conditions assuming that the *STRF* is computed and normalized according to Eq. (3.1). The *STRF* computed directly from Eq. (3.1) is an impulse response descriptor and thus it has units of output/input (spikes/sec/dB), where the neuron's output is measured as a rate (spikes/sec) and the input driving stimulus amplitude is measured in units of dB. For a linear neuron, the transfer function characteristics are independent of the input stimulus, and, thus, it is expected that the measured transfer characteristics of the neuron are identical for any given input. Although this *STRF* normalization captures the effective gain of the stimulus–response relationship, it is nonetheless instructive to consider alternate normalizations for the *STRF* magnitude.

Given that the output of sensory neurons is generally measured as a spike rate, it is desirable to consider a normalization for the *STRF* which more closely describes a sound's driving efficiency in units that are intuitive from a neurophysiological perspective. An alternate normalization for the *STRF* is obtained by considering the average response output produced for the average stimulus presented. The normalized rate *STRF*, $STRF_r = \sigma_s \cdot STRF(\tau, X_k)$, corresponds to the neuron's *STRF* (units of spikes/sec/dB) scaled by the input stimulus standard deviation, σ_s (units of dB). The normalized rate *STRF* thus describes the average firing rate change (units of spikes/sec) about the neuron's mean firing rate that is produced by presenting a sound of average

differential intensity (i.e. σ_x) within the neuron's receptive field. Unless otherwise stated we consider $STRF_r$ throughout.

As a means to relate the mean firing rate of a neuron to the average difference output as predicted from the neuron's $STRF_r$, we again consider the linear model neuron of Eqs. (2.3) and (2.4) where the filter for the k^{th} input channel is related to the neuron's $STRF_r$ by $h_k(\tau) = STRF_r(\tau, X_k)$. We would like to derive an equation for the expected output standard deviation or, equivalently, the neuron's firing rate variance that is predicted by its $STRF_r$. Thus the desired metric should provide a measure of the energy in the response that is captured by the neuron's $STRF_r$. For the linear model neuron the predicted firing rate variance is expressed as

$$\sigma_r^2 = E[(r(t) - r_0)^2] = \sum_{j=1}^N \sum_{k=1}^N E[(r_j(t) - r_0) \cdot (r_k(t) - r_0)] \quad (3.11)$$

where $r(t)$ is the predicted firing rate of the neuron (chapter 2; Eq. (2.3)), r_0 is the neuron's mean firing rate, and $r_k(t)$ is the predicted output for the k^{th} filter channel,

$h_k(t)$. Substituting Eq. (2.4) into Eq. (3.11) we get

$$\begin{aligned} E[(r_j(t) - r_0) \cdot (r_k(t) - r_0)] &= E\left[\int s_j(t - \tau_1) h_j(\tau_1) d\tau_1 \cdot \int s_k(t - \tau_2) h_k(\tau_2) d\tau_2\right] \quad (3.12) \\ &= \int \int E[s_j(t - \tau_1) s_k(t - \tau_2)] \cdot h_j(\tau_1) h_k(\tau_2) d\tau_1 d\tau_2 \\ &= \int \int R(\tau_1 - \tau_2, X_j - X_k) \cdot h_j(\tau_1) h_k(\tau_2) d\tau_1 d\tau_2 \end{aligned}$$

where $R(\tau_1 - \tau_2, X_j - X_k) = \sigma_s^2 \cdot \text{sinc}(2F_{Max}(\tau_1 - \tau_2)) \cdot \text{sinc}(2\Omega_{Max}(X_j - X_k))$ for the ripple noise and dynamic ripple stimuli (chapter 2, Eq. (2.56) and Eq. (2.75)). If the neurons spectro-temporal integration scale is slower than the fastest spectro-temporal components present in the ripple noise and dynamic ripple stimuli, the stimulus's autocorrelation function can be approximated by a spectro-temporal impulse:

$R(\tau_1 - \tau_2, X_j - X_k) \approx \sigma_s^2 \cdot \delta(\tau_1 - \tau_2) \cdot \delta(X_j - X_k)$. Substituting into Eq. (3.12) yields

$$E[(r_j(t) - r_0)^2] \approx \int \int \sigma_s^2 \delta(\tau_1 - \tau_2) \cdot h_j(\tau_1) h_j(\tau_2) d\tau_1 d\tau_2 = \sigma_s^2 \cdot \int h_j(\tau)^2 d\tau \quad (3.13)$$

for $k=j$ and

$$E[(r_j(t) - r_0) \cdot (r_k(t) - r_0)] = 0 \quad (3.14)$$

for $k \neq j$.

Combining with Eq. (3.11) the firing rate variance which is captured by the neuron's *STRF* is expressed as

$$\sigma_r^2 = \sigma_s^2 \sum_{k=1}^N \int h_k(\tau)^2 d\tau = \sum_{k=1}^N \int \text{STRF}_r(\tau, X_k)^2 d\tau \quad , \quad (3.15)$$

where σ_r can now be computed directly from the $STRF_r$, by computing its RMS value.

Although the expected $STRFs$ for the ripple noise and the dynamic ripple conditions have identical shape and amplitude for a linear model neuron, the expected normalized rate $STRF$, $STRF_r$, and the expected driving efficiency differ slightly in magnitude for these conditions. Consider the $STRF_r$ and the response variance that is captured by the $STRF$, σ_r (Eq. (3.15)). The expected ratio of these metrics for the moving ripple and ripple noise conditions

$$\sigma_{r_{RN}} / \sigma_{r_{DR}} = STRF_{r_{RN}} / STRF_{r_{MR}} = \sigma_{s_{RN}} / \sigma_{s_{DR}} \quad (3.16)$$

is strictly determined by the ratio of standard deviations for both stimulus conditions for an ideal model neuron. Note that these ratios assume identical values.

Despite the fact the dynamic ripple and ripple noise signals have spectro-temporal autocorrelations with identical shape and identical maximum and minimum amplitude values, their spectro-temporal standard deviation are slightly different. The moving ripple has a standard deviation of $\sigma_{s_{DR}} = M / \sqrt{8}$, where M is the maximum modulation depth of the signal (units of decibels), and the ripple noise has a standard deviation of $\sigma_{s_{RN}} = M / \sqrt{12}$. Hence the expected differential response for a hypothetical linear neuron that is captured by the neuron's rate $STRF_r$, should in theory differ by a multiplicative factor of $\sigma_{s_{RN}} / \sigma_{s_{MR}} = \sqrt{2/3} = 0.81$. Thus, the measured $STRF_r$ for the dynamic ripple condition should be roughly 20 % stronger in magnitude than for the

ripple noise condition.

Given the expected response properties for a linear model neuron, we can measure how the response statistics of inferior colliculus neurons differ from those expected for the null hypothesis. To quantify the observed differences between the dynamic ripple and ripple noise conditions, we consider three metrics which describe the mean and differential response activity of a neuron. Given two conditions to be tested, A and B , we measured the mean spikes rates, r_A and r_B , the measured $STRF_r$ RMS values, σ_{r_A} and σ_{r_B} , and the peak to peak $STRF_r$ magnitudes, ΔRF_A and ΔRF_B , where $\Delta RF = \max(STRF_r) - \min(STRF_r)$. The peak to peak $STRF_r$ magnitude, ΔRF , and the $STRF_r$ RMS values, σ_r , provide a direct measure of the neuron's driven phase-locked activity (in units of spikes/sec) that is captured by the neuron's $STRF_r$, whereas the mean spike rate provide a measure of the average activity, regardless of whether the activity is specific or nonspecific.

Statistics for these three response parameters are shown for the dynamic ripple and the ripple noise stimulus in Fig. 7. The measured mean firing rates, $STRF_r$ peak to peak magnitudes, and RMS values were significantly correlated ($r = 0.86 \pm 0.07$, $r = 0.81 \pm 0.08$, $r = 0.6 \pm 0.1$ respectively) for the dynamic ripple and ripple noise conditions. Thus, on the average, neurons that responded strongly (weakly) to the ripple noise likewise responded strongly (weakly) to the dynamic ripple. The neurons' phase-locked activity as measured by σ_r and ΔRF also showed similar trends although these trends seemed to be less apparent for neurons with very low spike rates (< 2.0 spikes/sec).

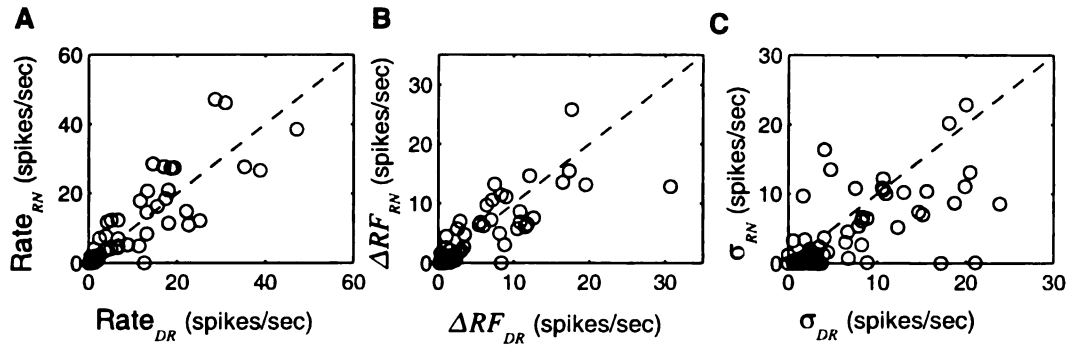


Figure 7: Scatter plot of the measured response parameters for single neurons in the ICC shown for the ripple noise and dynamic ripple conditions. Ripple noise versus the dynamic ripple mean firing rate shown for single neurons (A). *STRF* peak to peak magnitudes (units of spikes/sec) shown for the dynamic ripple versus ripple noise (B). Scatter of the *STRF*, RMS value shown for both stimulus conditions (C).

The interrelations among the three response parameters is shown in Fig. 8 for both the dynamic ripple (blue) and ripple noise (red) conditions. All parameters show a significant correlation (Fig. 8 A–C; dynamic ripple: $r=0.82\pm0.07$, $r=0.88\pm0.06$, $r=0.95\pm0.04$; ripple noise: $r=0.91\pm0.05$, $r=0.74\pm0.09$, $r=0.86\pm0.07$). On the average, neurons that had high (low) firing rates also had high (low) values of σ_r and ΔRF_r . The overall range of values of r_0 , σ_r , and ΔRF_r were highly overlapped for the ripple noise and dynamic ripple conditions. The largest deviates from this trend were observed for neurons with low firing rates (<2 spikes/sec) where the

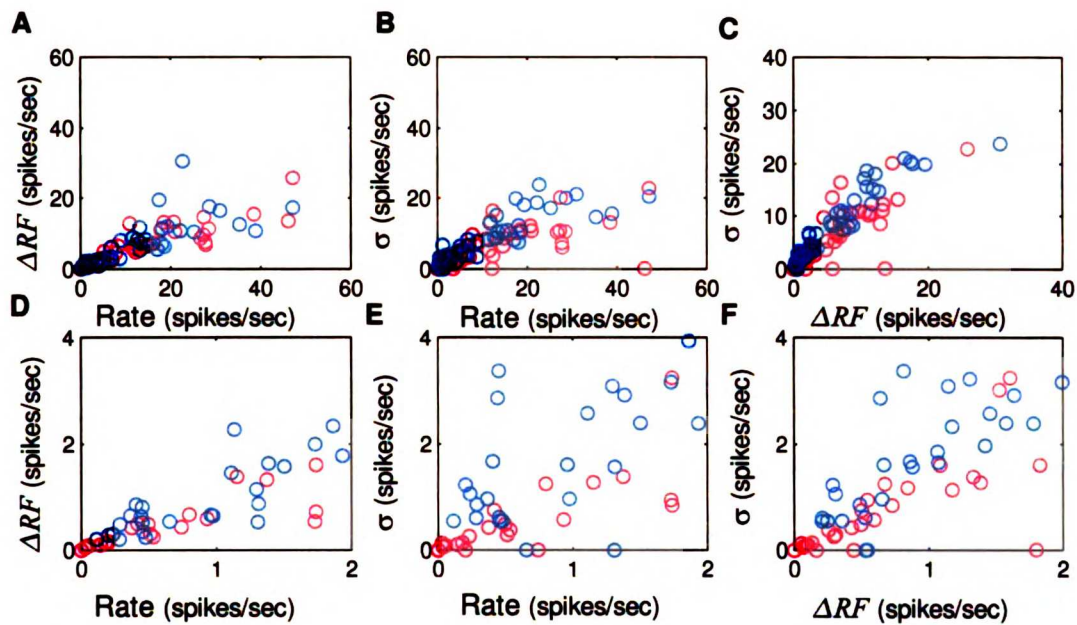


Figure 8: Response statistics for the dynamic ripple (red) and ripple noise (blue) conditions showing the interrelation among the mean firing rate, $STRF$, RMS value, σ , and the peak to peak $STRF$, amplitude, ΔRF . Scatters plots for the three response parameters shown for all neurons (A–C). All parameters are significantly correlated and cover similar ranges for the ripple noise and dynamic ripple conditions. Thus the overall population response strength was similar for both conditions. For neurons with mean spikes rates of less than 2 spikes/sec (D–F; same as A–C but zoomed in), however, the ripple noise mean and difference response rates (σ , and ΔRF) are considerably weaker.

ripple noise condition seemed to elicit weaker responses (smaller values for the three parameters; see Fig. 8 D–F). The large amount of overlap for the ripple noise and dynamic ripple conditions indicate that the overall activity for these two stimulus conditions is comparable for the neuronal population. The data as presented in Fig. 7 and

Fig. 8, however, does not indicate that the dynamic ripple and ripple noise provide equal functional driving force for any given single neurons. Note that although the range of values for these three parameters are comparable, individual neurons could show large changes in their response parameters between the dynamic ripple and ripple noise conditions. Consequently the data of Fig. 8 does not convey the extent of dissimilarity in these parameters for single neurons when tested for both stimuli.

To quantify differences in the mean firing rate and σ_r for individual neurons, we constructed two metrics that quantify the percent difference in response strength between the considered conditions (condition *A*=moving ripple and condition *B*=ripple noise). *STRF* amplitude differences are characterized by the amplitude similarity index (*ASI*)

$$ASI_{A,B} = s \cdot \left[\left(\frac{\sigma_{r_A}}{\sigma_{r_B}} \right)^s - 1 \right] \times 100\% \quad (3.17)$$

where $s = \text{sign}(\sigma_{r_A} - \sigma_{r_B})$. This metric measures the percent change in *STRF* energy for condition *A* and *B*. The *ASI* metric assumes values between negative and positive infinity. A value of zero indicates that $\sigma_{r_A} = \sigma_{r_B}$ whereas values greater than zero indicate that $\sigma_{r_A} > \sigma_{r_B}$. Values less than zero alternately indicate that $\sigma_{r_A} < \sigma_{r_B}$. The magnitude of $ASI_{A,B}$ is numerically equivalent to the percent difference between σ_{r_A} and σ_{r_B} where the sign of $ASI_{A,B}$ indicates an increase in the *STRF* energy referenced

on condition *A* (for negative values) or *B* (for positive values).

A similar metric was also used to characterize the mean response rate differences for the two experimental conditions. We consider the rate similarity index (*RSI*)

$$RSI_{A,B} = s \cdot \left[\left(\frac{r_A}{r_B} \right)^s - 1 \right] \times 100\% \quad (3.18)$$

where $s = \text{sign}(r_A - r_B)$. This metric is numerically identical to the *ASI* where the mean rates for conditions *A* and *B* are substituted for the *STRF* norms for those conditions. The *RSI* and *ASI* differ since the *RSI* measures mean rate changes between stimulus conditions whereas the *ASI* measures differences in stimulus driven activity (note that the *STRF* is a direct measure of the stimulus phase-locked differential spike rate produced by a given stimulus pattern).

Statistics for the mean and difference spike rates are shown in Fig. 9. For display purposes, values of *ASI* or *RSI* with magnitudes greater than 1000% are set to 1000%. Response rates and *STRF* energies were substituted for the dynamic ripple into r_A and σ_{r_A} and for the ripple noise into r_B and σ_{r_B} . Values of *ASI* and/or *RSI* greater than zero indicate that the dynamic ripple sound produced higher firing rates and/or differential response rates. The vast majority of neurons ($n=45$) had *ASI* and *RSI* values centered about zero ($|ASI| < 500\%$ and $|RSI| < 500\%$) indicating that the firing rates and *STRFs* were comparable in strength for both conditions. A large number of neurons ($n=16$) either showed significantly higher firing rates ($RSI > +500\%$) and/or significantly

higher *STRF* energies ($ASI > +500\%$) for the dynamic ripple sound. Three of these

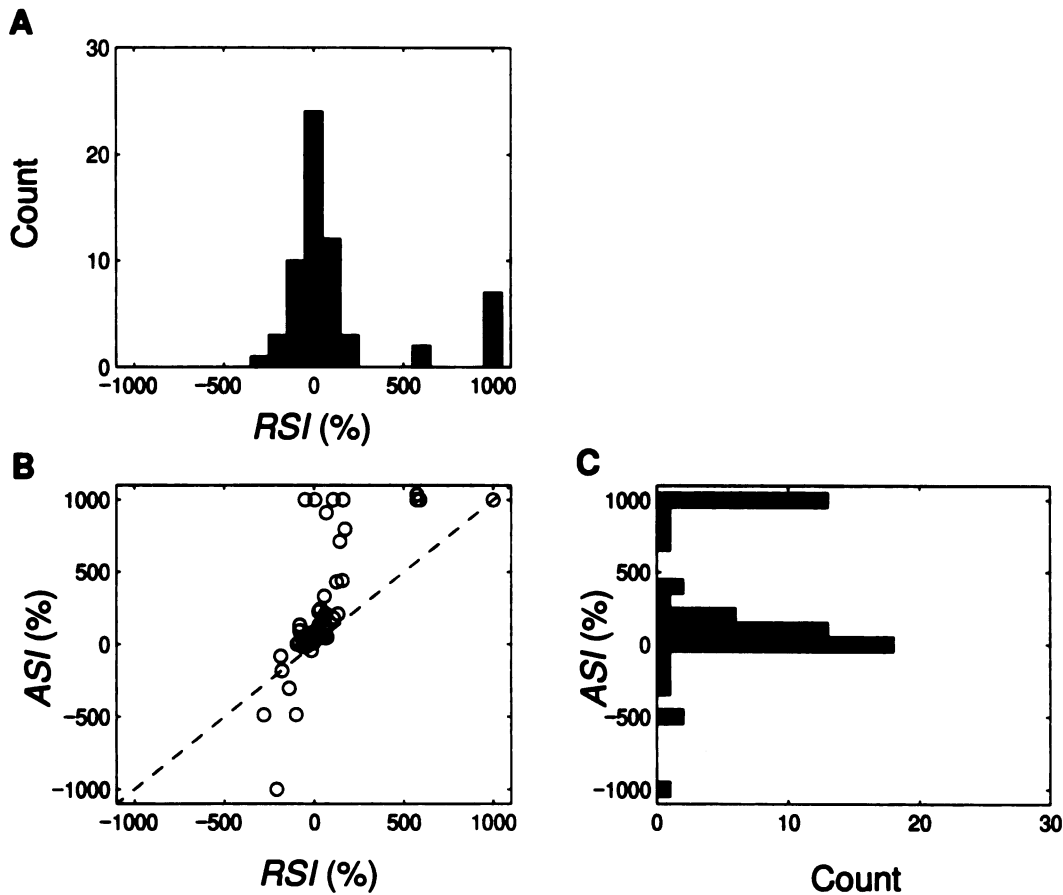


Figure 9: Rate and amplitude sensitivity index statistics for phase-locked neurons. *ASI* (A) and *RSI* (C) histograms show a large central peak centered about zero. These neurons consisted of s-neurons which responded with similar firing rates and produced similar *STRF* strengths for the dynamic ripple and ripple noise conditions. A large number of neurons additionally showed large *ASI* and/or *RSI* values. Neurons which had values for either of these metrics that were numerically greater than 1000% are shown with the *ASI* or *RSI* values set to 1000%. These largely consisted of f-neurons which responded most specifically to the dynamic ripple stimulus. Scatter plot for *ASI* and *RSI* (B) shows the interrelation among these metrics. For neurons that had small *ASI* and

RSI magnitudes ($< 500\%$), the two metrics were highly correlated ($r = 0.79 \pm 0.056$; bootstrap). The correlation coefficient for the two metrics was significantly lower for the remaining neurons ($r = 0.4 \pm 0.12$, bootstrap: $p < 0.05$).

neurons had *ASI* values in between 500% and 1000% while the remaining neurons had *ASI* values greater than 1000%. Of these, six neurons had values of *RSI* and *ASI* identically $+\infty\%$ since they responded to the dynamic ripple stimulus but produced zero spikes for the ripple noise sound. These are shown collectively as a single point centered about $ASI = +1000\%$ and $RSI = +1000\%$. Two additional neurons had values of *ASI* and *RSI* that were greater than +500%. The remaining neurons in this category had comparable firing rates for the dynamic ripple and ripple noise whereas the phase-locked component of their response was significantly stronger for the dynamic ripple sound ($ASI > +500\%$). One neuron alternately had a large negative *ASI* value ($ASI = -7570\%$) despite the fact that its overall firing rate was similar for the two stimulus conditions ($RSI = -208\%$), indicating that its envelope phase-locked response was stronger for the ripple noise.

3.8 Quantifying Response Specificity to Structured Sound Patterns

The fact that some neurons respond preferentially to the dynamic ripple sound and are unresponsive to the ripple noise sound suggest that these neurons may require a high degree of local sound structure, as found in FM sweeps and spectral resonances, to efficiently activate them. Given that precise *STRFs* are generated for these neurons when

using the dynamic ripple and no *STRFs* are produced for the ripple noise stimulus (large positive *ASI* values) for these neurons is consistent with this observation. Here we further address issues related to neuronal response selectivity by quantifying the specificity of a neuron's response to particular sounds pattern. This analysis serves as a consistency check, allowing us to relate observed mean- and difference-spike rate properties of the response (between the dynamic ripple and ripple noise), to the neuron's overall response specificity.

The feature detector hypothesis was first proposed by Barlow (1953) during his early works of the amphibian retina. His ideas eventually lead to the doctrine that sensory neurons encode behaviorally relevant stimuli. In his early works, Barlow noted that the responses of ganglion cells in the amphibian retina responded specifically to small flashes of light. This elementary response pattern was sufficient and led Barlow to the proposal that neurons in the amphibian retina function as "fly detectors". The stereotyped "on-off" response patterns of single neurons at this early level of processing could in theory serve as a stimulus trigger for invoking the frog's "strike and swallow" feeding behavior. Although this simple hypothesis is attractive, due to its simplicity and its direct link between sensory coding and behavior, numerous competing theories of sensory coding, include network theories and the idea that single neurons encode stimulus information as linear filters (Jones and Palmer 1987; Deangelis *et al.* 1993; Versnel and Shamma 1998), predominate. With possible exceptions of the echolocating bat's and birdsong's auditory systems, few convincing examples of true feature selectivity have been described. Currently, it is not clear to what degree the feature selectivity plays a role in sensory coding in part because a general definition and

quantitative methodologies for quantifying feature selectivity are lacking. Here we specifically consider: To what degree do single neurons behave like linear filters? Or do single neurons more closely resemble ideal feature detectors? Are neurons that function like feature detectors common and, if so, what functional purpose do these serve?

Response selectivity is generally quantified by measuring a neuron's firing rate for a given sound. Selectivity to a native sound is determined by performing a spectral and/or temporal modifications in which a behaviorally relevant sound is reversed (Glass and Wollberg 1983; Doupe 1997), stretched (Wang *et al.* 1995), and/or filtered (Theunissen and Doupe 1998). Other paradigms and/or stimulus modifications are also possible: these may include testing for response specificity to combinations of sound components (Suga, O'neil, and Manabe 1978; Olsen and Suga 1993a 1993b; Doupe 1997), signals in noise, and testing for specificity to conspecific calls (Ohlemiller, Kanwal, Suga 1996; Doupe 1997). The disparity of the response between the native and modified stimulus is then used to quantify the overall degree of response selectivity. In general such procedures generally require a priori knowledge of the relevant sound component to be tested. Although these methodologies have proven useful for studying echolocating bats and songbirds, similar attempts have failed to reveal highly specific neurons in primate species (Winter and Funkenstein 1973; Glass and Wollberg 1983) despite the fact that these animals exhibit a diverse audio-vocal behavior (Ploog 1981). For most animals, including cats and possibly primates, there is no a priori basis for choosing a particular sound or stimulus component. Thus the search for feature selective neurons is limited by the fact that the probe stimulus is not well defined for most animals.

The presented procedure for quantifying feature selectivity differs in that no a priori knowledge is necessary for defining the degree of feature selectivity of a neuron. Secondly, the presented procedure assigns a numerical value to the observed response pattern where the devised metric is defined in reference to the ideal feature detector and that of a randomly firing neuron.

Given a large subset of possible stimuli an ideal feature detector neuron would respond to one and only one sound component. Thus if one were to average over all the stimuli that evoked a neuronal event (in order to estimate the neuron's *STRF* or its impulse response) one inevitably obtains an *STRF* that identically matches the sound pattern of interest. For such an idealized scenario the response covariance (Rieke *et al.* 1997), a measure of the overall variance of the sound patterns that produce a neuronal response, is zero since the sound patterns are identical for each neuronal event. Although this ideal scenario is unlikely to occur in a physiologic system, it nonetheless serves as a useful reference for defining response specificity.

It is required that the probing stimulus be persistently exciting and spectro-temporally rich so that it probes numerous stimulus possibilities over sufficiently long test period. The motivation for this requirement is twofold: first a sufficiently rich stimulus is required so that estimation of the neuron's *STRFs* is performed in a statistically sound and unbiased manner. Furthermore, this requirement is necessary so that the response specificity is tested with respect to a large subset of possible sound patterns. The dynamic ripple and ripple noise stimuli satisfy this basic requirements. We determine the degree of feature sensitivity relative to the subset of all possible sound patterns that are present in the dynamic ripple and/or ripple noise stimulus space.

Recall that the *STRF* corresponds to the average of all the sound patterns preceding every spike. In principle this corresponds to a first-order estimate of the "optimal" first-order sound pattern that evoked neuronal responses for the given neuron and does not tell us anything about the variability of the constituent sound patterns that compose the *STRF*. Here we ask how variable are these sound patterns on a trial to trial basis: are these neurons feature selective, responding exclusively to sound patterns that identically match their *STRFs* or are they non-specific? What are the basic differences in response specificity between the described f- and s-neurons? The presented procedures, seeks to derive the variance in the constituent sound patterns that is attributed to the spectro-temporal content or shape, and is independent of amplitude and contrast.

The computation of the feature selectivity index (*FSI*) metrics is described in Figs. 10 and 11. Consider the statistically significant vectorized receptive field, RF_p , where $p < 0.002$ is the level of confidence in determining the significant *STRF*. Given this vectorized *STRF* the spectro-temporal sound pattern which preceded the k^{th} spike and is spatially overlapped with the vectorized receptive field, RF_p , is designated as S_k . We ask how similar is this sound pattern to the neuron's *STRF*?

The degree of similarity between a sound pattern and the neuron's *STRF* is derived by computing the correlation coefficient or similarity index (*SI*) between the k^{th} sound pattern, S_k , and the neuron's vectorized receptive field, RF_p . This operation is expressed as

$$SI_k = \frac{\langle RF_p, S_k \rangle}{\|RF_p\| \|S_k\|} \quad (3.19)$$

Presumably if there is a high degree of similarity between the neuron's receptive field and the constituent sound pattern, the similarity index will be close to unity. Alternately, if the similarity index is near zero, the constituent sound pattern bear no resemblance to the neuron's *STRF*. This operation is repeated for every neuronal event, resulting a similarity index for each spike. The similarity index measure differs from the variance calculation used to determine the response covariance since it only accounts for variability that is inherent in the spectro-temporal shape independent of amplitude differences (Fig. 10).

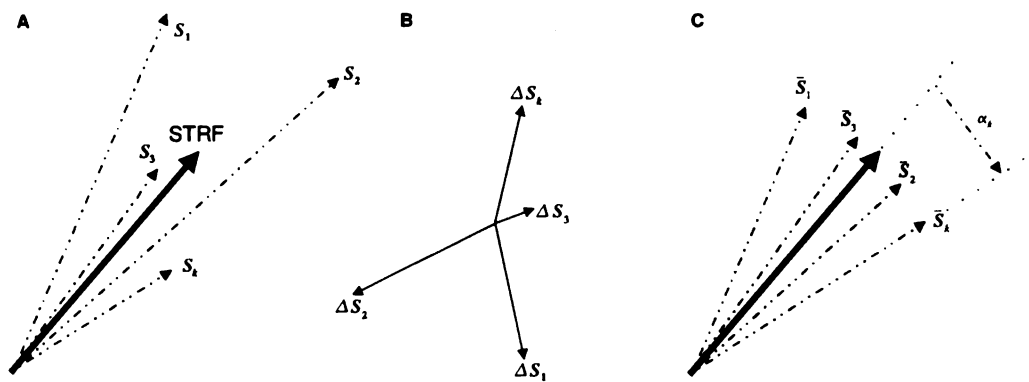


Figure 10: Abstract representation of the *STRF* and the relationship between the covariance and the similarity index. The *STRF* is expressed as the average vector defined in an $N_t \times N_x$ dimensional space (A): where $N_t=400$ and $N_x=230$ correspond to the number of temporal and spectral samples used to compute the *STRF* respectively. Here the vectors S_k designate the sound patterns used to construct the

STRF. The covariance is computed by computing the difference vector (B),

$\Delta S_k = STRF - S_k$, and averaging its squared values, ΔS_k^2 . This measure therefore accumulates differences between the *STRF* and the constituent sound patterns, S_k , that arise from *magnitude* and *shape* differences (associated with the direction of the *STRF* and S_k). The similarity index is computed by normalizing all sound vectors and the *STRF* so that they have unity variance (C). This measure therefore only takes into account shape differences between the *STRF* and the individual sound patterns (irrespective of magnitude differences). The similarity index is defined as $SI_k = \cos(\alpha_k)$, where α_k is the angle between the *STRF* and the k^{th} sound pattern.

Given the array of similarity indices, $SI = [SI_1, SI_2, \dots, SI_N]$, for a total of N neuronal responses, the overall specificity of a neuron is assessed by considering the overall response statistics for the dynamic ripple and/or the ripple noise sound ensemble. Given this array, one can construct a similarity index histogram (*SIH*) by collapsing the array of similarity indices, SI , into a probability histogram (Fig. 11). This measure provides an estimate for the similarity index distribution, $p(SI)$, for a given sound and experimental condition. The relative degree of feature selectivity is determined by considering the degree of skewness of the similarity index distribution relative to the expected distribution for an ideal feature detector neuron and to a null response condition in which a neuron fires randomly.

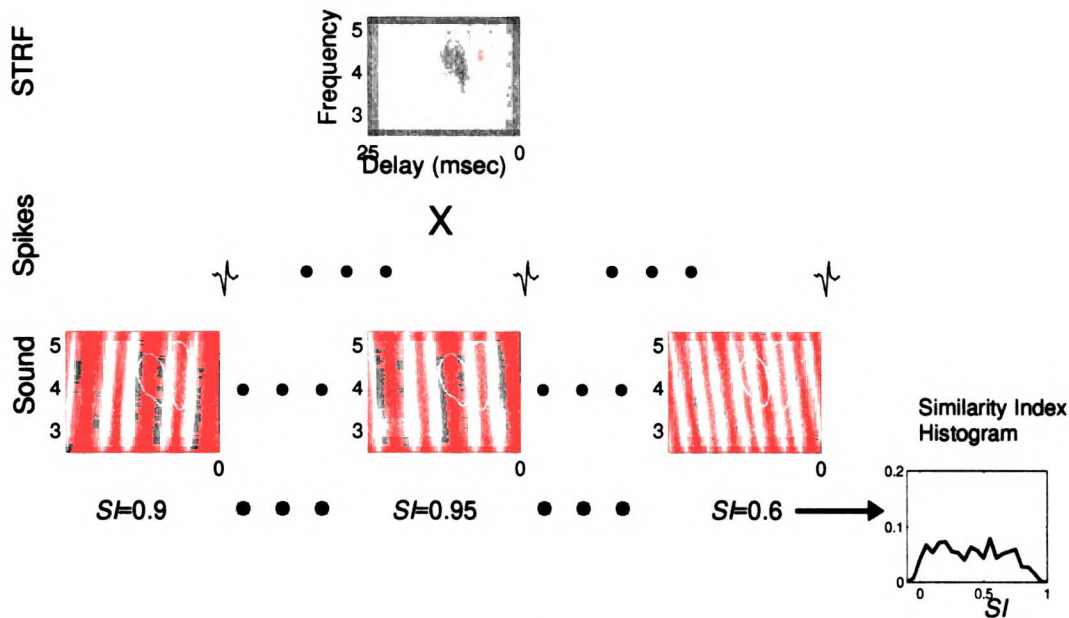


Figure 11: Computing the similarity index histogram. The neuron's significant *STRF* is used to define the neuron's "optimal" sound pattern. Selectivity is tested against this reference sound pattern. For each occurrence of a spike, the pre-event spectro-temporal envelope is extracted (third row). These sound segments are individually compared with the neuron's significant *STRF* ($p < 0.002$). Using the sound segments that are spatially overlapped with the neuron's significant *STRF* (top), designated by white contour, the correlation coefficient or similarity index is computed between the *STRF* and each sound segment. For each spike, the similarity index provides a metric with numerical values between -1 and 1 . Zero indicates that the *STRF* and sound pattern bear no resemblance whereas unity indicates maximal similarity. Values near -1 alternately indicate that the sound pattern resembles the *STRF* but has a sign inversion. The similarity index histogram is obtained by collapsing the array of similarity indices obtained for each spike into a histogram (bottom right). This descriptor provides an accurate picture of the neuron's response specificity for the stimulus ensemble under consideration (shown for the dynamic ripple sound).

For an ideal feature detector, the trial to trial variability between the neuron's *STRF* and the sound patterns that evoked neuronal responses is zero. Such a neuron responds *if and only if* the presented sound pattern identically matches the neuron's *STRF*. For such an idealized neuron, the trial to trial similarity index is unity. The corresponding *SIH* will thus consist of a single impulse centered about unity:

$p_{fd}(SI) = \delta(SI - 1)$, where $p_{fd}(SI)$ is the expected distribution function for the ideal feature detector.

In contrast, we also consider a null condition in which a neuron fires randomly and the spiking output bears no direct relationship to the input sounds. A purely random neuron produces a zero valued *STRF* since the neuronal response is not functionally related to the driving stimulus. Consequently, the similarity index distribution consists of a single impulse centered about zero: $p_r(SI) = \delta(SI)$, where $p_r(SI)$ designates the similarity index distribution for the random neuron.

Although this ideal scenario is appealing as a reference for defining selectivity, we will instead consider the *SI* distribution obtained for a random set of spikes and the non zero-valued *STRF* for the neuron under investigation. In general, this distribution is centered about zero but it is slightly broader than the obtained distribution for the described scenario (Fig. 14). We consider this distribution, as opposed to the ideal distribution for the random neuron, because the *SI* distribution for the neuron under investigation, $p(SI)$, is obtained directly from the significant *STRF* which is in all cases non zero-valued. If, for example, a subset of the neuron's spikes are non-specific, the *SI* value for these spikes will be derived using an *STRF* which is not zero. Although this subset of spikes is not causally related to the input sound patterns, the resulting *SI*

values can differ from zero because the *STRF* is real valued. Hence, the contribution from random spikes in the neuron's response have *SI* values which differ from zero. For the null condition, we therefore consider the distribution of *SI* values obtained using the neuron's significant *STRF*, RF_p , and a random set of spikes since this distribution closely resemble the properties of non-specific spikes in the neuronal data.

The *SI* distribution for a randomly firing neuron, $p_r(SI)$, is thus experimentally derived for each individual *STRF* by randomly choosing a set of 12,000 spikes and constructing the *SI* distribution. Because of the differences in *STRF* shape and the relative energy of its excitatory and inhibitory domains for individual neurons, the *SI* distribution differs for each *STRF* and thus it must be reestimated for each neuron. An example of the experimentally derived distribution is shown in Fig. 14 (additional examples are shown in Figs. 12 and 13). Although the ideal and experimentally derived distributions are both centered about zero, the experimental distribution is slightly broader and has values other than zero.

3.9 Similarity Index Histogram of s- and f-Neurons

Similarity index histograms were computed for all neurons that showed a significant *STRF* in order to determine the relative degree of specificity for the dynamic ripple and ripple noise conditions. As described, the similarity index histogram quantifies the specificity of the response directly by comparing the neuron's response similarity index statistics to those expected for an ideal feature detector and a random firing neuron. In section 3.7, s- and f-neurons were categorized according to the measured response strength differences between the dynamic ripple and the ripple noise. The similarity

index histogram method serves as a consistency check, allowing us to compare the results of section 3.7 directly with a theoretically sound measure of selectivity.

Fig. 12 and 13 shows the similarity index histograms for the s- and f-neurons of Figs. 4 and 5 respectively. *SI* histograms are shown for both the dynamic ripple (left column) and the ripple noise (right column) as continuous lines. Control *SI* histograms are also shown for the random firing assumption (red-dotted lines).

S-neurons have *SI* histograms that are largely overlapped with the random control condition. With the exception of two of the forty-five s-neurons (one of these neurons is shown in Fig. 12 C (dynamic ripple) and D (ripple noise)), this trend was evident for both the dynamic ripple and the ripple noise signals. These neurons have *SI* distributions that resembles those observed for f-neurons (compare to Fig. 13). For all neurons *SI* distributions for the dynamic ripple were, in general, narrower than for the ripple noise condition. This is also evident in the random control distribution, $p_r(SI)$ for the examples of Fig. 12, indicating that this is a stimulus dependent property of the similarity index histogram.

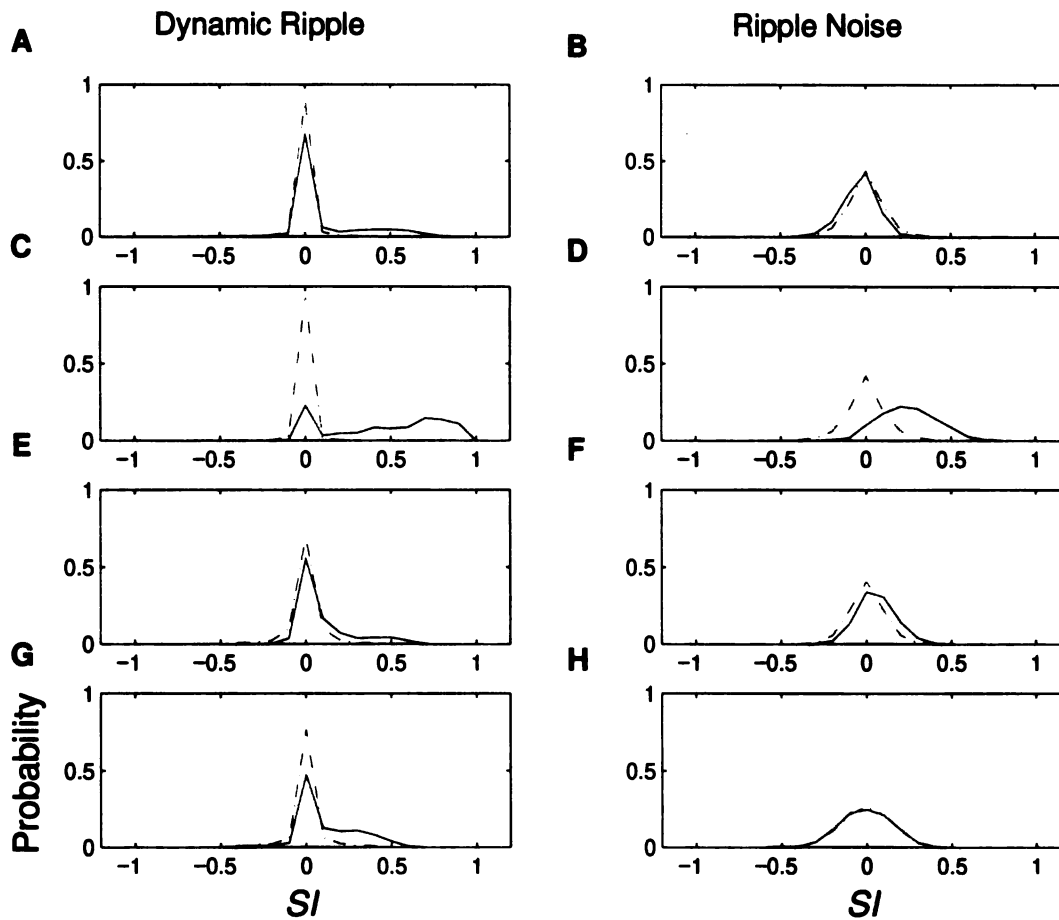


Figure 12: Similarity index distributions for the s -neurons of Fig. 4 – estimated for the dynamic ripple (A, C, E, H) and for the ripple noise (B, D, F, H). For most s -neurons (43/45) the derived similarity index distributions (continuous line) are highly overlapped with those of the random control simulation (red, dashed-dotted line). This is true for both the dynamic ripple (left column) and the ripple noise (right column) stimuli. Only two of the classified (using the procedure of section 3.7) s -neurons showed similarity index distributions that were shifted towards +1 relative to the random control condition (one of these shown: C/D).

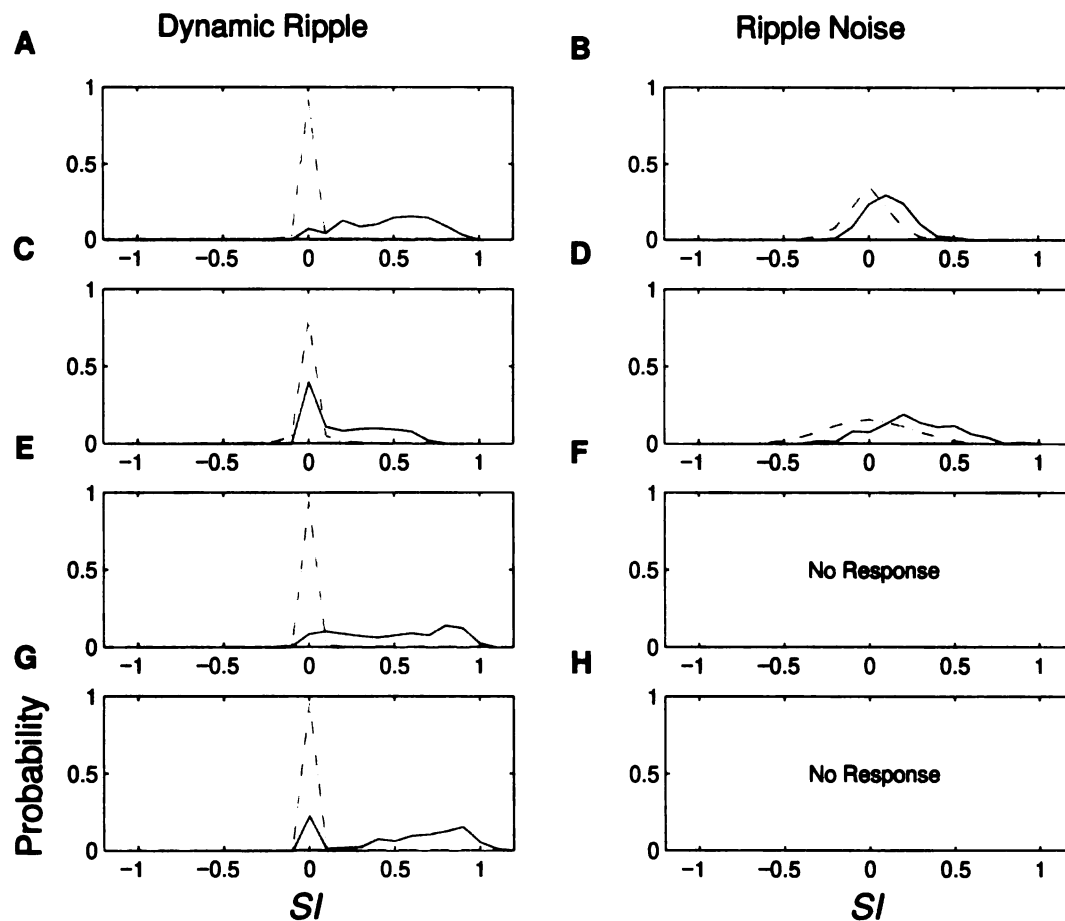


Figure 13: Similarity index distributions for the f -neurons of Fig. 5 – estimated for the dynamic ripple (A, C, E, H) and for the ripple noise (B, D, F, H). Similarity index histograms for the dynamic ripple are positively skewed towards +1 and have little overlap with the random control SI distribution. Ripple noise SI histograms (B and D) are similar to the random control SI distribution, suggesting that these neurons respond with higher selectivity to the features present in the dynamic ripple. The neurons of E and G did not respond to the ripple noise (F and H).

By comparison, a large number of f -neurons had SI histograms for the dynamic ripple that were positively skewed (9/17) towards +1 and were not overlapped with the random firing condition. Such neurons are depicted in Fig. 13 (A, C, E, G). Comparing

with the ripple noise condition, these neurons either did not respond to the ripple noise (F and H) or produced *SI* histograms that resemble the random control distribution, but were only slightly positively skewed towards +1, thus suggesting that they responded more precisely to the dynamic ripple sound.

3.10 Quantifying Feature Selectivity

The similarity index histogram devised in section 3.8 allows us to measure the variability of the input sound patterns that evoke neuronal responses. These can be compared directly to the results expected for an ideal "feature detector" neuron and for a random neuron. Comparisons among the different cell classes in the ICC shows that the similarity index histogram of some cells are skewed relative to the random condition – indicative of higher selectivity. Here we devise a numerical metric that distinguishes and quantifies the relative selectivity of a neuron directly from the similarity index histogram and the reference similarity index histograms.

Using the ideal feature detector and the randomly firing neuron's *SI* distributions as a reference point, we devise three feature selectivity index (*FSI*)

$$FSI_1 = \frac{\int_{-1}^1 P_r(SI) - P(SI) dSI}{\int_{-1}^1 P_r(SI) - P_{fd}(SI) dSI} = \frac{\int_{-1}^1 P_r(SI) - P(SI) dSI}{\int_{-1}^1 P_r(SI) dSI} \quad (3.20)$$

$$FSI_2 = \frac{\mu - \mu_r}{\mu_{fd} - \mu_r} \quad (3.21)$$

$$FSI_3 = \frac{m - m_r}{m_{fd} - m_r} \quad (3.22)$$

where $P_r(SI) = \int_{-1}^{SI} p_r(x) dx$ and $P(SI) = \int_{-1}^{SI} p(x) dx$ are the cumulative SI distribution functions (CDF) for a random spiking condition and for the neuron under study respectively, $P_{fd}(SI) = u(SI-1)$ is the CDF for the ideal feature detector neuron, and $u(x)$ is the unit step function. The symbols μ and m correspond to the mean value and the median value of the SI distributions. The subscripts r and fd designate the random and feature detector control neurons respectively.

The construction and significance of these metrics is depicted in Fig. 14. In all instances the CDF has a monotonically increasing sigmoidal shape. The transition points at which the CDF reaches a value of 0.5 (the median value) designates the value of the SI at which the PDF has accumulated half of the total probability. Example CDFs for the considered conditions $P_r(SI)$, $P_{fd}(SI)$, and $P(SI)$ are shown in Fig. 14 B. For the random spiking condition (dashed-dotted line) the median value of SI is reached near zero whereas for the ideal feature detector (dashed line) it is reached at unity. For the neuron under study (continuous line) this transition generally occurs at an intermediate value of SI , usually between zero and unity. For an ideal feature detector, this transition point occurs at $SI=+1$. Likewise, the mean value of the respective distributions obey similar rules assuming that the distributions for the null, feature detector, and experimental condition are not skewed (they are symmetric). The FSI (FSI_2 and FSI_3) derived directly from the mean (median) value of these three distributions designates distance between the means (median) of the null and experimental condition (red arrow; Fig. 14 A) normalized by the distance between the mean (median) of the null and feature

detector condition (blue arrow; Fig. 14 A). Since $\mu_r \approx 0$ ($m_r \approx 0$), $\mu_{fd} = 1$ ($m_{fd} = 1$), and $0 < \mu_p < 1$ ($0 < m_p < 1$) these *FSI* metrics takes values between zero and unity. Values near zero indicate that the neuron's response is non-specific whereas values near unity indicate a high degree of specificity to sound patterns that resemble the neuron's *STRF*.

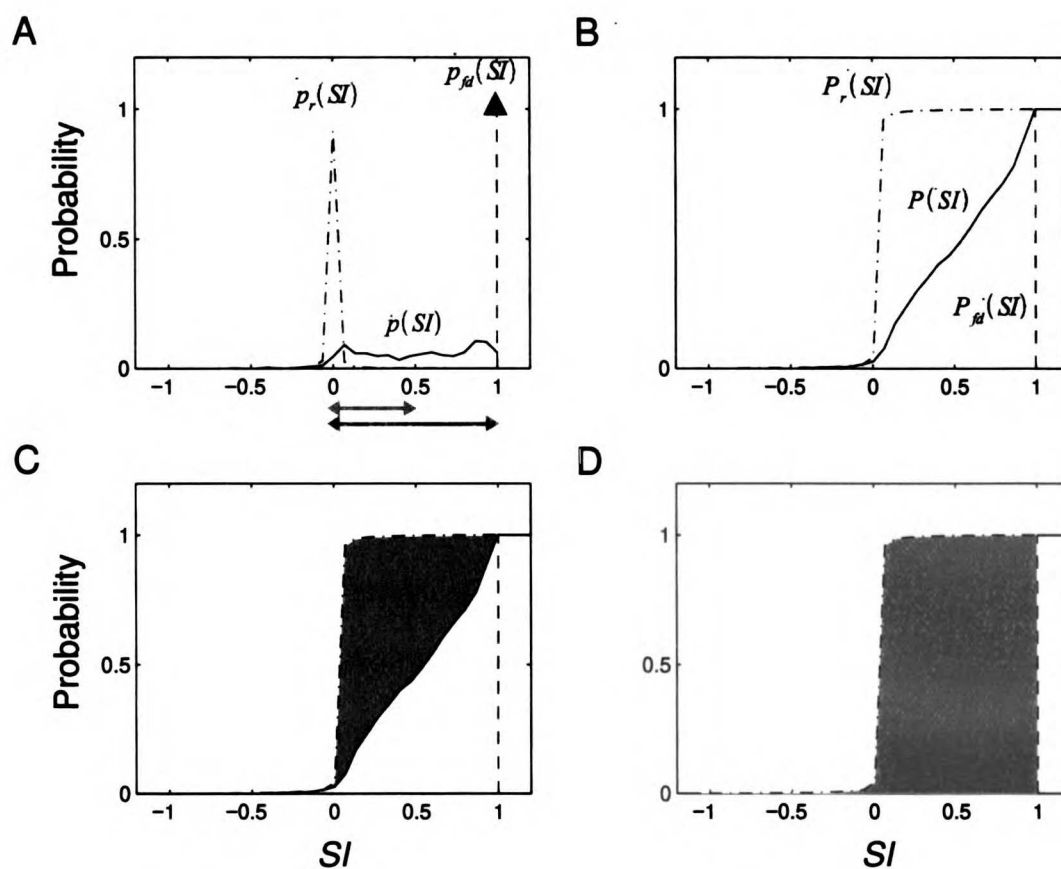


Figure 14: Defining the feature selectivity index (*FSI*). The definition of the *FSI* is derived from the neuron's *SI* distribution function, $p(SI)$, and the expected *SI* distribution for an ideal feature detector, $p_{fd}(SI)$, and a randomly firing neuron, $p_r(SI)$, both of which serve as reference points (A). The reference *SI* distribution

1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100

for an ideal feature detector neuron is an impulse centered at +1. The reference distribution for a randomly firing neuron is centered about zero. The tested neuron's SI distribution, $p(SI)$, is generally broad taking values between zero and unity. The simplest versions of the FSI metric (FSI_2 and FSI_3) are derived directly from the mean or median values of the respective SI distributions. These metrics are derived by computing the ratio of the distances between the means (median) of the tested neuron and the random neuron (shown as a red arrow, A), and between the feature detector and random neuron (shown as a blue arrow, A). An alternate FSI metric is derived by considering the cumulative SI distribution functions (FSI_1 , B). This metric is derived by measuring the area in-between the tested neuron's CDF and the random neuron's CDF (red region, C) normalized by the area in-between the ideal feature detector and random neuron's CDF (blue region, D).

Although, the value of these simple metrics serves as an indicator of the overall response characteristics of a neuron, a more accurate metric is obtained by considering the entire shape of the respective CDF curves. This is done by considering the accumulated area between the respective CDF curves to derive the metric FSI_1 . The denominator of Eq. (3.20) designates the area in-between the sigmoidal CDF curves for the random spiking neuron and the ideal feature detector neuron conditions (Fig. 14 D; blue region). The numerator alternately corresponds to the accumulated area in-between the CDFs for the random spiking neuron and the neuron under study (Fig. 14 C; red region). Thus, the FSI_1 metric has numerical value between zero and unity. FSI_1 values near unity are indicative of a neuron with "feature detector" like qualities whereas values of FSI_1 near zero are indicative of a neuron which fires with little precision.

3.11 Binaural FSI Metric

The described procedure for deriving the feature selectivity index of a given neuron was derived on the assumption that the neuron's *STRF* was measured monaurally. For neurons that responded only to one ear (contra or ipsi), the described procedure can be applied directly to the sound channel that produced a significant *STRF*. If, however, the given neuron has a significant contralateral and an ipsilateral *STRF* the described procedure can in theory significantly underestimate the actual feature specificity of the given neuron.

As an example, consider a highly selective neuron which responds independently for the contralateral and ipsilateral ears. Given that the input stimuli are bilaterally independent, a fraction of the overall spike train consists of events that are specific only for the contralateral ear. The remaining spikes are specific only for the ipsilateral ear. Given independence of the contra and ipsilateral stimuli and independence of the responses to both ears, the spikes that show specificity for the contralateral ear are nonspecific for the ipsilateral ear and vice versa. Thus when one constructs the feature selectivity histogram for the contralateral *STRF*, it will show a large peak centered about zero (non-specific) arising from neuronal events that are actually specific for the ipsilateral ear and vice versa (see Fig. 15). Thus to properly quantify the specificity of the response for such a neuron one needs to take into account the interdependence between the contralateral and ipsilateral ear.

As a first step to remedy the limitations of computing a monaural *FSI* measurement for a binaural response condition, we consider joint *FSI* measurements for the contralateral and ipsilateral ears. Given a set of neuronal responses for a bilateral

stimulus condition, the outlined procedure of section 3.8 and 3.10 can be used to compute the FSI first for the contralateral ear, FSI_c , and subsequently for the ipsilateral ear, FSI_i . The overall specificity of a neuron can then be described as binaural feature selectivity vector, $FSI_b = [FSI_c \ FSI_i]$, where the subscript b designate a binaural feature selectivity measurement. The subscripts c and i designate the contralateral and ipsilateral ears respectively. To assign a numerical value to this vector the vector norm is computed:

$$\|FSI_b\| = \sqrt{FSI_c^2 + FSI_i^2} \quad (3.23)$$

This metric corresponds to the vector distance relative to reference binaural feature selectivity index value of $[0 \ 0]$. Since the contralateral and ipsilateral FSI can assume values between zero and unity, the quantity $\|FSI_b\|$ assumes values between zero and $\sqrt{2}$.

We further extend the procedure for estimating the monaural SI distribution, $p(SI)$, and the corresponding monaural FSI by considering the joint SI distribution for the contralateral and ipsilateral responses. For each neuronal event the similarity index is computed as described in section 3.8 for both contralateral and ipsilateral ears. Thus for each spike two similarity index values, one for the contralateral and the other for the ipsilateral ears, are derived. The two values of the similarity index, SI_c and SI_i , are then used to construct the joint similarity index distribution, $p(SI_c, SI_i)$.

The joint cumulative SI distributions is computed as

$$P(SI_c, SI_i) = \int_{-1}^{SI_c} \int_{-1}^{SI_i} p(x_c, x_i) dx_c dx_i .$$

This procedure is repeated for the null conditions of a random firing neuron, $P_r(SI_c, SI_i)$, and the ideal feature detector

neuron, $P_{fd}(SI_c, SI_i)$. For the later case, the joint SI distribution

$$p_{fd}(SI_c, SI_i) = \delta(SI_c - 1) \delta(SI_i - 1)$$

takes the form a two dimensional impulse centered at the vector location $[SI_c \ SI_i] = [1 \ 1]$. The corresponding joint CDF is given by

$$P_{fd}(SI_c, SI_i) = u(SI_c - 1) \cdot u(SI_i - 1)$$

where as before $u(x)$ is the unit step function.

Examples of the monaural PDFs, the joint PDFs, and the corresponding joint CDFs are shown for the neuron of Fig. 1 C/D in Fig. 15. The monaural PDF for this neuron, $p(SI_c)$ and $p(SI_i)$, has a non-specific peak centered at SI values near zero

for the contralateral ear. A careful look at the joint SI PDF, $p(SI_c, SI_i)$, reveals that this "non-specific" responses is actually specific for the ipsilateral ear. The joint CDF,

$$P(SI_c, SI_i)$$

takes the shape of a two dimensional sigmoidal surface (Fig. 16). For reference, the joint PDF and CDF are also shown for the null condition, $p_r(SI_c, SI_i)$, in Fig. 15 B and Fig. 16 B respectively.

AMERICAN
INDUSTRIAL
CORPORATION

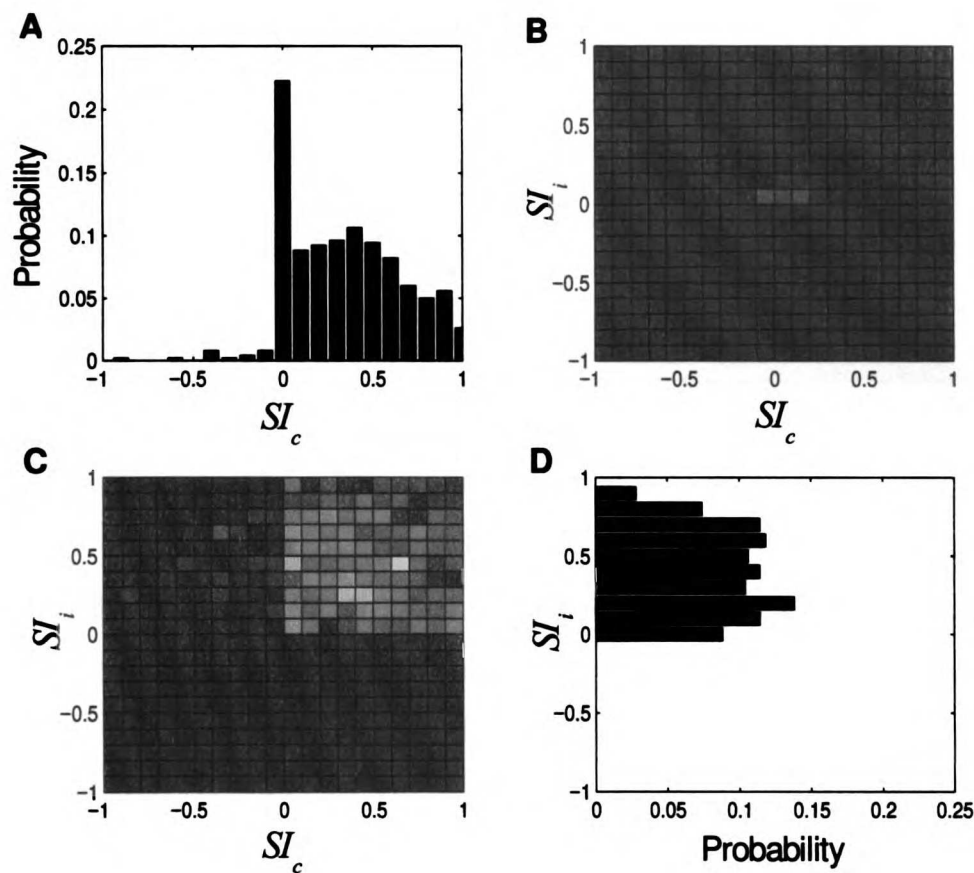


Figure 15: Relationship between the binaural SI PDF, $p(SI_c, SI_i)$ (C), and the monaural SI PDFs, $p(SI_c)$ (A) and $p(SI_i)$ (D). Contralateral (A) and ipsilateral (D) monaural SI PDFs shown for the binaural f-neuron of Fig. 1 C/D. Both the contralateral and ipsilateral responses have SI PDFs which are suggestive of feature sensitive type responses. The contralateral SI PDF has a "non-specific" response peak centered at $SI_c=1$. The joint SI PDF (C) displays the interrelation between the contralateral and ipsilateral responses. The joint SI PDF reveals that the "non-specific" contralateral response is actually specific for the ipsilateral ear. Unlike the joint PDF for the tested neuron, which has contralateral and ipsilateral SI values that are skewed towards unity, the joint SI PDF for the random control condition (B), $p_r(SI_c, SI_i)$, is narrowly distributed (impulsive) and centered about $SI_c=0$ and $SI_i=0$.

ANNOUNT JOURN

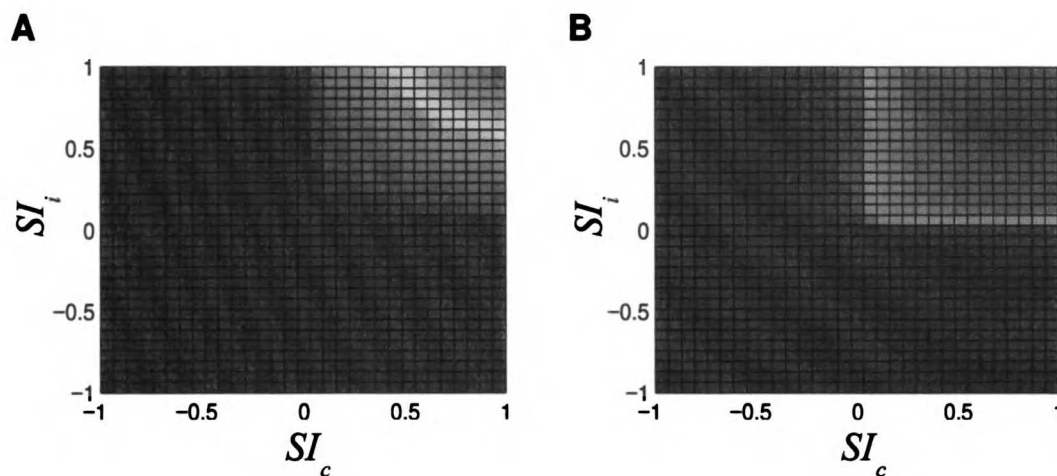


Figure 16: The joint SI cumulative distribution function (CDF) is obtained directly from the neuron's joint SI PDF by integrating across the contralateral and ipsilateral axis. An example CDFs is shown for the neuron of Fig. 1 C/D (same neuron shown in Fig. 15),

$P(SI_c, SI_i)$ (A), and for its random control condition, $P_r(SI_c, SI_i)$ (B). The neuron's joint CDF (A) resembles a sigmoidal surface with a broad transition occurring at SI_c and SI_i values between zero and unity. The random control CDF (B) also resembles a sigmoidal surface but its transition is sharp and centered about $SI_c=0$ and $SI_i=0$. This neuron has a high FSI value (0.62 for the dynamic ripple), consistent with its feature sensitive response properties.

The binaural feature selectivity index is constructed as for the monaural case by considering the volumes in-between the CDFs for the ideal feature detector neuron, the null random neuron, and the neuron under investigation. The binaural FSI is obtained as

$$FSI_b = \frac{\int_{-1}^{-1} \int_{-1}^{-1} P_r(SI_c, SI_i) - P(SI_c, SI_i) dSI_c dSI_i}{\int_{-1}^{-1} \int_{-1}^{-1} P_r(SI_c, SI_i) - P_{fd}(SI_c, SI_i) dSI_c dSI_i} \quad (3.24)$$

1907 12 11

$$= \frac{\int_{-1}^{-1} \int_{-1}^{-1} P_r(SI_c, SI_i) - P(SI_c, SI_i) dSI_c dSI_i}{\int_{-1}^{-1} \int_{-1}^{-1} P_r(SI_c, SI_i) dSI_c dSI_i}$$

where the integrated volume in-between SI values of -1 and $+1$ is zero for the ideal feature detector neuron: $\int_{-1}^1 \int_{-1}^1 P_{fd}(SI_c, SI_i) dSI_c dSI_i = 0$. Since the SI distribution for the feature detector condition is centered at the vector location $[SI_c \ SI_i] = [1 \ 1]$, the metric FSI_b can assume values in between zero and $\sqrt{2}$.

3.12 Distribution of FSI Values

To determine the degree of response specificity for the ripple noise and dynamic ripple conditions, values of FSI_b were estimated for all neurons for both conditions. Histograms and a scatter plot are shown in Fig. 17. The distribution of FSI_b values for the dynamic ripple stimulus is shown in Fig. 17 A. The distribution for this condition is bimodally distributed with a threshold that was visually determined at $FSI_b = 0.4$. The vast majority of neurons ($n=51$) had low FSI values (mean value of 0.18) whereas 11 neurons had significantly higher values of FSI ($FSI_b > 0.4$ and mean value of 0.56).

In Fig. 17 B, the distribution of FSI_b values for the ripple noise shows a large amount of scatter when compared to the distribution for the dynamic ripple. For the ripple noise condition, only seven neurons had values of FSI_b greater than 0.4. Of these, 1 neuron had an FSI_b that was much higher for the ripple noise (0.75 ripple

AMERICAN
1000
1000

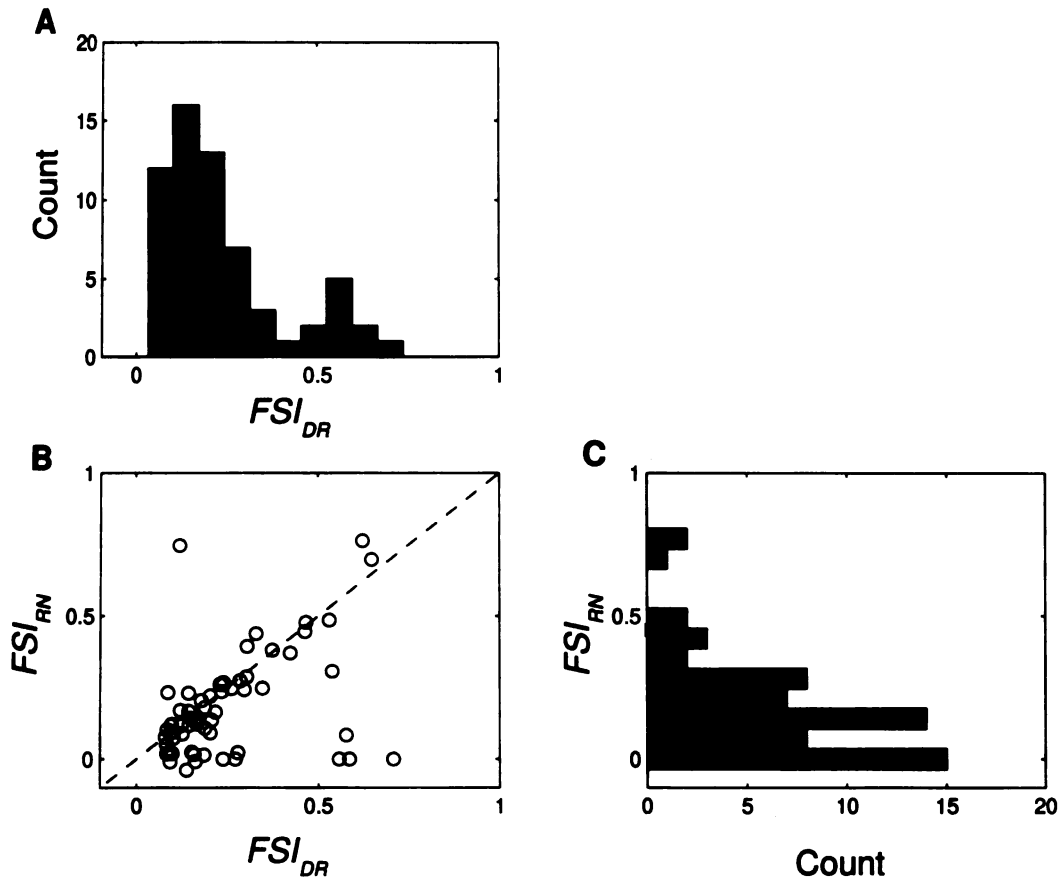


Figure 17: Measured FSI_b values for the ripple noise and dynamic ripple conditions.

(A) The distribution of FSI_b values for the dynamic ripple condition is bimodally distributed. (C) The distribution of FSI_b values for the ripple noise stimulus is highly scattered. (B) Scatter plot showing the interrelation of FSI_b values for the ripple noise and dynamic ripple.

noise versus 0.12 for the dynamic ripple). By this measure this neuron seemed to respond more precisely to the ripple noise. At odds with this observation, however, was the fact that this neuron had a significantly higher firing rate and difference response strength for the dynamic ripple condition (Rate=12.62 spikes/sec, $\sigma_r=0.84$ spikes/sec,

and $\Delta RF=8.4$ spikes/sec) versus the ripple noise (rate=0.03 spikes/sec, $\sigma_r=0.015$ spikes/sec, and $\Delta RF=0.05$ spikes/sec). Of the remaining neurons, $n=48$ responded with similar specificity for both conditions and a smaller subset of neurons ($n=7$) had FSI_b values for the dynamic ripple condition that were more than ten times larger compared to the ripple noise.

3.13 Independence of the Binaural SI Distribution

The chosen reference position for the ideal feature detector neuron,

$$[SI_c \ SI_i]=[1 \ 1] ,$$

requires that such a neuron respond if and only if the optimal contralateral and ipsilateral sounds are presented simultaneously. Thus, for such a neuron to respond, the sound features which identically match the neuron's binaural *STRFs* must be presented. Given that ICC neuron's respond most strongly to the contralateral inputs (e.g. Irvine and Gao 1990) it is unlikely that such binaural specificity will be encountered. Thus expecting a value of FSI_b as high as $\sqrt{2}$ is likely an unfair expectation.

To further investigate the degree of binaural specificity, the independence for the contralateral and ipsilateral responses was measured by constructing a separable FSI_b measure FSI_b^* . We ask: Are the joint PDFs and CDFs separable independent functions for the contralateral and ipsilateral ears? Do neurons require structured inputs that coactivate both ears, or are the contralateral and ipsilateral responses independent? To address this question, we consider Eq. (3.24) and assume that the joint similarity index

CDF is separable: $P(SI_c, SI_i) = P(SI_c)P(SI_i)$, implying that the SI_c and SI_i are independent of each other. By construction it is clear that $P_r(SI_c, SI_i) = P_r(SI_c)P_r(SI_i)$ since the input sounds for the contralateral and ipsilateral ears are independent and since the neuronal spike trains were randomly generated for this condition. Substituting into Eq. 3.24, the expected FSI_b for a neuron with independent joint FSI distribution is given by

$$FSI_b^* = \frac{\int_{-1}^{-1} P_r(SI_c) dSI_c \cdot \int_{-1}^{-1} P_r(SI_i) dSI_i - \int_{-1}^{-1} P(SI_c) dSI_c \cdot \int_{-1}^{-1} P(SI_i) dSI_i}{\int_{-1}^{-1} P_r(SI_c) dSI_c \cdot \int_{-1}^{-1} P_r(SI_i) dSI_i} \quad (3.25)$$

By default we expect that $FSI_b = FSI_b^*$ if and only if the neuron's joint PDF is separable. Thus, if this condition is satisfied, the measured values of SI_c and SI_i are independent of each other.

Fig. 18 shows a scatter plot between FSI_b and FSI_b^* . The two parameters are highly correlated ($r = 0.985 \pm 0.015$, linear regression: $B=0.98$, $A=0.01$) indicating that the binaural interactions for the similarity index parameter are statistically independent between the two ears.

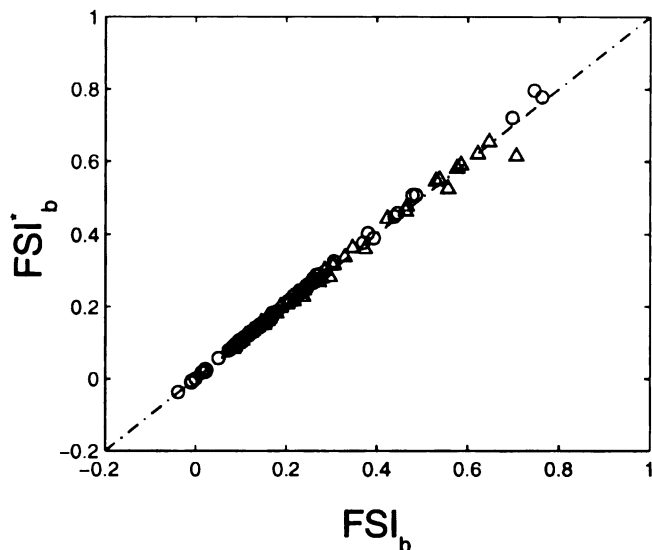


Figure 18: Separability of the binaural similarity index histogram. Binaural feature selectivity index, FSI_b , versus the separable binaural feature selectivity index, FSI'_b . Shown for the moving ripple (triangles) and the ripple noise (circles) stimulus. Both metrics are highly correlated ($r^2=0.97$) with unity slope (linear regression $B=0.98 \pm 0.01$, $p < 0.001$ confidence interval) indicating that the contralateral and ipsilateral binaural *SI* measures are statistically independent of each other.

3.14 The Ripple Transfer Function

As can be seen from the preceding chapters, the *STRF* is a useful descriptors that allows one to determine the response area of a neuron. In addition, this descriptor is useful for describing the patterning of spectral and temporal excitation, inhibition, binaurality, spectro-temporal specificity, and the overall response strength to complex spectro-temporal stimulus as demonstrated in the preceding sections. Thus, if one is interested in characterizing the response area of a neuron and its spectro-temporal

patterning along the sensory epithelium the *STRF* is likely the most viable descriptor.

As with any descriptor, however, the *STRF* also has both theoretical and practical limitations. Although a number of response parameters (such as best frequency, temporal delay, response strength, and binaurality) can be derived directly from the *STRF*, higher-order response parameters such as the neuron's best ripple frequency and temporal modulation rate, are not directly accessible. To evaluate a neuron's response preference with respect to these higher-order stimulus parameters, we therefore consider the spectro-temporal ripple transfer function (*RTF*) (Klein *et al.* 2000).

The ripple transfer function is a two dimensional *Fourier* representation of the *STRF*. The *RTF* describes the input-output characteristic of a linear neuron to any arbitrary ripple signal and combination of ripple parameters F_m and Ω . Formally the *RTF* exists only for the special case of a linear integrating neuron. However, as for the *STRF*, its applicability is not limited to linear systems.

Recall that a neuron's *STRF* is analogous to its impulse response described along time and along the sensory epithelium (i.e. the frequency axis). Alternately, the *RTF* describes neuron's transfer function characteristics as a function of the temporal modulation rate, F_m , and spectral envelope frequency, Ω . Together the *STRF* and *RTF* form a Fourier transform pair

$$STRF(t, X) \quad \leftarrow \mathfrak{F}_2 \rightarrow \quad H(F_m, \Omega) \quad (3.26)$$

where the forward and inverse transforms are defined by:

$$\mathfrak{F}_2[\cdot] = \iint \cdot e^{-j2\pi(F_m t + \Omega X)} dt dX \quad \text{and} \quad \mathfrak{F}_2^{-1}[\cdot] = \iint \cdot e^{j2\pi(F_m t + \Omega X)} dF_m d\Omega \quad \text{respectively.}$$

Upon performing this transformation the time axis of the *STRF* is converted to the temporal modulation rate (F_m) axis of the *RTF* and the spectral axis (X) is converted to the ripple frequency axis (Ω) of the *RTF* (see Fig. 19).

When using the *RTF*, the neuron's input-output characteristic is expressed directly as a complex quantity and is described as a function of F_m and Ω . For a linear neuron the *RTF* can be expressed in the general form

$$H(F_m, \Omega) = M(F_m, \Omega) \exp[-j\Phi(F_m, \Omega)] \quad (3.27)$$

where $M(F_m, \Omega)$ is the magnitude response and $\Phi(F_m, \Omega)$ is the phase response.

The magnitude of the *RTF* determines the strength of the response as a function of spectral, Ω , and temporal, F_m , frequencies for a given input stimulus of fixed energy. If, for example, the magnitude response of a neuron has a gain of three units, for a given ripple combination, the response to that ripple stimulus will be three times as strong in amplitude as the incoming input signal. Alternately, the phase response determines the spectro-temporal patterning observed directly in the *STRF* as a function temporal and spectral frequency. As an example, consider the off-on neuron of Fig. 4 A/B. One can imagine a similar neuron, with identical spectro-temporal shape, but with an inverted on-off response profile. Such a neuron has identical magnitude response and differs only in its spectro-temporal phase profile by a constant of π radians (180°).

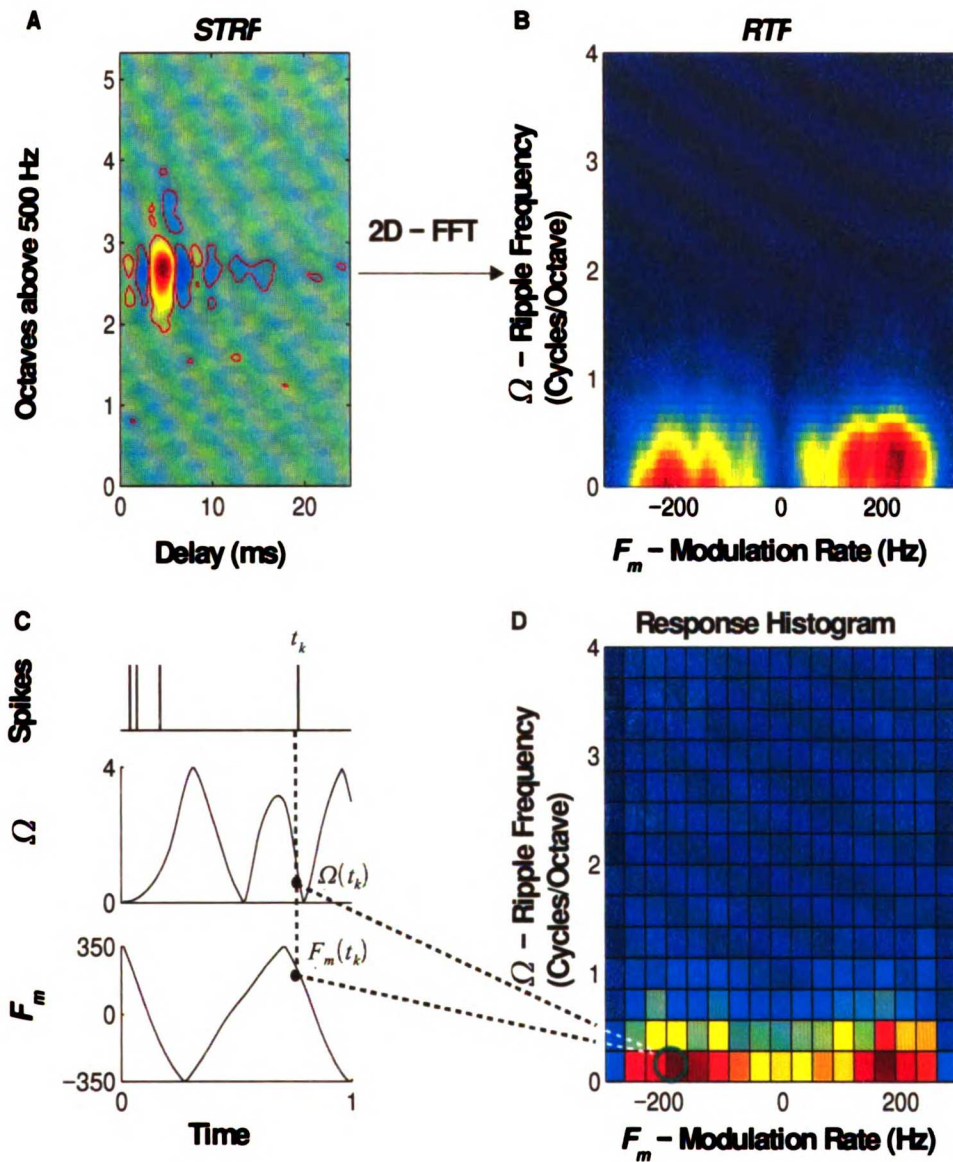


Figure 19: The *ripple transfer function (RTF)* (B) is a two dimensional *Fourier* representation of the *STRF* (A). The *RTF* is derived directly from the significant *STRF* via a two dimensional Fourier transformation. This descriptor depicts time-locked energy in the response as a function of temporal modulation rate, F_m , and spectral envelope frequency, Ω . Red indicates parameter combinations which evoked a strong

time-locked response whereas blue indicates a weak response. We also computed the *conditional-response histogram* (D) in order to quantify nonlinear response characteristics that are not time-locked. The response histogram is similar to the *RTF* but it is computed directly from the instantaneous dynamic ripple parameters ($\Omega(t)$ and $F_m(t)$) by counting responses to each parameter combinations (C) (only possible for the dynamic ripple). For each neural event, such as the one depicted by the dotted line in (C), the spectral and temporal dynamic ripple parameters, $\Omega(t_k)$ and $F_m(t_k)$, are determined at the time instance of the neural spike, t_k . The values of Ω and F_m are then used to increment the corresponding bin in the joint histogram by +1 (D). For a linear system, the *RTF* magnitude response and the response histogram are virtually identical. The two descriptors differ since the response histogram can access nonlinear information.

Conceptually the magnitude response of the ripple transfer function describes the spectro-temporal filtering characteristics of the neuron. Regimes in the magnitude response, $M(F_m, \Omega)$, with non-zero values designate combinations of F_m and Ω that efficiently activate the neuron. Likewise, a given neuron responds weakly to spectro-temporal combinations wherever the magnitude response is zero-valued. Using this simplified description, any given neuron can therefore be described as a spectro-temporal filter which rejects sounds with certain parameter combinations and accepts all others.

To visualize the neuron's response as a function of F_m and Ω , we consider the magnitude response. The magnitude response is obtained directly from the *RTF*,

$H(F_m, \Omega)$, as

$$M(F_m, \Omega) = \|H(F_m, \Omega)\| = \sqrt{H(F_m, \Omega) \cdot H(F_m, \Omega)^*} \quad (3.28)$$

where $H(F_m, \Omega)^*$ is the complex conjugate *RTF* and the neuron's *RTF* is derived from its *STRF* via a 2-D Fourier transform.

Given that the *RTF* defines the spectro-temporal filtering operation of a neuron, the overall patterning of the response can be determined by independently considering the spectral and temporal dimensions. Thus the neuron's stimulus-response transfer function characteristics can be described as a lowpass and/or bandpass filtering operation along the spectral and/or temporal dimension. Using this convention to categorize spectro-temporal response properties, a total of four filtering combinations are possible: LL, LB, BL, BB, where the first letter designates the temporal filtering operation and the second letter designates the type of spectral selectivity. The symbols L and B designate lowpass and bandpass filtering operation respectively. If one additionally considers the possibility that the neuron's *STRF* can be spectro-temporally separable or inseparable a total of eight possible combinations are presented, LL, LB, BL, BB, LLS, LBS, BLI, BBI, where the third symbol, S or I, designates separable and inseparable respectively.

Example *STRFs* and their corresponding *RTFs* are shown for the possible different scenarios in Fig. 20. Spectral bandpass filtering neurons have alternating patterns of excitation and inhibition along the spectral axis (F and D). Likewise, neurons that show bandpass *RTFs* along the modulation rate axis, F_m , axis (A, C, D, and E)

have *STRFs* with "on-off" (A, D, and E) or "off-on-off" (C) temporal profiles. Neurons with lowpass spectral (A, C, E, and F) and/or temporal (B and F) *RTFs* lack alternation patterns of "on" and "off" components along the appropriate dimension. Finally, inseparable neurons (D and E) have asymmetric *RTFs* and oblique *STRFs* – indicative of a preference for upward going (E, negative F_m) or downward going (D, positive F_m) ripple profiles.

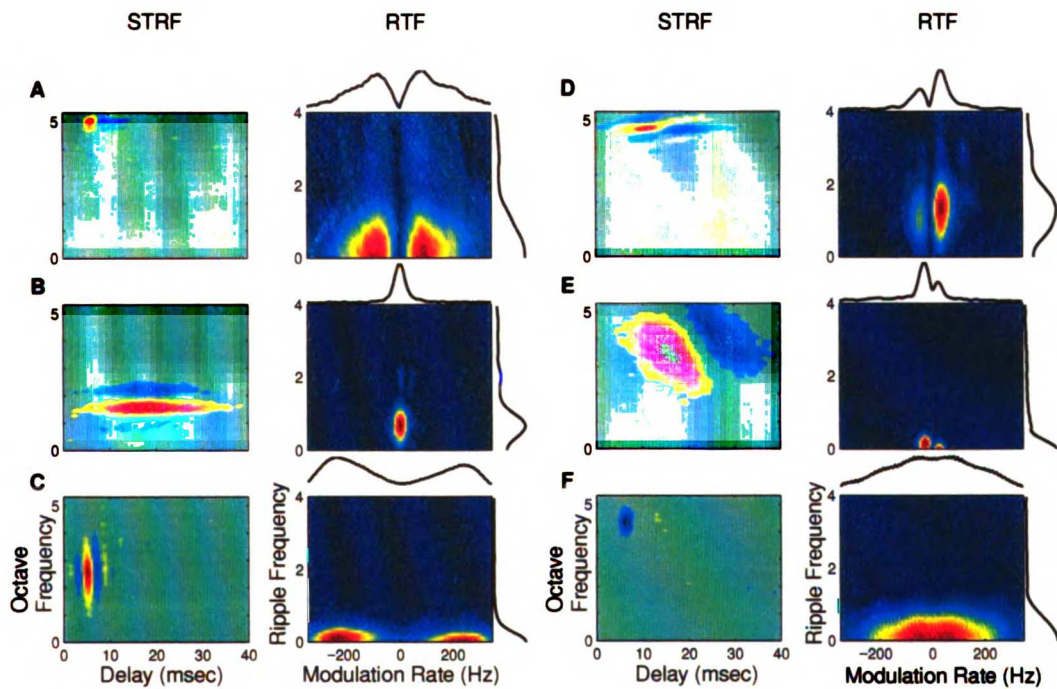


Figure 20: Example *STRFs* (left) and the corresponding *RTF* (right) – chosen to reflect the diversity of ripple transfer functions observed for the population. Projections of the *RTF* are shown for reference along the modulation rate and ripple frequency axis. Neurons with BL response selectivity (A and C) are characterized by an "on-off" temporal profile and lack alternating patterns of excitation and inhibition along the spectral axis of the *STRF*. The neuron of (B) shows almost orthogonal LB response characteristics to the neurons of (A and C). Its temporal patterning is characterized by

an "on" component and its spectral "off-on-off" profile. Cells with inseparable *STRFs* have asymmetric *RTFs* (D and E). The neuron of D responds best to downward going ripple profiles (positive F_m) and has a BBI response pattern. The neuron of (E) has a BLI response to upward going ripple profiles (negative F_m). Neurons with LL responses (F) lack alternating patterns of excitation and inhibition along the spectral and temporal axis. Shown for a neuron with purely inhibitory "off" response profile.

3.15 Non-phase-locked Neurons (C-Neurons)

A basic requirement for computing an *STRFs* is that the neuron under investigation time-lock to the stimulus spectro-temporal envelope. The term phase- or time-locking is used to describe a neuron's ability to follow, on a cycle to cycle basis, the amplitude modulations of the stimulus. Sinusoidal amplitude modulations studies show that many auditory neurons in the ICC and along the auditory pathway phase-lock to the stimulus modulation waveform (Plomp 1983; Rees and Møller 1983; Schreiner, Urbas, and Mehrgardt 1983; Jones and Palmer 1987; Rees and Møller 1987; Krishna and Semple 2000). Consequently, it is of no surprise that a large number of neurons in this study phase-locked to the spectro-temporal envelope and consistently produced statistically significant *STRFs*.

Up to now, we have described only neurons that show statistically significant *STRFs*, indicative of quasi-linear processing and time-locking to the spectro-temporal envelope. In some instances ($n=22$ of 84), however, the estimation of the *STRF* failed to produce statistically significant *STRFs* ($p<0.002$) with a distinct spectro-temporal patterning, despite a significant overall firing rate (mean firing rate=7.5 spikes/sec). Two

possible explanations are presented: first it is possible that the observed neurons were spontaneously firing and did not respond in a time-dependent manner to the spectro-temporal modulations of the dynamic ripple and ripple noise stimulus. Alternately, it is also possible that these neurons respond in a highly nonlinear manner to these sounds and the *STRF* is an insufficient functional descriptor. Such would be the case for even-order nonlinearities, dynamic nonlinearities, and nonlinearities arising from combination products. If this is so, we can overcome this limitation of the *STRF* by devising an alternate functional descriptor that can capture the nonlinear stimulus-response transformation performed by these neurons.

As described in the previous section, an alternate approach for characterizing the stimulus-response relationship of a neuron is to compute the ripple transfer function of the neuron directly in the spectro-temporal frequency domain. This method, however, is limited by the fact that the described procedure for computing the magnitude response function, $M(F_m, \Omega) = \sqrt{H(F_m, \Omega)H^*(F_m, \Omega)}$, requires a significant *STRF* for its computation. Since the described neurons do not produce significant *STRFs* this procedure can not be used directly.

A closely related method for estimating the magnitude response function involves performing a spike-triggered average with respect to the stimulus parameters for the dynamic ripple stimulus to construct a conditional-response histogram (refer to Fig. 19). This method accumulates signal parameters rather than the spectro-temporal stimulus waveform and is, therefore, insensitive to spike timing jitter, unlike the conventional *STRF* method. Given that the time varying parameters, $\Omega(t)$ and $F_m(t)$, for the dynamic ripple stimulus are known a priori we use these to estimate the conditional

distribution function $P(F_m, \Omega | t_n)$. This distribution describes the probability of observing a set of ripple parameters, F_m and Ω , given the presence of a spike at time t_n . For a linear time-invariant system this distribution function should be identical to the systems magnitude response.

We approximate this distribution by discretizing the spectro-temporal ripple domain into bins of resolution $\Delta F_m \times \Delta \Omega$. For a sequence of neuronal spikes, t_n , we perform a spike-triggered average

$$P_{kl} = \sum_{n=1}^N I(k \Delta F_m \leq F_m(t_n) < (k+1) \Delta F_m) \cdot I(l \Delta \Omega \leq \Omega(t_n) < (l+1) \Delta \Omega) \quad (3.29)$$

where P_{kl} is the discrete version of $P(F_m, \Omega | t_n)$ and $I(\cdot)$ is the identity function.

The identify function takes a value of unity whenever the condition inside its argument is satisfied. Otherwise it assumes a value of zero. Thus for any given bin of P_{kl} , the response histogram is incremented by +1 if and only if the instantaneous parameters,

$F_m(t_n)$ and $\Omega(t_n)$, fall within the required intervals,

$k \Delta F_m \leq F_m(t_n) < (k+1) \Delta F_m$ and $l \Delta \Omega \leq \Omega(t_n) < (l+1) \Delta \Omega$, at the time of the neuronal spike, t_n (see Fig. 19 for further explanation).

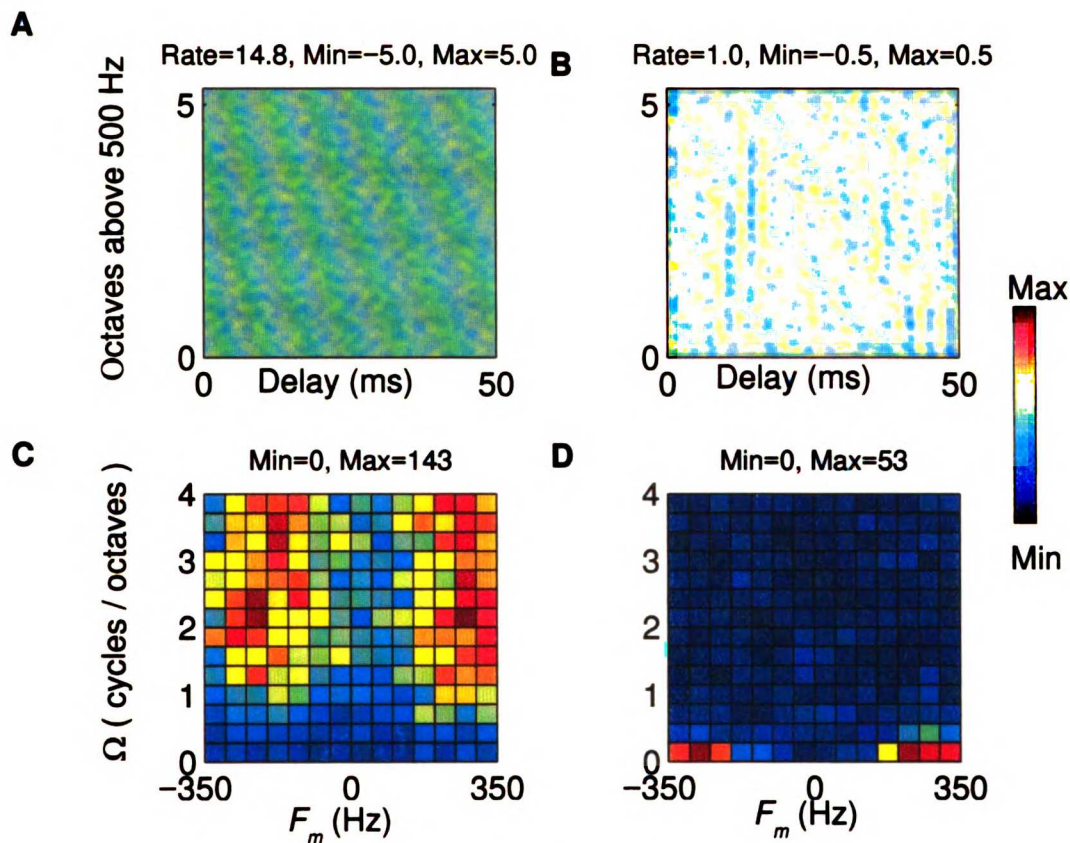


Figure 21: C-neurons are characterized by a highly nonlinear stimulus-response relationship. The *STRF* for such cells is absent or weak providing little or no information ($p < 0.002$). Two such *STRFs* are shown in A and B. Note that no significant spectro-temporal patterns stand out. Since the *STRFs* provide no information about the response characteristics of these neurons, spectro-temporal feature selectivity for these neurons was established by examining the conditional-response histogram, which always showed a tuned response not observed directly from the *STRF* or the *RTF*. The response-conditional histograms is shown in C and D for the c-units of A and B. The c-neuron shown in C, responded best to high ripple densities (> 1 cycle/octave) and temporal modulation rates greater than 50 Hz. The neuron shown in D responded best to low ripple densities (< 0.25 cycles/octave) and to fast temporal modulation rates (~ 200 –300 Hz). These responses can not be visualized using the *STRF*.

Using the conditional–response histogram, we tested whether the observed neurons do respond selectively to complex sound attributes but do not meet the necessary stimulus–response requirements for producing *STRFs*. Surprisingly, the conditional–response histogram for these neurons revealed strong responses to particular stimulus parameter combinations (Fig. 21 C and D) despite the lack of linear time–locking to the spectro–temporal envelope (resulting in no *STRF*). Thus the responses of these neurons can not follow the fast spectro–temporal modulations of the stimulus envelope (up to 350 Hz) but were able to track very slow changes of the stimulus parameters (1.5 Hz for the temporal modulation rate and 3 Hz for the ripple density).

For the examples of Fig. 21, the *STRFs* are absent (A and B) despite a significant overall firing rate. If one were to judge these neurons based on the *STRF* and mean firing rate alone one inevitably concludes that these neurons are spontaneously firing and do not serve any functional purpose for encoding information about complex sound attributes. A more careful evaluation of the stimulus–response relationship, however, revealed that this is not so, since the neurons show tuned responses when described using the conditional–response histogram. Because of the analogous functional properties to complex cells in the visual cortex, which likewise do not produce *STRFs* (see discussion), these neurons are referred to as c–neurons.

3.16 Spectro–Temporal Population Statistics

To evaluate the processing capabilities of the three identified cell types (s–, f–, and c–neurons), we measured the best ripple density and best modulation rate parameters for each neuron. The best ripple density and temporal modulation rate are defined by the

"hot-spot" in the *RTF* with strongest magnitude. Since most neurons responded symmetrically to upward going (negative F_m) and downward going (positive F_m) ripples, two values of the best parameters were extracted (one for each quadrant of the *RTF*). For neurons that did not phase-lock to the spectro-temporal envelope, and consequently did not produce *STRFs* and *RTFs*, these parameters were estimated directly from the conditional-response histogram for the dynamic ripple.

The three identified cell types differ not only in the described response characteristics but also in their encoded spectro-temporal parameter range. A scatter plot of the preferred temporal modulation rate, F_m , and preferred spectral envelope frequency, Ω , (Fig. 22) reveals that time-locked neurons (s-cells and f-cells) appear to trade temporal for spectral resolution. Cells that responded selectively to fast temporal modulations (high F_m) were generally spectrally broad (low values of Ω), while narrow spectral resolutions (high Ω) were seen only in cells that were temporally slow (low F_m). This trade-off was not seen for c-neurons, which displayed a wide range of spectral and temporal preferences.

NOTHING

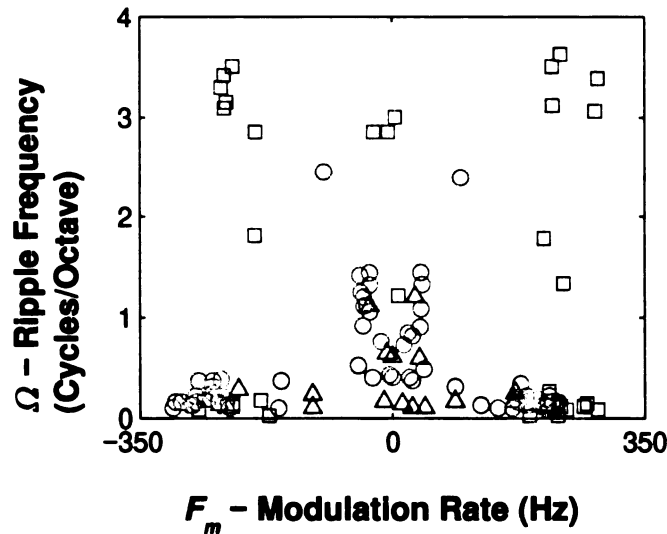


Figure 22: Scatter plot of the preferred ripple frequency and best temporal modulation rate (A) for the observed neural response types (\circ s-neurons, Δ f-neurons, \square c-neurons). S- and f- neurons showed overlapping best response areas. C-neurons, which lack the time-locked responses observed for s- and f-neurons, responded preferentially to sound instances with fast temporal modulation rates, F_m , and/or high ripple frequencies, Ω .

3.17 Discussion

The main goal of this study was to understand how the time-varying spectrum of complex sounds, such as speech and animal vocalizations, are represented and processed at the level of the feline auditory midbrain. We describe a number of nonlinear processing strategies employed for the analysis of the time-varying spectrum, which have traditionally been described as occurring only in higher-order cortical stations and only for acoustically specialized animals. Aside from the overall utility of using the *STRF* for describing the "linear" stimulus-response relationship of central auditory neurons, we have demonstrated how a principled approach, grounded on linear (Klein *et*

ADDITIONAL FORM

al. 2000) and nonlinear systems theory (chapter 2), can be utilized directly for identifying complex nonlinearities. Furthermore, we demonstrate the importance of stimulus design and ecological considerations for identifying such higher-order response properties.

Unlike most approaches, which try to identify nonlinearities directly by performing higher-order analysis of the stimulus-response function, we used an A/B comparison paradigm using structurally complex signals with known statistical properties. The rationale behind this approach was to use complex broadband stimuli to efficiently activate the auditory neuronal network using known structural components that are physiologically (Schreiner and Calhoun 1994; Calhoun and Schreiner 1998; Kowalski, Depireux and Shamma 1996; Klein *et al.* 2000), psychophysically (Van Veen and Houtgast 1983 1985), and possibly ecologically relevant. Simpler sounds, such as spectro-temporal tone pips (de Charms, Blake, and Merzenich 1998; Theunissen *et al.* 2000), and sum of ripple sounds sound (the ripple noise; Klein *et al.* 2000), lack many of the structural components present in natural (e.g. spectral resonances and FM sweeps) sounds and may not efficiently activate relevant nonlinearities. Thus employing higher-order analysis schemes for such sounds can fail to identify relevant aspects of the stimulus-response function due to insufficient activation (Theunissen *et al.* 2000). Although natural sounds can be used to overcome some of these limitations (Theunissen *et al.* 2000), they are currently limiting from an analysis perspective since they are in general difficult to quantify due to their parametric complexity.

Using the dynamic ripple and ripple noise stimulus, we show that neurons in the inferior colliculus can be classified as phase-locking and non phase-locking neurons.

Phase-locking neurons showed varying degrees of response specificity to the dynamic ripple and the ripple noise sounds. Accordingly these neurons were subdivided into s- and f-neurons. S-neurons are characterized by high firing rates and comparable rates (RSI; Fig. 9) to the moving ripple and ripple noise sounds. Given the similarity of the *STRFs*, both in energy (ASI; Fig. 9) and shape (correlation coefficient; Fig. 6) for these neurons suggests that they behave more or less as expected for a linear system.

Alternately, f-neurons responded most strongly to the dynamic ripple noise and either showed very weak responses or no responses to the ripple noise sound (ASI; Fig. 9). Despite the general low firing rate (mean spike rate=1.4 spikes/sec dynamic ripple and 0.2 spikes/sec for the ripple noise), it was surprising that receptive fields for the dynamic ripple were above chance, extremely precise, and noise free.

In addition to comparing the responses to the dynamic ripple and the ripple noise stimuli, we devised a secondary measure that allowed us to test for feature selectivity by quantifying the variability of the neuronal response to these sounds. The feature selectivity index (*FSI*) quantifies degree of specificity to a particular sound feature given the occurrence of a spike. The *FSI* is derived directly from a covariance-like measure (the similarity index histogram) of the stimulus preceding neuronal responses. Unlike other conventional measures of variability, which quantify the variability of the *output* spike train (e.g., see Rieke *et al.* 1997), the described metric measures the variability at the *input*. Furthermore, the *FSI* measure of selectivity does not require any apriori assumptions about the relevant stimulus or feature, and can be derived directly with the data set used to derive the *STRF*.

The population *FSI* distribution for the dynamic ripple sounds was bimodally

distributed with means of 0.18 and 0.56. More than half of the reported f–neurons (9/17) showed high values of *FSI* ($FSI > 0.4$) whereas only (2/45) of the reported s–neurons had similar high *FSI* values. These findings are consistent with the basic hypothesis that feature selectivity, at least for these instances (9/17 f–neurons), manifests itself directly in the average firing rate, differential response rate of the *STRF*, and the disparity of firing rate between the dynamic ripple and ripple noise sounds. For the described s–neurons, their low *FSI* values (43/45 had values less than 0.4; mean value 0.18) is consistent with their described quasi–linear processing capabilities, since a linear system can respond to sound patterns that *do not* precisely match the system's impulse response, in this case the *STRF*. This is also evident in the similarity index histograms of Figs. 12 and 13 obtained for f– and s–neurons. F–neurons have *SI* histograms that are highly skewed towards +1, indicating that the sound patterns used to construct the *STRF* were largely similar on a trial to trial basis. Alternately, s–neurons had *SIH* that resemble the random control condition, suggesting that the contributing sound patterns are not "identical" to the neuron's *STRF*. This finding can also be interpreted as arising from different ratios of specific versus non–specific spikes for the s– versus f–neurons. However, this is unlikely given that in all instances s–neurons produced strong, statistically significant *STRFs* with differential spikes rates that are comparable in magnitude to the mean firing rate of the neuron (see Figs. 8 A and B).

If a neuron is highly feature selective, it is expected that the overall firing rate be limited by the number of occurrences of the neuron's preferred sound pattern. Presumably the more selective the neuron, the fewer the number of patterns that meet the requirements to invoke a response. Thus, it is not unprecedented that f–neurons

responded with low firing rates, indicating that few patterns meet the requirements to invoke a response. Furthermore the disproportionate response specificity to the dynamic ripple, is explained by the fact that this sound has correlated structural components that are not present in the ripple noise and which may ultimately be more relevant for such neurons.

The motivation for the simple A/B comparison between the dynamic ripple and ripple noise sounds is grounded on the fact that these two sounds have identical long-term statistics (see chapter 2; section 2.15 and 2.18). Consequently, a hypothetical linear neuron would produce identical *STRFs* and similar response rates for both sounds (see section 2.4; Eq. 2.5). Although the long-term autocorrelation function is identical for both sounds, the dynamic ripple has strong local spectro-temporal correlations whereas the ripple noise is locally weakly correlated (chapter 2; Fig. 8 and 9). Thus, at the time-scales that are relevant for neuronal integration (in the order of a few to tens of milliseconds for ICC neurons), the dynamic ripple sound has maximal power concentrated at a particular set of ripple parameters. The fact that f-neurons are efficiently activated under these conditions is suggestive of a threshold like nonlinearity. This is supported by intracellular studies in the visual cortex which show that supra-threshold receptive fields are more precise and spatially localized in comparison to sub-threshold receptive fields (Moore and Nelson 1998; Bringuier *et al.* 1999).

It is important to note that the described nonlinearities can not be identified directly with pure tones and other simple narrow-band stimuli. In all of these experiments, all neurons were initially identified audio-visually with pure tones and clicks. Despite the selectivity to the dynamic ripple, the described f-neurons also

responded to pure tone stimuli. This seems at odds with the described "feature selectivity" of these neurons. However, one must take into account that the dynamic ripple and ripple noise sounds represent a distinct operating condition for the auditory neuronal network (Miller and Schreiner 2000). It is very likely that the observed response selectivity is an effect induced by the complex broadband excitation pattern of the dynamic ripple and ripple noise stimuli (in comparison to the focal excitation pattern of a pure tone). This is supported by studies in the auditory cortex with the dynamic ripple sound which show an induced modification of the operational and dynamic state of the auditory neuronal network (Miller *et al.* 2000; Miller and Schreiner 2000).

In the visual system, the fundamental operations performed by visual cortex neurons were not revealed until the pivotal discovery of simple and complex cells by Hubel and Wiesel (1962). Since these results were largely attributed to the fact that complex stimuli (bars and edges) were used (as opposed to simple spots of light) it is no wonder why the visual field has taken the notion of using complex visual stimuli quite seriously. It is likely that similar steps will be necessary for understanding the functional rules which the auditory system uses for natural sounds processing. Accordingly, the observed findings have direct implications for natural sound processing since the broadband excitation patterns of the ripple sounds likely represents a more realistic scenario of auditory processing – providing joint activation of excitatory and inhibitory neuronal inputs with a complex spectro-temporal activation pattern. Further studies and more direct comparisons with pure tone tuning curves and AM stimuli are necessary to elucidate on these points. Such a study may likewise be necessary to disambiguate the projection patterns to the ICC from

subcollicular neurons (Ramachandran, Davis, and May 1999), and to relate the known neuronal types with the described spectro-temporal processing abilities of these neurons .

Unlike the described populations of phase-locking neurons, a significant number of neurons showed weak or no phase-locking to the stimulus spectro-temporal envelope. Despite the general lack of time-locking to the stimulus spectro-temporal envelope, the described neurons were shown to respond selectively to the stimulus spectral and temporal parameters.

This represents an interesting functional analogy between the visual and auditory systems since the described response properties are analogous to those of complex cells in the primary visual cortex which have even order nonlinearities, do not linearly time-lock to spatio-temporal visual patterns and, consequently, do not produce linear spatio-temporal receptive fields (Emerson *et al.* 1987; Szulborski and Palmer 1990). Whether the described neuronal responses arise from even-order nonlinearities and similar projecting patterns of input as for visual complex cells (Alonso and Martinez 1998; DeAngelis 1999), still needs to be determined. Preliminary analysis (not shown), however, suggests that these neurons may indeed have interleaved on and off response subfields (temporally oriented), analogous to visual complex cells in layer II/III (Szulborski and Palmer 1990; DeAngelis 1999). Because of this analogy, we refer to these as non phase-locking or c-neurons.

Sinusoidal AM studies in the inferior colliculus indicate that sinusoidal AM tuning characteristics may be best defined for some neurons by a rate code as opposed to a synchrony code (Rose and Capranica 1985; Epping and Eggermon 1986; Langner and Schreiner 1988; Schulze and Langner 1997). The exact function of this basic

transformation is not clear although it may be necessary because of the limited capacity of the auditory cortex to follow fast temporal modulations beyond about 50 Hz. This may in turn be attributed partly to intrinsic properties of the cortical cell membrane (Eggermont 1999) and functional transformations of the corticothalamic network (Creutzfeldt, Hellweg, and Schreiner 1980). Consequently, it is possible that these temporal encoding limitations of the auditory cortex give rise to a spatially distributed rate code for temporal modulations (Schreiner and Langner 1988). Such a representation is already partly present at the level of the central nucleus of the inferior colliculus (Schreiner and Langner 1998; Langner and Schreiner 1998).

A secondary hypothesis for the observed segregation of phase-locked (f-neurons and s-neurons) and non-phase-locked encoding (c-neurons), is provided by the fact that the described c-neurons respond to a distinct range of spectro-temporal parameters (see Fig. 22) in comparison to the phase-locked neurons. These differences may be necessary for encoding various ranges of perceptually relevant temporal modulations. For example, the following rate abilities of the auditory cortex are precisely overlapped with the range of temporal modulation rates that give rise to the perception of slow temporal rhythms (Royer and Garner 1966 1970). It has been suggested that AI neurons can encode information about fast and slow temporal modulations, such as those that give rise to the percept of pitch and rhythms respectively, using a spatially distributed rate-place code (Schulze and Langner 1997). Thus, it is possible that auditory cortex neurons encode slow temporal modulations directly using a temporal code, whereas they simultaneously encodes fast temporal modulations, using a rate code. Given the encoded ranges for phase-locked and non-phase-locked c-neurons, it is likewise possible that similar

segregation occurs at the level of the ICC. These neuronal encoding differences may have direct implications for how various perceptual quantities are encoded.

3.18 Conclusion

This study contributes a general approach that was able to reveal classes of neurons with linear and nonlinear response characteristics along with the underlying spectro-temporal acoustic structure that activates these nonlinear responses. The comparison of neuronal responses to structured and unstructured noise stimuli (with identical low-order statistics) combined with a parametric characterization of the neuronal responses allowed us to identify several nonlinear neuronal response classes which conventional reverse correlation stimuli, such as unstructured noises (spectro-temporal m-sequences and white noise), and *STRF* mapping techniques alone can not reveal.

Because of the distinct response properties to structured and unstructured sounds, these physiologically defined neural classes can, in principle, relay information about different types of natural sounds to higher auditory areas in codes that are either sparse, dense, synchronized, and/or desynchronized. The possible stream of information arising from f-neurons is temporally (low spike rate) sparse and highly synchronized. Consequently, it is well suited for detecting temporal transitions, biologically relevant sounds, and for feature segmentation. In contrast, s-neurons can provide a dense and continuous flow of time-locked activity which is ideal for general processing. C-neurons can alternately provide a dense and desynchronized rate code for spectral and temporal stimulus features that are beyond the range of time-locked neural activity at this stage of

processing. The three distinct coding strategies employed in parallel at this level of processing in addition to differences in the preferred spectro–temporal stimulus parameters may be prerequisite for processing higher–order stimulus features and establishing distributed spatial representations (Schreiner 1998; Rauschecker 1998). Together, such distinct nonlinear processing modes can offer computational advantage for acoustic feature decomposition, for signal segregation, and for tasks which are otherwise intractable with simple linear filtering strategies.

The fact that these functionally distinct neural types occur in a subcortical station is surprising, in light of the general views of auditory and visual processing where such higher–order functions are either reserved for, or most prominent in, cortical stations. The visual system, however, is anatomically different than the auditory pathway and has no integrative structure analogous to the ICC. The ICC receives convergent projections from more than 20 neural types in at least 10 ascending brainstem nuclei as well as descending projections from the auditory cortex (Fay and Popper 1992). Consequently, the ICC is ideally situated to play a key functional role of neuronal integration and acoustic transformation as reflected in the identified response types. The notion that the central auditory system is adapted for processing structural features of natural sounds is well supported by behaviorally guided studies in the bat and songbird auditory systems (Suga and Jen 1976; Suga, O’neil, Manabe 1978; Margoliash 1983; Margoliash and Fortune 1992; Olsen and Suga 1993a 1993b; Dope 1997) although these methods are not easily transferred to species that lack obvious behavioral specializations. Our findings demonstrate that a general and systematic analysis of central auditory function in a purportedly non–specialized mammal is possible, revealing specializations for sound

3.19 References

- J.M. Alonso and L.M. Martinez. Functional connectivity between simple cells and complex cells in cat striate cortex. *Nature Neurosc.* **1** (5), 395–403 (1998).
- A. M. H. J. Aersten, J. H. J. Olders, and P. I. M. Johannesma. Spectro–temporal receptive fields in auditory neurons in the grass frog: analysis of the stimulus–event relation for tonal stimulus. *Biological Cybernetics* **38**, 235–248 (1980).
- A. M. H. J. Aersten, J.H.J. Olders, and P. I. M. Johannesma. Spectro–temporal receptive fields in auditory neurons in the grass frog: analysis of the stimulus–event relation for natural stimulus. *Biological Cybernetics* **30**, 195–209 (1981).
- H. Attias, and C.E. Schreiner. Temporal Low Order Statistics of Natural Sounds. *Advances in Neural Information Processing Systems* **10**, 27–33 (1998a).
- H. Attias, and C.E. Schreiner. Coding of Naturalistic Stimuli by Auditory Neurons. *Advances in Neural Information Processing Systems* **10**, 103–109 (1998b).
- A. Anzai, I. Ohzawa, R.D. Freeman. Neural mechanisms for processing binocular information: I. Simple cells. *J. Neurophysiol.* **82** (2), 891–908 (1999).
- H.B. Barlow. Summation and inhibition in the frog retina. *J. Physiol. (London)* **119**, 69–78 (1953).
- V. Binguier, F. Chavane, L. Glaesr, Y. Frégnac. Horizontal propagation of visual activity in the synaptic integration field of are 17 neurons. *Science* **283**, 695–699 (1999).
- B.M. Calhoun and C.E. Schreiner. Spectral Envelope Coding in Cat Primary Auditory Cortex: Linear and non–linear effects of stimulus characteristics. *European J. of Neurosci.* **10**, 926–940 (1998).
- L.H. Carney and C.D. Geisler. A temporal analysis of auditory nerve fibers responses to spoken stop consonant–vowel syllables. *J. Acoust. Soc. Amer.* **79**, 1896–1914 (1986).
- J.H. Casseday, D. Ehrlich, and E. Covey. Neural Tuning for Sound Duration. Role of Inhibitory Mechanisms in the Inferior Colliculus. *Science* **264**, 847–850 (1994).
- C.K. Chui. *An Introduction to Wavelets*. Academic Press, San Diego, 1992.
- O. Creutzfeldt, F.C. Hellweg, and C.E. Schreiner. Thalamocortical

transformation of responses to complex auditory stimuli. *Experimental Brain Research* **39**, 87–104 (1980).

G.C. Deangelis, I. Ohzawa, R.D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex: II. Linearity of temporal and spatial summation. *J. Neurophysiol.* **69** (4), 1118–1135 (1993).

G.C. DeAngelis, G.M. Ghose, I. Ohzawa, R.D. Freeman. Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *J. Neurosci.* **19** (10), 4046–4064 (1999).

B. Delgutte and N.Y.S. Kiang. Speech coding in the auditory nerve. I. Vowel-like sounds. *J. Acoust. Soc. Am.* **75** (3), 867–879 (1984).

A.J. Doupe, Song- and Order-Selective Neurons in the Songbird Anterior Forbrain and their Emergence during Vocal Development. *J. of Neurosci.* **17** (3), 1147–1167 (1997).

R.C. Emerson, M.C. Citron, W.J. Vaughn, and S.A. Klein. Nonlinear directionally selective subunits in complex cells of cat striate cortex. *J. Neurophysiol.*, **58**, 33–65 (1987).

W.J.M Epping and J.J. Eggermont. Sensitivity of neurons in the auditory midbrain of the grassfrog to temporal characteristics of sounds. II. Stimulation with amplitude modulated sound. *Hear Reas* **24**, 55–72 (1986).

J.J. Eggermont. The magnitude and phase of temporal modulation transfer functions in cat auditory cortex. *J. Neurosci.* **19** (7), 2780–8 (1999).

J.J. Eggermont. Wiener and Volterra Analyses applied to the Auditory System. *Hear. Reas.* **66**, 177–201 (1993).

R.R. Fay and A.N. Popper (Eds). *The Mamalian Auditory Pathway: Neuroanatomy*. Springer Verlag, New York, (1992).

C.D. Geisler and T. Gamble. Responses of "high-spontaneous" auditory-nerve fibers to constant-vowel syllables in noise. *J. Acoust. Soc. Amer.* **85**, 1639–1652 (1989).

S.V. Girman, Y. Sauve, and R.D. Lund. Receptive field properties of single neurons in rat primary visual cortex. *J. Neurophysiol.* **82** (1), 301–311, 1999.

I. Glass and Z. Wollberg. Auditory cortex response to sequences of normal and reversed squirrel monkey vocalizations. *Brain. Behav. Evol.* **22**, 13–21 (1983).

J.M. Goldberg and P.B. Brown. Response of binaural neurons of dog superior olivary complex: an anatomical and electrophysiological study. *J. Neurophysiol.* **31**, 639–656 (1969).

D.H. Hubel and T.N. Wiesel. Receptive Fields, Binaural Interaction and Functional Architecture in the Cat's Visual Cortex. *J. Physiol. Lond.* **160**, 106–154 (1962).

D.R.F. Irvine and G. Gao. Binaural interaction in high-frequency neurons in the inferior colliculus of the cat. Effects of variations in sound pressure level on sensitivity to interaural intensity differences. *J. Neurophysiol.* **63**, 570–591 (1990).

J.P. Jones and L.A. Palmer. An evaluation of the two dimensional Gabor filter model of simple receptive fields in the cat striate cortex. *J. Neurophysiol.* **58**, 1233–1258 (1987a).

J.P. Jones and L.A. Palmer. The two dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1187–1211 (1987b).

D.J. Klein, D.A. Depireux, J.Z. Simon, and S.A. Shamma. Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. *J. Comp. Neurosci.* **9**, 85–111 (2000).

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

B.S. Krishna and M.N. Semple. Auditory Temporal Processing: Responses to Sinusoidally Amplitude-Modulated Tones in the Inferior Colliculus. *J. Neurophysiol.* **84**, 255–273 (2000).

M.S. Lewicki. Bayesian modeling and classification of neural signals. *Neural computation* **6**, 1005–1029 (1994).

G. Langner and C.E. Schreiner. Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J. Neurophysiol* **60**, 1799–1822 (1988).

D. Margoliash. Acoustic parameters underlying the response of song-specific neurons in the white-crowned sparrow. *J. Neurosci.* **3**, 1039–1057 (1983).

D. Margoliash and E.S. Fortune. Temporal and harmonic combination-sensitive neurons in the zebra finch's Hvc. *J. Neurosci.* **12**, 4309–4326 (1992).

P.Z. Marmarelis and K.I. Naka. Identification of Multi-Input Biological

- Systems. *IEEE Transactions on Biomedical Engineering* **21** (2), 88–101 (1974).
- L.M. Miller, M.A. Escabí, C.E. Schreiner. Synchronous oscillation in the thalamocortical system and the effects of naturalistic ripple stimuli. In: *Computational models of Auditory Function*. S. Greenberg, M. Slaney (eds.), IOS Press (2000).
- L.M. Miller and C.E. Schreiner. Stimulus-based state control in the thalamocortical system. *J. Neurosci.* **20**, 7011–7016 (2000).
- A.R. Møller and A. Rees. Dynamic properties of the responses of single neurons in the inferior colliculus of the rat. *Hear. Res.* **24**, 203–215 (1986).
- C.I. Moore and S.B. Nelson. Spatio-temporal subthreshold receptive fields in the vibrissa representation of rat primary somatosensory cortex. *J. Neurophysiol.* **80** (6), 2882–2892 (1998).
- I. Nelken, P.J. Kim, E.D. Young. Linear and nonlinear spectral integration in type IV neurons in the dorsal cochlear nucleus. II. Predicting responses with the use of nonlinear models. *J. Neurophysiol* **78**, 800–811 (1997).
- I. Nelken, Y. Rotman, and O.B. Yosef. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* **37**, pp. 154–157 (1999).
- I. Nelken and O.B. Yosef. Processing of complex sounds in cat primary auditory cortex. *Proceedings of the NATO Advanced Study Institute on Computational Hearing*. Eds: S. Greenberg and M. Slaney (1998).
- K.K. Ohlemiller, J.S. Kanwal, N. Suga. Facilitative responses to species-specific calls in the cortical FM-FM neurons of the mustache bat. *NuroReport* **7**, 1749–55 (1996).
- J.F. Olsen and N. Suga. Combination sensitive neurons in the medial geniculate body of the mustache bat: encoding of target range information. *J. Neurophysiol.* **65**, 1254–1274 (1991a).
- J.F. Olsen and N. Suga. Combination sensitive neurons in the medial geniculate body of the mustache bat: encoding of relative velocity information. *J. Neurophysiol.* **65**, 1275–1296 (1991b).
- R. Plomp. Pitch of complex tones. *J. Acoust. Soc. Am.* **41**, 1526–1533 (1967).
- R. Plomp. Timbre as a multidimensional attribute of complex tones. *Frequency analysis and Periodicity Detection in Hearing*. Edited by R. Plomp and G.F. Smoorenburg, Sijthoff Linden (1970).

R. Plomp. The Role of Modulations in Hearing. *Hearing Physiological Bases and Psychophysics*. Edited by R. Klinke and R. Hartmann, Springer Verlag, New York, 270–276 (1983).

D. Ploog. Neurobiology of primate audio–vocal behavior. *Brain Research Reviews* **3**, 35–61 (1981).

L.C.W. Pols, L.J.T. Kamp, and R. Plomp. Perceptual and physical space of vowel sounds. *J. Acoust. Soc. Am.*, **46**, 458–467 (1969).

R. Ramachandran, K.A. Davis, and B.J. May. Single–Unit Responses in the Inferior Colliculus of Decerebrate Cats I. Classification Based on Frequency Response Maps. *J. Neurophysiol.* **82**, 152–163 (1999).

J.P. Rauschecker. Cortical processing of complex sounds. *Curr. Opin. Neurobiol.* **8**, 516–521 (1998).

A. Rees and A.R. Møller. Response of neurons in the inferior colliculus of the rat to AM and FM tones. *Hear. Res.* **10**, 301–330 (1983).

A. Rees and A.R. Møller. Stimulus properties influencing the response of inferior collicular neurons to amplitude–modulated sounds. *Hear. Res.* **27**, 129–143 (1987).

A. Rees and A.R. Palmer. Rate–intensity functions and their modification by broadband noise for neurons in the guinea pig inferior colliculus. *J. Acoust. Soc. Am.* **83**, 1488–1498 (1988).

A. Rees and A.R. Palmer. Neuronal response to amplitude–modulated and pure tone stimuli in the guinea pig inferior colliculus, and their modification by broadband noise. *J. Acoust. Soc. Am.* **85**, 1978–1994 (1989).

D.S. Reich, F. Mechler, K.P. Purpura, J.D. Victor. Interspike intervals, receptive fields, and information encoding in primary visual cortex. *J. Neurosci.* **20** (5), 1964–1974 (2000).

F. Rieke, D. Warland, R. de Ruyter van Steveninck, W. Bialek, *Spikes: Exploring the Neural Code*, The MIT Press, 1997.

R.M. Roark and M.A. Escabí. B–spline design of maximally flat and prolate spheroidal–type FIR filters. *IEEE Trans. on Signal Processing* **47** (3), 701–716 (1999).

G.J. Rose and R.R. Capranica. Sensitivity to amplitude modulated sounds in the

anuran auditory nervous system. *J. Neurophysiol.* **53**, 446–465 (1985).

F.L. Royer and W.R. Garner. Perceptual organization of nine–element auditory temporal patterns. *Percept. Psychophys.* **7**, 115–120 (1970).

F.L. Royer and W.R. Garner. Response uncertainty and perceptual difficulty of auditory temporal patterns. *Percept. Psychophys.* **1**, 41–47 (1966).

R.G. Szulborski and L.A. Palmer. The Two Dimensional spatial structure of nonlinear subunits in the receptive fields of complex cells. *Vision Research* **30** (2), 249–254 (1990).

M.B. Sachs and E.D. Young. Encoding of steady–state vowels in the auditory nerve: Representation in terms of discharge rate. *J. Acoust. Soc. Am.* **66** (2), 470–479 (1979).

S.A. Shamma. Speech processing in the auditory system I: The representation of speech sounds in the response of the auditory nerve. *J. Acoust. Soc. Am.* **78** (5), 1612–1621 (1985).

C.E. Schreiner. Spatial Distribution of Responses to Simple and Complex Sounds in the Primary Auditory Cortex. *Audiology Neuro–Otology* **3**, 104–122 (1998).

C.E. Schreiner, B.M. Calhoun. Spectral envelope coding in the cat primary auditory cortex: Properties of ripple transfer function. *Auditory Neuroscience*, vol. 1, pp. 39–61 (1994).

C.E. Schreiner and G. Langner. Periodicity coding in the inferior colliculus of the cat. II. Topographic Organization. *J. Neurophysiol* **60**, 1799–1822 (1988).

C.E. Schreiner, J.V. Urbas. and S. Mehrgardt. Temporal resolution of amplitude modulation and complex signals in the auditory cortex of the cat. *Hearing–Physiological bases and Psychophysics*, R. Klinke and R. Hartmann (eds.). New York, Springer–Verlag, 169–175 (1983).

H. Schulze and G. Langner. Periodicity coding in the primary auditory cortex of the Mongolian gerbil (*Meriones unguiculatus*): Two different coding strategies for pitch and rhythm?, *J. Comparative Neurophysiol.* **181** (6), 651–663 (1997).

N. Suga and P.H. Jen. Disproportionate tonotopic representation for processing CF–FM sonar signals in the mustache bat auditory cortex. *Science* **194**, 542 (1976).

N. Suga, W.E. O’neil and T. Manabe. Cortical neurons sensitive to particular combinations of information bearing elements of bio–sonar signals in the

mustache bat. *Science* **200**, 778–781 (1978).

R.G. Szulborski and L.A. Palmer. The two dimensional spatial structure of nonlinear subunits in the receptive fields of complex cells. *Vision Research* **30** (2), 249–254 (1990).

F.E. Theunissen and A.J. Doupe. Temporal and spectral sensitivity of complex auditory neurons in the nucleus HVC of male zebra finches. *J. Neuroscience* **18** (10), 3786–3802 (1998).

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

T.M. Van Veen and T. Houtgast. Spectral Sharpness and Vowel Dissimilarity. *J. Acoust. Soc. Am.* **77** (2), 628–634 (1985).

T.M. Van Veen and T. Houtgast. On the Perception of the Spectral Envelope. *Hearing Physiological Bases and Psychophysics* (R. Klinke and R. Hartmann Eds.), Springer Verlag, New York, 277–281 (1983).

H. Versnel and S.A. Shamma. Spectral–ripple representation of steady–state vowels in primary auditory cortex. *J. Acoust. Soc. Amer.* **103** (5), 2502–2514 (1998).

J.D. Victor and K.P. Purpura. Spatial phase and the temporal structure of the response to gratings in V1. *J. Neurophysiol.* **80** (2), 554–571 (1998).

R.V. Voss and J. Clarke. 1/f noise in music and speech. *Nature.* **258**, 317–318 (1975).

X.Wang, M. Merzenich, R. Beitel, and C.E. Schreiner. Representation of Species–Specific Vocalization in the Primary Auditory Cortex of the Common Marmoset: Temporal and Spectral Characteristics. *J. Neurophysiol.* **24** (6), 2685–2706 (1995).

P. Winter and H.H. Funkenstein. The effects of species–specific vocalizations on the discharge of auditory cortical cells in the awake squirrel monkey (*Saimiri sciureus*)., *Exp. Brain. Res.* **18**, 489–504 (1973).

Y. Yeshurun, Z. Wollberg, N. Dyn, N. Allon. Identification of MGB cells by Volterra kernels. I Prediction of responses to species specific vocalizations. *Biol. Cybern.* **51**, 383–390 (1985).

E.D. Young. What’s the Best Sound. *Science* **280**, 1402–1403 (1998).

Natural Contrast and Intensity Processing

Abstract

Human perception is remarkable both for the range of stimulus intensities that we can perceive sounds and for our ability to detect small intensity difference. Humans, for example, perceive sounds of absolute intensities which span a range of more than 110 dB. Yet the human ear is extremely sensitive to intensity differences and can detect changes of as little as 0.5 dB (Miller 1947; Harris 1963; Jesteadt *et al.* 1977; Florentine *et al.* 1987) throughout most of this range of absolute intensities. It is well accepted that the auditory sense utilizes its large operating range for loudness coding (e.g. Evans and Palmer 1980; Ehret and Merzenich 1988; Viemeister 1988; Eggermont 1989). Human psychophysics also indicates that the auditory system can exploit large intensity fluctuations related to contrast, in time and along the sensory epithelium, thereby improving intelligibility, discriminability, and detection thresholds of speech (Van Veen and Houtgast 1983 1985). The neurophysiological correlates of dynamic range, intensity discrimination, and contrast are poorly understood. Here we demonstrate that, on the average, natural sounds have logarithmically distributed spectro-temporal amplitude fluctuations. Single unit neuronal recordings in the cat auditory midbrain demonstrate that auditory neurons efficiently utilize the spectro-temporal energy distributions observed in natural sounds. When exposed to dynamic broad-band stimuli with logarithmic intensity gradations, midbrain neurons show contrast tuning and improved spectro-temporal coding at time scales comparable to the neurons' receptive field. This finding suggests that the operating range of auditory neurons is physically matched to the statistical structure inherent in natural sounds. Such a neural adaptations makes use of structural regularity of natural sounds and, likely, underlies human perceptual abilities.

4.1 Introduction

A central hypothesis of sensory coding asserts that sensory systems efficiently make use of statistical structure inherent in naturally occurring signals. The possibility that sensory systems are adapted for encoding natural signals has been a topic of discussion since the early works of Barlow (Barlow 1953; Barlow 1961). Recent works have revealed that naturally occurring visual (Ruderman and Bialek 1994; Dong and Atick 1995; Ruderman 1997) and acoustic signals (Voss and Clarke 1975; Voss and Clarke 1978; Attias and Schreiner 1998a; Nelken *et al.*; Escabí 2000 Chapt. 1) show robust statistical properties such as scale invariant contrast statistics and $1/f$ power spectrum. Although numerous works have looked at these statistical characteristics of natural signals, only a few studies have addressed how such statistics can be used for efficient sensory coding (Rieke *et al.* 1995; Dan *et al.* 1996; Attias and Schreiner 1998b; Nelken *et al.*; Stanley *et al.* 1999). Direct application of information theoretic approaches has revealed that sensory neurons respond most efficiently to sensory signals with natural statistics, although the exact mechanisms responsible for these observations have not been studied in detail. Here we address issues of contrast and intensity coding in the central auditory system using signals that emulate the statistical characteristics observed in natural sounds.

Contrast is a fundamental component of all sensory signals which the brain uses to encode stimulus information. By definition contrast is the overall range of intensity gradations which coexist along time and along the sensory epithelium normalized by the mean stimulus intensity or luminance. In vision, contrast corresponds to the range spatio-temporal gradations along retinotopic space whereas in audition we can consider

the spectral–temporal intensity gradations which excite the basilar membrane. In natural vision and hearing, our senses are exposed to sensory stimuli which span many orders of magnitude in their mean and instantaneous intensities. Since energy gradations, either along time or the sensory epithelium, represent much of the information–bearing components of sensory signals, it is expected that the auditory system utilizes structure present in natural signals for efficient sound encoding. Despite these general facts, auditory and visual scientist generally use stimuli with linear amplitude gradations and of limited dynamic range.

In audition, the ability of the auditory system to encode amplitude differences has been studied almost exclusively in the context of intensity discrimination and loudness coding (Palmer and Evans 1979; Evans and Palmer 1980; Ehret and Merzenich 1988; Viemeister 1988; Eggermont 1989) for pure tones and white noise. Little is know as to how the spectral, temporal, and intensity dimensions of a complex sounds are jointly represented by auditory neurons. Auditory neurons, for example, often show a monotonic input–output relationship where the mean response rate increases linearly as a function of sound pressure level (SPL) (Palmer and Evans 1979; Evans and Palmer 1980) over a range of stimulus intensities. Such neurons typically show rate level functions with 30–60 dB operating range (Palmer and Evans 1979; Evans and Palmer 1980). Neurons at central auditory stations can additionally show tuned input–output response curves (Eggermont 1989; Ehret and Merzenich 1988). All of these studies use static stimuli (i.e. noise and pure tones) which lack many of the relevant acoustic features common to natural sounds (i.e. spectro–temporal fluctuations). Such studies maintain that input–output curves of auditory neurons largely convey intensity information. According to this

general model, temporal modulations and spectral gradations are encoded independently of level.

Using a *naturalistic* ripple noise stimuli that emulate statistical characteristics of natural sounds (see chapter 1) along with spectro-temporal reverse correlation methods we show that inferior colliculus neurons utilize the large dynamic range of natural sounds for efficient sound encoding. By comparing responses to ripple noise stimuli of different dynamic range we observe contrast tuning, increased spike rates, and reduced variability of the spiking output. The auditory system can therefore potentially use contrast information as a secondary acoustic cue to encode stimulus information. Additionally it is shown that auditory neurons exploit the dynamic range and the relative spectro-temporal energy distribution observed in natural stimuli to accurately extract spectral and temporal information at time scales comparable to the neuron's spectro-temporal receptive field. These findings suggest that neurons in the central auditory system are matched, by virtue of their operating range, to analyze acoustic stimuli with similar logarithmically distributed amplitude fluctuations.

4.2 Contrast Statistics of Natural Sounds

Visual contrast is defined as the percent deviation relative to the mean intensity of a spatial sinusoid grating. Mathematically it is expressed as $C = (I_{\max} - I_{\min}) / (I_{\max} + I_{\min})$ where I_{\max} and I_{\min} correspond to the maximum and minimum stimulus intensities (Albrecht 1995; Nordmann, Freeman, and Casanova 1992; Troy *et al.* 1998). In the auditory literature the analogous quantity is the modulation depth or modulation index, $\beta = (A_{\max} - A_{\min}) / A_{\max}$ where A_{\max} and A_{\min} correspond to the maximum and minimum

stimulus amplitudes. Such a description suffices for the case of sinusoidal, square wave, and other simple stimulus gradations since these waveforms are fully specified by their minimum and maximum intensities. For natural signals, where the amplitude gradations can cover several orders of magnitude, such descriptions fail to fully characterize amplitude fluctuations since they only take into account the minimum and maximum envelope intensities. They do not tell us anything about intermediate values and higher-order amplitude statistics of the modulation signal. To overcome this we adopt a more general definition of contrast to denote the probability distribution of the relative amplitude gradations.

A large ensemble of natural sounds was analyzed which included human speech (excerpts from Hamlet), music (pop and classical), environmental sounds (wind, rain, thunder, etc.), animal vocalization (primate, bird, cat, crickets etc.) and mixtures of the latter two. All sounds were obtained from digitally recorded or remastered media (see methods). These sounds were taken as representative examples of the vast acoustic biotope (Smolders *et al.* 1979) which mammals and humans are typically exposed to. For comparison, white noise was included in this analysis as a control. For all sounds the relative spectro-temporal envelopes, $\bar{S}(t, f_k)$ and $\bar{S}_{dB}(t, f_k) = 20 \log_{10}(\bar{S}_{dB}(t, f_k))$, were computed and the corresponding envelope contrast distributions, $C = p(\bar{S})$ and $C_{dB} = p(\bar{S}_{dB})$, were estimated for thousands of sound segments (see methods).

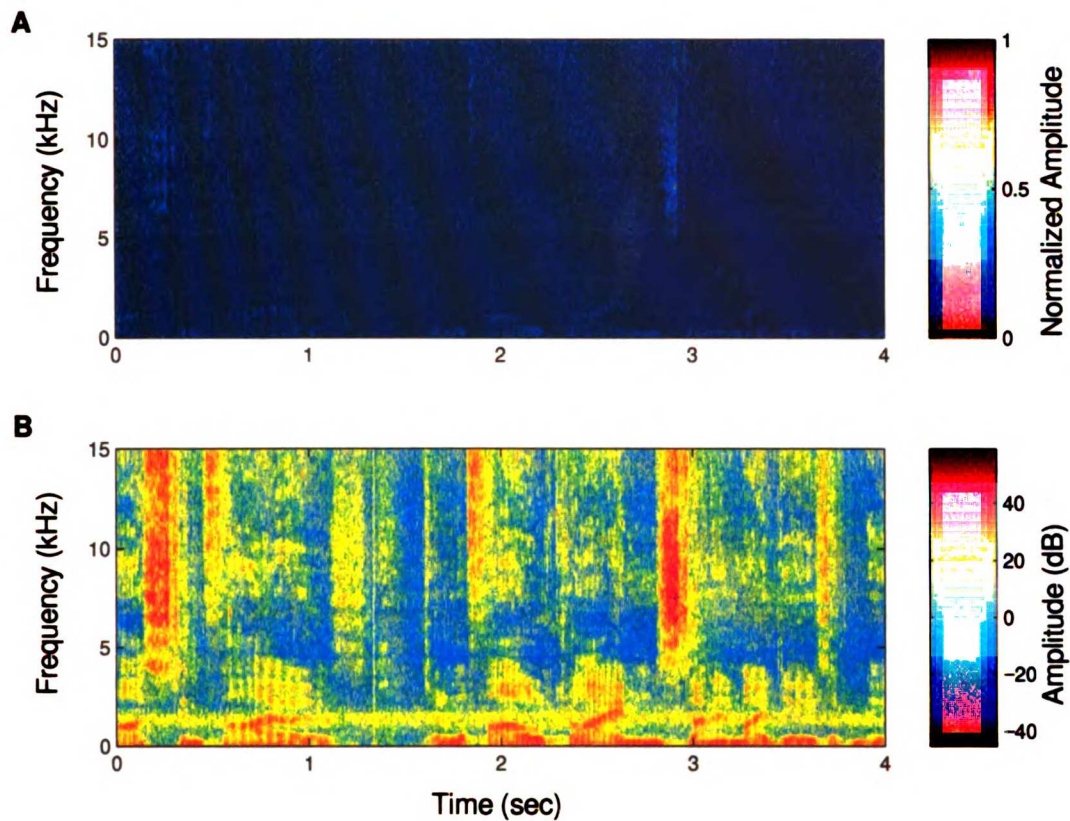


Figure 1: Detrended spectrographic envelope for a short speech segment. A brief speech segment is shown using a linear amplitude (A) and a decibel amplitude convention (B). The linear amplitude spectro-temporal envelope shows little detail and most of the signal values are concentrated near zero. The decibel spectro-temporal envelope has more detail and has amplitude fluctuations that span a large dynamic range of more than 50 dB.

Fig. 1 shows the decibel and linear amplitude spectro-temporal envelopes for a human speech segment. The linear amplitude spectro-temporal envelope (Fig. 1 A) shows little detail and largely consists of amplitude values near zero (blue). The measured linear modulation depth for this speech segment is exceptionally high (99.9994 %), whereas the measure standard deviation, σ , is relatively small (0.019 normalized

amplitude units for an amplitude range that spans 0 to 1). Together these two descriptors provide a conflicting and misleading description of the envelope fluctuations. The large modulation index suggests that the sound components for this segment span a large range of the 0 to 1 linear amplitude dimension, whereas the small standard deviation suggest that it only covers a small portion of this linear amplitude space. By comparison, the decibel amplitude spectro-temporal envelope (Fig. 1 B) shows significant more structure. A close inspection of the logarithmic decibel envelope,

$\bar{S}_{dB}(t, f_k) = 20 \log_{10}(\bar{S}(t, f_k))$, reveals that the speech signal has spectral and temporal amplitude fluctuations that span several orders of magnitude (approximately 50 dB, Fig. 1 B, see colorscale).

To quantify these observations, we computed the linear and decibel contrast distributions for all sounds by collapsing all pixel values of the linear and decibel spectro-temporal envelopes, respectively, into a probability histogram. These are shown collectively for all sound ensembles in Figs. 2 and 3. The linear amplitude distribution was obtained by normalizing the spectro-temporal envelope so that it has a maximum value of unity, $\bar{S}_{Lin}(t, f) = \bar{S}(t, f) / \max(\bar{S}(t, f))$, therefore obeying the general convention used to define a modulation signal (Cohen 1995). For all natural sounds the linearly defined envelope has a skewed amplitude distribution such that loud (near unity) sound segments are sparse whereas soft segments (near zero) are much more common (Fig. 2). In contrast, white noise (Fig. 2 F) has a linear amplitude distribution which is broadly distributed and partially symmetric. Upon performing a logarithmic decibel transformation of the envelope to construct the decibel contrast distributions,

$C_{dB} = p(\bar{S}_{dB})$, the relative amplitude gradations of natural sounds are roughly symmetric, have an average standard deviation of 10.9 dB, and span an overall range of more than 25 dB (Fig. 3) for the natural sounds ensembles. Traditional definitions of contrast, such as the modulation depth or the envelope standard deviation, fail to characterize such higher-order statistics associated with the shape and the overall range of the envelope gradations.

The transformed logarithmic decibel amplitude (\bar{S}_{dB}) magnifies the soft and moderately loud sound segments relative to the very loud sounds. Thus one can discern the fine detail in the amplitude distribution over several orders of magnitude. This descriptor is perceptually motivated since the perception of loudness and intensity discrimination thresholds are ordered on a decibel space (Miller 1947; Stevens 1957; Harris 1963; Stevens 1972; Jesteadt, Wier, and Green 1977; Viemeister and Bacon 1988). For all sounds the distribution of logarithmic-contrast is broadly distributed. To quantify the range of relative amplitudes we measured the average spread of the distribution, σ_{dB} . With the exception of the background sounds, all natural sounds had relatively large standard deviation values: 11.0 dB for speech, 13.3 dB for vocalizations, 7.4 dB for background sounds, 11.2 dB for pop-music and 11.8 dB for the classical music ensemble. By comparison, the white noise control ensemble has a small standard deviation of only 5.6 dB. The overall range of values as determined by the 95th percentile range also covered a large extent of the decibel amplitude space: a total range of 53 dB for speech, 52 dB for vocalizations, 44 dB for background sounds, 46 dB for pop-music and 54 dB for the classical music ensemble.

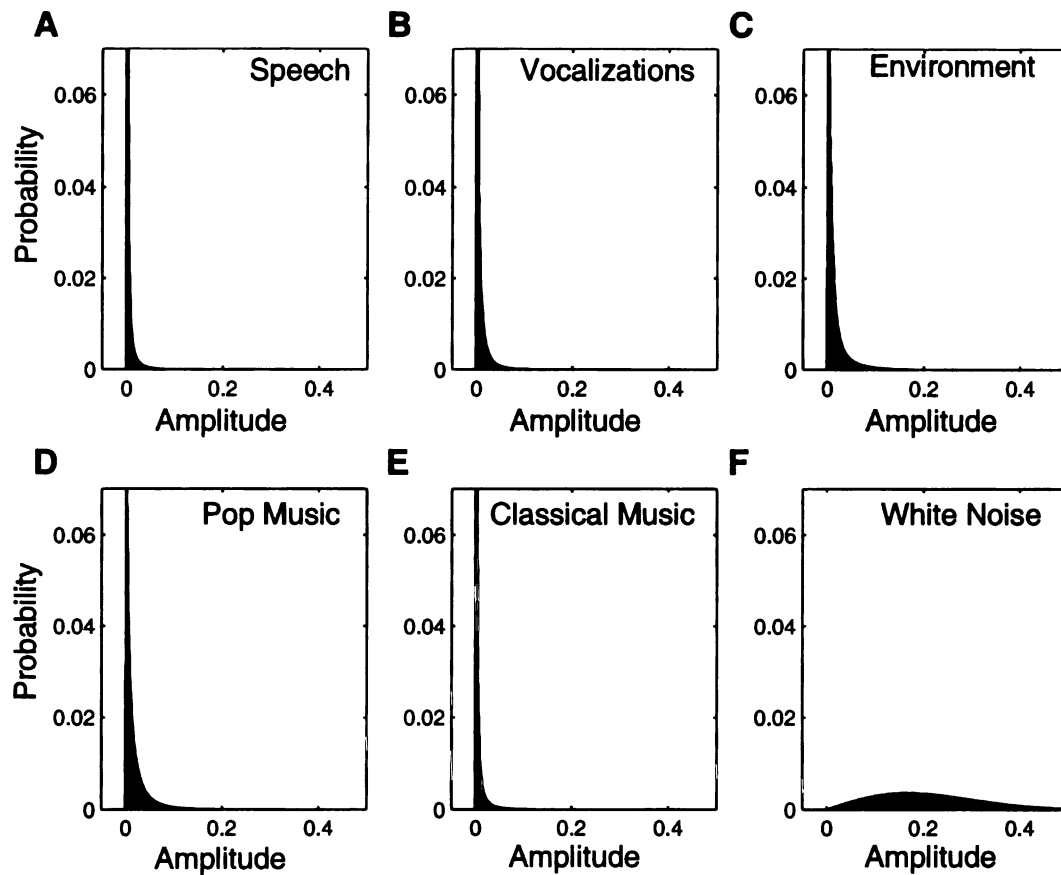


Figure 2: Linear contrast statistics for natural sound ensembles and white noise. The linear amplitude distribution, $p(\bar{S})$, for speech (A), animal vocalizations (B) (both primate and nonprimate sources), background sounds (C) (e.g. wind, running water, etc.), pop-music (D) classical music (E) and white noise (F). All sounds are normalized so that the spectro-temporal envelope has a maximum amplitude of unity. The linear amplitude distribution of all natural sounds is skewed such that soft sound segments (near zero) occur with high probability. By comparison, white noise has a broad distribution (F).

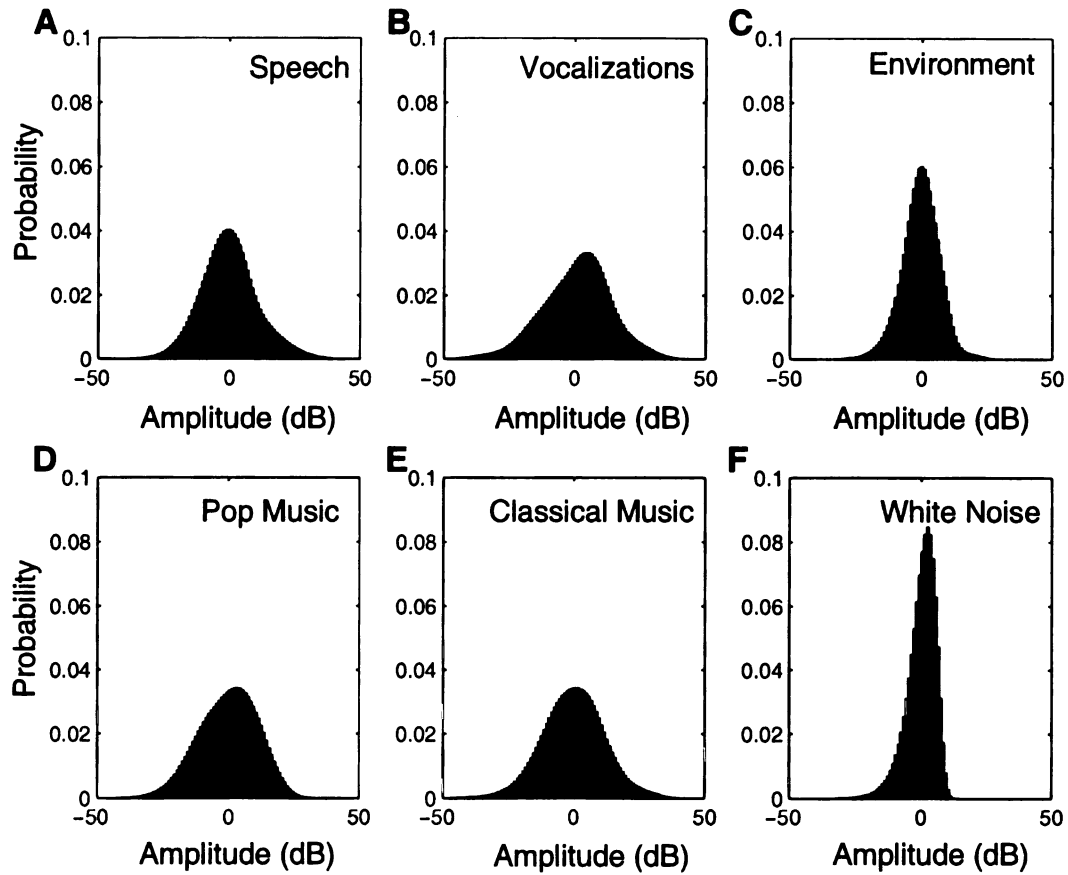


Figure 3: Decibel contrast statistics for natural sound ensembles. The decibel amplitude distribution, $p(\bar{S}_{dB})$, for speech (A), animal vocalizations (B) (both primate and nonprimate sources), background sounds (C) (e.g. wind, running water, etc.), pop-music (D) classical music (E) and white noise (F). All natural sound ensembles have Gaussian-like decibel distributions. Of these, environmental sounds has the narrowest distribution indicating that the overall range of spectro-temporal fluctuations are significantly smaller than for speech, vocalizations, and music. By comparison, white noise has the narrowest distribution indicative of a narrow range of spectro-temporal amplitude fluctuations (F).

The statistical homogeneity of the *shape* of contrast distribution across the four natural sound ensembles suggests that logarithmic amplitude fluctuations are an invariant acoustic property across natural stimuli (Attias and Schreiner 1998a). Natural sounds are therefore characterized by exponential-like amplitude distributions and Gaussian-like log-contrast which extends over average dynamic range of 14–25 dB (i.e. $2\sigma_{dB}$) and an overall range of values of roughly 50 dB. This fundamental property of natural sounds closely resemble natural image statistics which show similar spatial amplitude fluctuations (Ruderman and Bialek 1994; Dong and Atick 1995; Ruderman 1997).

4.3 Contrast Versus Intensity Response Characteristics

To test the possibility that the central auditory system is adapted for such higher-order amplitude statistics, we designed *naturalistic* ripple noise stimuli that mimic the logarithmic amplitude fluctuations observed in natural sounds (15, 30, 45, or 60 dB dynamic range) and an *artificial* control stimuli (linearly distributed contrast, modulation index=0.968) (see methods, Figs. 4 and 5). Both the naturalistic and artificial stimuli have identical spectro-temporal envelope content and differ only in their amplitude statistics (see methods). Recordings were performed from $n=63$ single neurons (su) and $n=40$ multi units (mu) in the central nucleus of the inferior colliculus. Fifteen-second sound segments were presented in a pseudo-random order for the five contrast conditions and for five RMS sound pressure levels (SPL) over a range of 50 or 75 dB (step size of 10 or 15 dB respectively). Each 15 sec sound segment was presented four times for a total of 60 seconds at any intensity-contrast condition. Intensity versus contrast response curves were derived for each neuron by measuring the mean spike rate

at all operating conditions (see methods).

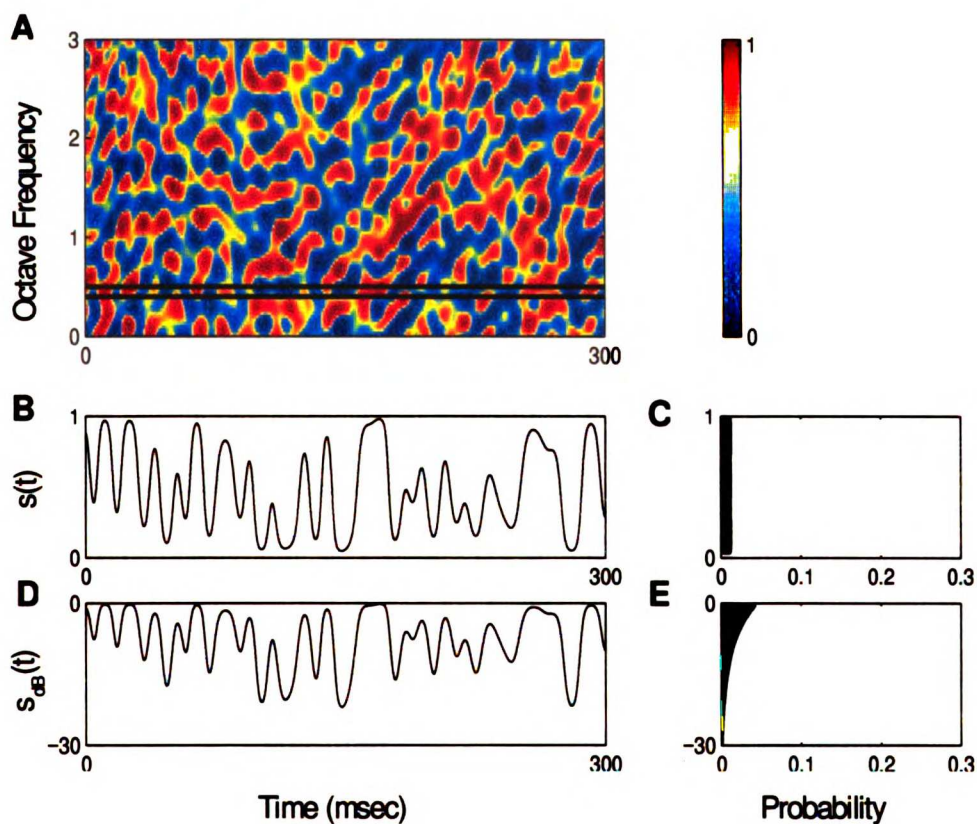


Figure 4: Artificial (linear) control ripple noise stimulus (A). The ripple noise spectro-temporal envelope has random intensity modulations along time and along the spectral dimension of the stimulus. A spectral cross-section is shown on a linear amplitude dimension (B) and on a decibel amplitude dimension (D). The linear amplitude waveform, $s(t)$, is uniformly distributed (C) and thoroughly covers the linear amplitude dimension. The corresponding decibel amplitude waveform (D) for this sound, $s_{dB}(t) = 20 \log_{10}(s(t))$, has a skewed amplitude values (E).

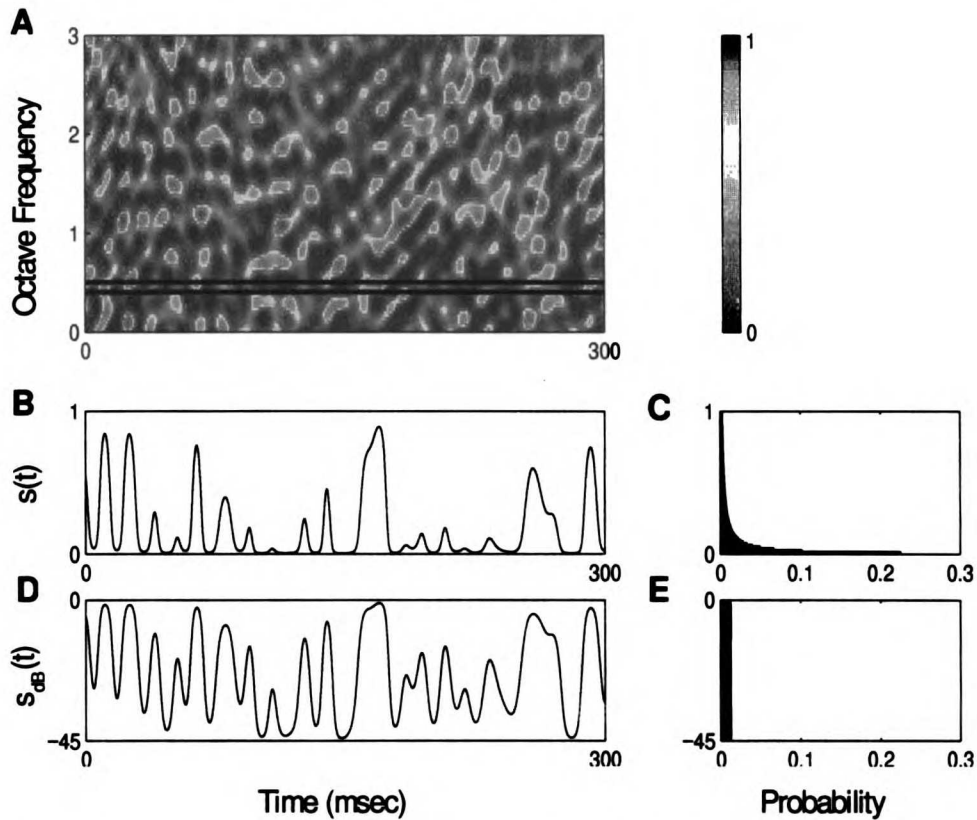


Figure 5: Naturalistic (logarithmic) ripple noise spectro-temporal envelope (45 dB) mimics the spectro-temporal envelope fluctuations observed in natural sounds (A). The naturalistic ripple noise envelope has identical spectro-temporal content as the artificial ripple noise envelope of Fig. 4. These two sounds differ only in their amplitude statistics. A spectral cross-section is shown on a linear amplitude dimension (B) and on a decibel amplitude dimension (D). The linear amplitude waveform, $s(t)$, is skewed following an exponential distribution (C) which resembles that of natural sounds. The corresponding decibel amplitude waveform (D), $s_{dB}(t) = 20 \log_{10}(s(t))$, follows a uniform amplitude distribution (E) which thoroughly covers the decibel amplitude dimension.

Intensity–contrast response curves are shown for nine single neurons in Fig. 6. As is well known from intensity coding experiments all neurons showed monotonic or non–monotonic response characteristics as a function of stimulus intensity (i.e. along the SPL axis in Fig. 6). Similar dependencies were observed for the contrast axis. Response characteristics can be increasing–monotonic (Fig. 6 A–C), tuned (Fig. 6 D–F), decreasing–monotonic (Fig. 6 I) or independent (Fig. 6 G–H) of the stimulus contrast statistics. Many units showed increasing–monotonic (significance: at least $p < 0.05$) response characteristics ($n=37$) as a function of the contrast dynamic range. In such cases the mean spike rate was minimal for the linearly distributed amplitude gradations and maximal for logarithmic contrast statistic (average=168% and median=78% firing rate increase). The mean spike rate for the *artificial* (linear amplitude) and *naturalistic* (logarithmic amplitude) contrast of 15 dB dynamic range was similar for all neurons tested (at least $p > 0.05$). Upon increasing the envelope dynamic range above 15 dB the mean spike rate increased monotonically for such neurons. Examples are provided in Fig. 6 A–C. A significant increase in firing rate was observed for the three neurons (A:

$r_{Lin} = 0.21$ spikes/sec and $r_{60} = 0.95$ spikes/sec, $p < 0.0001$; B: $r_{Lin} = 0.0$ spikes/sec and $r_{60} = 3.4$ spikes/sec, $p < 1 \times 10^{-66}$; C: $r_{Lin} = 0.15$ spikes/sec and $r_{60} = 2.15$ spikes/sec, $p < 2 \times 10^{-18}$).

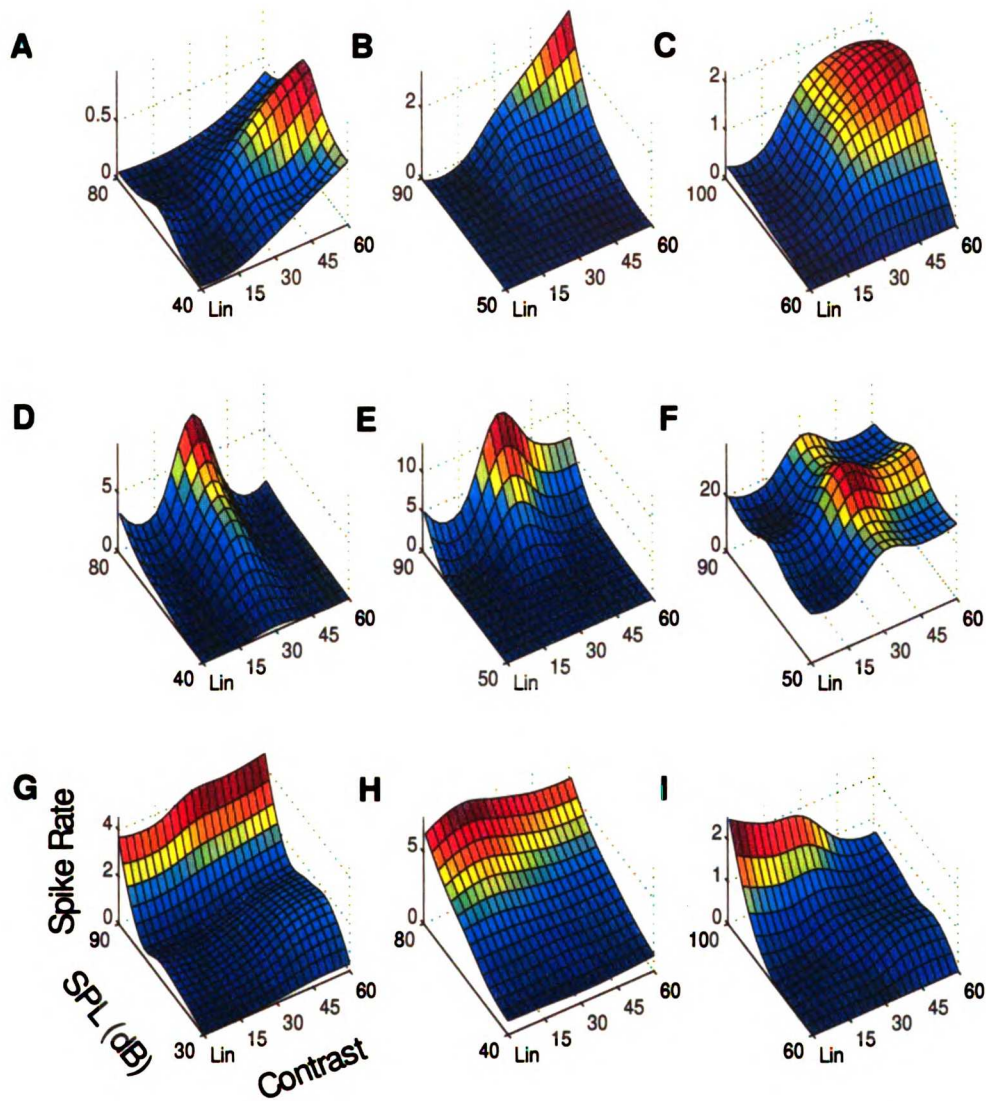


Figure 6: Contrast versus intensity response curves shown for nine single units. The ripple noise stimulus was presented in pseudo-random order at five contrast (*Lin*, 15, 30, 45, and 60 dB) and five intensity conditions (intensity spacing of 10 or 15 dB) for a total of twenty-five combinations. The measured spike rate is shown as a function of the contrast and intensity parameters. The mean spike rate of most neurons displays monotonically increasing dependency along the contrast axis (A–C). For such neurons the mean spike rate was typically low for the *Lin* contrast and increased with increasing dynamic range for the decibel ripple noise sound. A subset of neurons alternately showed non-monotonic / tuned contrast dependency (D–F) where the mean spike rate

was highest for an intermediate value of the contrast parameter (either 30 dB or 45 dB).

The remaining neurons either had a decreasing monotonic response curve (I) or displayed no statistically significant dependency with contrast (G–H).

A large number of neurons ($n=47$) showed statistically significant (at least $p<0.05$) non-monotonic responses to logarithmic amplitude statistics of different dynamic range. Such neurons are shown in Fig. 6 D–F. Responses were minimal for linearly distributed contrast and maximal for logarithmic intensity fluctuations with a dynamic range of 30 or 45 dB. Upon increasing the dynamic range to 60 dB, the neuron's response is suppressed. On the average, a 34 % (multi unit =27 %) decrease in firing rate was observed (single unit median=25 %; multi unit median=26 %). The neuron depicted in Fig. 6 D has a significant reduction (91 % ; $p<1 \times 10^{-66}$) in firing rate ($r_{30}=9.7$ spikes/sec and $r_{60}=0.85$ spikes/sec). Although the observed non-monotonic relationships were statistically significant, the overall reduction in firing rate for the 60 dB contrast was in general small. Most neurons had a subtle reductions in firing rate. The neurons shown in Fig. 6 E and F had a reduction of 49 % ($r_{30}=14.7$ spikes/sec and $r_{60}=7.48$ spikes/sec; $p<2 \times 10^{-17}$) and 16 % ($r_{30}=39.0$ spikes/sec and $r_{60}=33.8$ spikes/sec; $p<0.001$). Only 4 single neurons and 3 multi-units showed a significant decrease in firing rate to less than half of their maximum response amplitude (which occurred for either 30 dB or 45 dB). Population histograms for the percent decrease in firing rate for the 60 dB condition relative to the 30 dB or 45 dB condition is shown in Fig. 7. An additional seven single neurons showed a decreasing trend in firing rate as a function of contrast. The firing rate for this subset of neurons was maximal for

the linearly distributed contrast and minimal for the 60 dB logarithmic contrast gradations (e.g. Fig. 6 I). The remaining neurons ($n=12$) did not show a statistically significant response pattern along the contrast axis (Fig. 6 G–H).

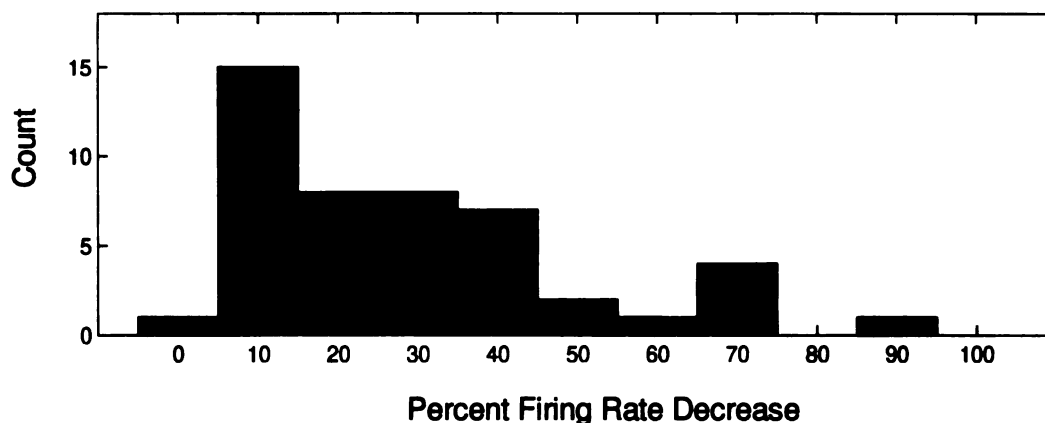


Figure 7: Reduction in firing rate for contrast non-monotonic units shown as a percent decrease relative to the maximum observed firing rate (either 30 dB or 45 dB condition). Most neurons showed only a subtle but significant reduction in firing rate for the 60 dB condition (tested for at least $p<0.05$).

4.4 Independence of Response to Intensity and Contrast Cues

The depicted contrast-intensity response curves of Fig. 6 demonstrate that stimulus intensity and contrast can, in principle, be encoded by the mean firing rate characteristics of individual neurons. The well accepted hypothesis that intensity is partly encoded by the mean firing rate of single neurons (Palmer and Evans 1979; Evans and Palmer 1980; Eggermont 1989) is consistent with this observation. What is presently not clear is how spectral and temporal fluctuations (which are themselves a form of intensity at very fine spectral and temporal scales) associated with the contrast characteristics of

the ripple sound are jointly encoded with intensity by individual or populations of neurons. It is possible that neuronal responses to these parameters covary with each other thus supporting the possibility that contrast and intensity are encoded together. An alternate and more attractive possibility is that contrast and intensity are processed independently of each other. To determine which of these two possibilities is most consistent with the observed intensity–contrast rate functions we considered a procedure that determines whether the intensity–contrast response curves form separable functions for these two parameters.

Intensity–contrast response curves were decomposed using a singular value decomposition procedure (Strang 1988). This procedure decomposes the contrast–intensity response curve into a weighted sum of functions that are each independent products of the contrast (C) and intensity (SPL) parameters. Mathematically the response function can be expressed as

$$R(C, SPL) = \sum_{k=1}^N \lambda_k \cdot u_k(C) \cdot v_k(SPL)$$

where $R(C, SPL)$ is the contrast–intensity response curve, λ_k is the k^{th} singular value, and $u_k(C)$ and $v_k(SPL)$ are functions of contrast and intensity respectively. If the contrast–intensity response curve is strictly a separable function of SPL and C , it is expected that above sum degenerates into a single term. For this unique scenario, the response of the neuron is expressed by the first term in the sum

$$R(C, SPL) = u_1(C) \cdot v_1(SPL) \quad , \quad \text{where } \lambda_1 = 1 \quad .$$

Note that the overall response for this

special case is simply a product of independent functions of C and SPL .

Fig. 8 demonstrates the general result observed for all of the studied neurons. A separable approximation ($\hat{R}(C, SPL) = u_1(C) \cdot v_1(SPL)$) of the contrast–response curve was obtained by considering only the first singular value. The separable approximation and the true contrast–intensity response curves are depicted in Fig. 8 for two single neurons. In both cases the separable approximation, $\hat{R}(C, SPL)$, captures most of the detail of the true response function $R(C, SPL)$, thus indicating that contrast and intensity responses are independent functions.

A direct measure of separability, is provided by considering the relative strength of the first singular value to the other singular values of the singular value decomposition. Thus we devise a separability index (S)

$$S = \frac{\lambda_1}{\sum_{k=1}^N \lambda_k}$$

which consists of the ratio of the first singular value, λ_1 , to the weighted sum of all the singular values (a total of $N=5$ since the measured contrast–intensity response function consists of a 5x5 matrix; 5 intensities versus 5 contrast conditions). This measure quantifies the overall fraction of the contrast–intensity response curve which the separable approximation, $\hat{R}(C, SPL)$, accounts for. Values near zero indicate that the

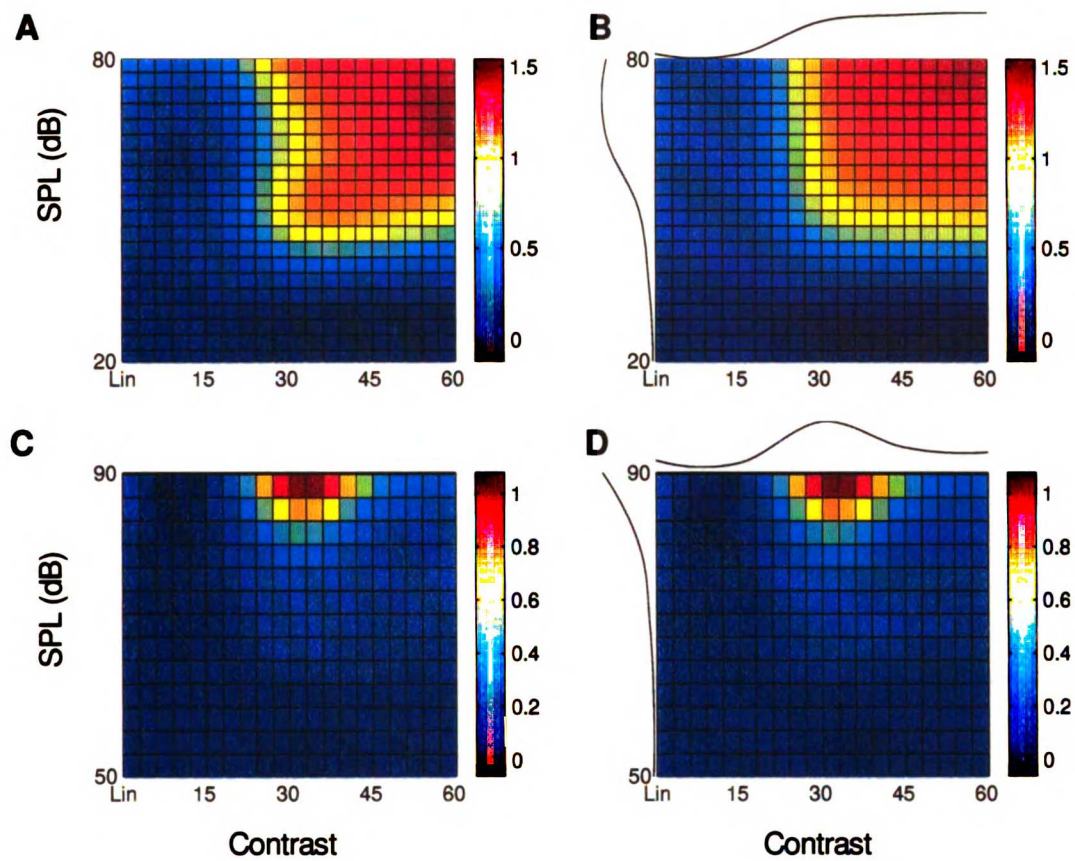


Figure 8: Separability of the contrast–intensity response function. Contrast–intensity response curves of a contrast–monotonic(A) and non–monotonic (C) single neurons.

Separable approximations, $\hat{R}(C,SPL)$ (B and D), closely approximate the true response curves of A and C. In both cases high separability index values are obtained (0.88 for B and 0.95 for D). The separable response components for contrast, $u_1(C)$, and SPL, $v_1(SPL)$, are shown above and to the left respectively of the separable response curves of panels B and D.

contrast–intensity response curve is a strongly non–separable function of these two parameters. Alternately, values near unity indicate that the measured response curve is separable. The examples of Fig. 8 exemplify this point. Both response curves are in

closely agreement to their separable approximations and consequently the measured separability index values are near unity (0.88 for the neuron of A–B and 0.95 for C–D). In the case of A, the approximation is visually not as good as for B. Accordingly, the separability index is slightly lower.

Looking at the separability index values of all single and multi units (Fig. 9) it is clear that the separable approximation accounts for most of the detail of the true response curve, $R(C,SPL)$ of all neurons. The measured separability index of $n=63$ single units and $n=40$ multi units was statistically greater than 0.75 (mean value = 0.898 ± 0.006 , t-test $p < 10^{-10}$). This finding thus indicates that the shape (i.e. monotonic, non-monotonic etc.) of the contrast firing rate response of individual single neurons is independent from their intensity response characteristics.

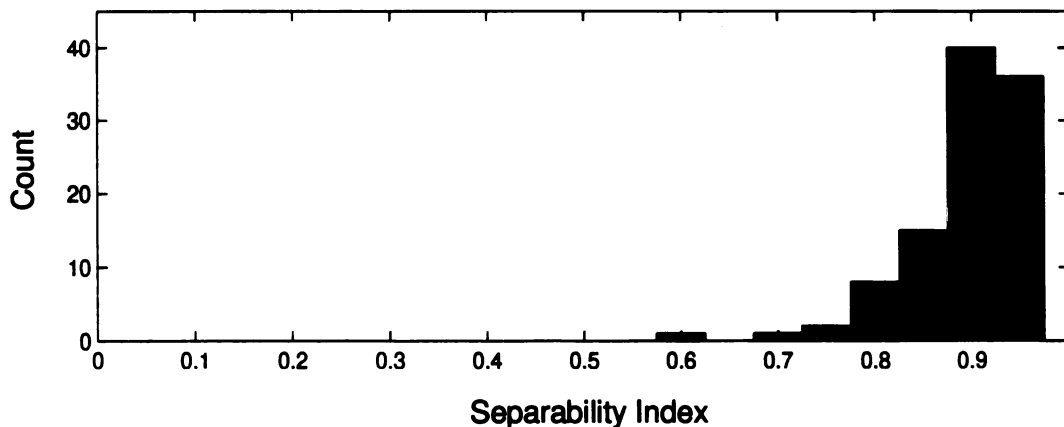


Figure 9: Separability statistics of the contrast–intensity response function. Histogram showing the separability index of $n=63$ single units and $n=40$ multi units. All neurons had very high separability index indicating that the response rate can be expressed as a separable function of contrast and intensity.

4.5 Effects of Envelope Statistics on Spectro–Temporal Coding

It is conceivable that the auditory system utilizes contrast information as a secondary acoustic cue since individual neurons can show tuned rate response curves to logarithmic contrast fluctuations. Yet for a large number of neurons the mean response rates were considerably larger for the naturalistic ripple noise (greater than 30 dB dynamic range) than for the artificial ripple noise (linearly distributed envelope statistics). This increased response rate for the naturalistic ripple noise sound suggest that ICC neurons utilize the increased dynamic range and the shape of the contrast distribution of natural sounds to encode some stimulus aspect. Do individual neurons, however, utilize the broad dynamic range in natural sounds to faithfully encode fine spectral and temporal sound components? Can individual neurons more accurately detect specific acoustic features under such "naturalistic" contrast conditions?

To test this hypothesis we computed the spectro–temporal receptive field (*STRF*) at different operating points of the contrast–intensity response curve (see methods). Ripple noise stimuli were presented at identical RMS intensity and two or more contrast conditions (*Lin* versus 30, *Lin* versus 60, 30 versus 60, or *Lin* versus 30 versus 60). Fig. 10 shows *STRFs* and the corresponding contrast–intensity response curves for three typical neurons. *STRFs* were computed at the operating points depicted by the circles on the contrast–intensity response curve (green=*Lin*, blue=30 dB, and red=60 dB). For all conditions, the shape of the *STRF* is qualitatively similar indicating that the neuron is responding to identical sound features during all contrast conditions. The neuron's mean firing rate and *STRF* amplitude, however, is significantly stronger (tested for $p < 0.01$) for the naturalistic than for the artificial ripple stimulus. Comparing the contrast–intensity

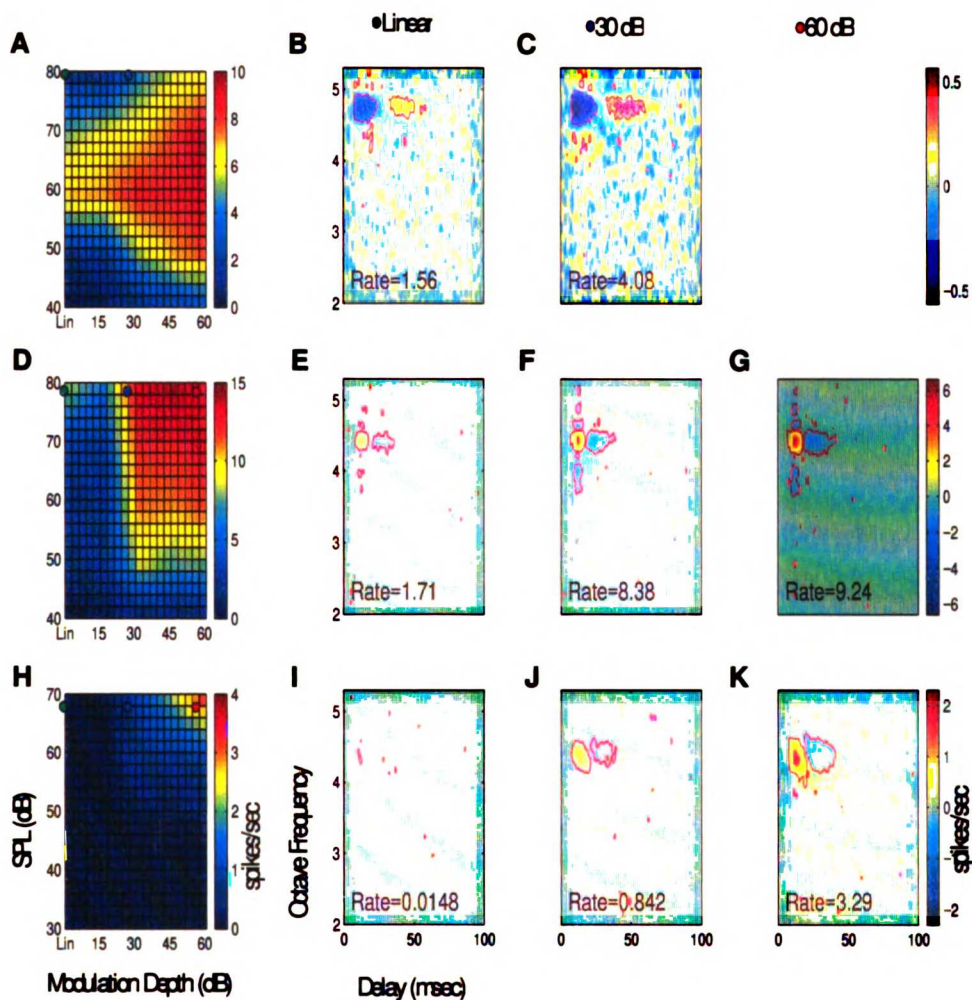


Figure 10: Relationship between the contrast-intensity response curve and the *STRF*. The contrast-intensity response curve is shown for three contrast monotonic neurons (A, D, and H). *STRFs* were computed at the contrast-intensity operating points designated by the colored circles (green=*Lin*, blue=30dB, and red=60 dB). B and C show the *STRFs* for the neuron depicted in A. The mean spike rate increased from 1.56 to 4.0 spikes/sec as the contrast was changed from *Lin* to 30dB. The amplitude of the *STRF* also increased in a similar manner. The *STRFs* for the neuron of D are shown in Figs. E-G. As expected from the contrast-intensity response curve, the mean firing rate increased monotonically as the contrast was increased from *Lin* to 60dB dynamic range. Likewise the *STRF* magnitude increases monotonically as the contrast is increased. The

neuron of panel H did not respond during the *Lin* condition but responded with increased efficacy to the 30 and 60 dB conditions.

response curves with the *STRF*, it is noted that the differential strength of the *STRF* (units of spikes/sec) is increased at contrast operating points where the mean spike rate is likewise increased. This observation indicates that the neuron utilizes the increased spike rate to encode phase-locked activity with respect to the stimulus spectro-temporal envelope. This response enhancement is typical for the vast majority of neurons.

It appears that changing the contrast operating point of the input stimulus alters the relative amplitude of the *STRF* and leaves the shape of the *STRF* unaffected, suggesting that the neuron responds to identical sounds components but with increased or decreased efficacy. To quantify this effect, we ascertained the amplitude and shape differences of the *STRF* as a function of the contrast and intensity operating point. We considered two metrics which independently quantify shape and amplitude differences of the *STRFs*. Given the experimental conditions A and B to be tested, we consider the vectorized RFs which consists of all sample values of $STRF_A$ and $STRF_B$ which exceed a significance test ($p < 0.002$) (see methods) for condition A or for condition B. The vectorized RFs, RF_A and RF_B , thus consists only of the sample values of $STRF_A$ and $STRF_B$ for which either of the *STRFs* exceeded the significance test. To quantify the similarity of the *STRFs* we consider the correlation coefficient or similarity index (SI) (DeAngelis *et al.* 1999; Reich *et al.* 2000)

$$SI_{A,B} = \frac{\langle RF_A, RF_B \rangle}{\|RF_A\| \cdot \|RF_B\|}$$

where RF_A and RF_B are the significant *STRF*s for condition A and condition B respectively, $\langle \cdot, \cdot \rangle$ corresponds to the vector inner product, and $\|\cdot\|$ designates the vector norm operator. The similarity index quantifies the *STRF* shape differences or similarity independently of *STRF* amplitude. The *SI* assumes a numerical value normalized to the range -1 to 1 . Values near 1 indicate maximal shape similarity between $STRF_A$ and $STRF_B$, whereas values near 0 indicate that the *STRF*s have nothing in common and are thus orthogonal. *SI* values near -1 indicate that both RFs have similar spectro-temporal patterns but differ by a sign inversion.

Amplitude differences are characterized by the amplitude similarity index

$$ASI_{A,B} = s \cdot \left[\left(\frac{\|RF_A\|}{\|RF_B\|} \right)^s - 1 \right] \times 100\%$$

where $s = \text{sign}(\|RF_A\| - \|RF_B\|)$. The *ASI* metric assumes values between negative and positive infinity. A value of zero indicates that $\|RF_A\| = \|RF_B\|$ whereas values > 0 indicate that $\|RF_A\| > \|RF_B\|$. Values < 0 alternately indicate that $\|RF_A\| < \|RF_B\|$. The magnitude of $ASI_{A,B}$ is numerically equivalent to the percent difference between $\|RF_A\|$ and $\|RF_B\|$ where the sign of $ASI_{A,B}$ indicates an increase in the *STRF*

amplitude referenced on condition A (for negative values) or B (for positive values). A similar metric was also used to characterize the mean response rate differences for two experimental conditions. We consider the rate similarity index (RSI)

$$RSI_{A,B} = s \cdot \left[\left(\frac{r_A}{r_B} \right)^s - 1 \right] \times 100\%$$

where $s = \text{sign}(r_A - r_B)$. This metric is numerically identical to the ASI where the mean rates for conditions A and B are substituted for the *STRF* norms for those conditions. The RSI and ASI differ since the RSI measures mean rate changes over the stimulus duration whereas the ASI measures stimulus driven activity (note that the *STRF* is a direct measure of the stimulus phase-locked differential spike rate produced by a given stimulus pattern relative to the mean spike rate).

The neuron of Fig. 10 E–G has relatively large value of *SI* for all contrast conditions ($SI_{60,30} = 0.97$, $SI_{60,Lin} = 0.92$, $SI_{30,Lin} = 0.91$) indicating that the neuron responded to identical spectro-temporal sound patterns at any given operating point. Despite the similarity in spectro-temporal shape, the RSI and ASI coefficients

indicate that the neuron respond with a higher spike rate ($RSI_{30,Lin} = 390\%$,

$RSI_{60,Lin} = 440\%$) and stronger differential response strengths ($ASI_{30,Lin} = 395\%$,

$ASI_{60,Lin} = 468\%$) for the 30 dB and 60dB condition relative to the *Lin* contrast.

Similar trends are observed for the neurons of Fig. 10 B–C ($SI_{30,Lin} = 0.86$,

$ASI_{30, Lin} = 116\%$, $RSI_{30, Lin} = 161\%$). The neuron of Fig. 10 I–K did not respond to the linear contrast condition but responded strongly to the 30 dB and 60 dB conditions. Consequently, this neuron has small SI values ($SI_{30, Lin} = 0.21$ and $SI_{60, Lin} = 0.28$) and large ASI ($ASI_{30, Lin} = 4,558\%$ and $ASI_{60, Lin} = 18,120\%$) and RSI values ($RSI_{30, Lin} = 5,589\%$ and $RSI_{60, Lin} = 22,130\%$) .

Similarity index population data is shown in Fig. 11 for $n=57$ single neurons and $n=75$ multi units. Multi unit and single unit data was polled together for the various contrast conditions (30 versus *Lin*, 60 versus *Lin*, and 60 versus 30) since they all followed similar trends. The vast majority of neurons had high SI values (mean value of 0.77, median value of 0.87) across multiple contrast conditions supporting the initial observations shown for the neurons of Fig. 10. Thus in all instances neurons responded to similar spectro–temporal sound features. A small number of neurons (12 single units and 7 multi units) had low SI values ($SI < 0.5$). These neurons were observed only for the 30 versus *Lin* or 60 versus *Lin* conditions. In all instances these neurons had low spike rates and did not produce a statistically significant *STRFs* ($p < 0.002$) during the *Lin* condition (see the example neuron of Fig. 10 I–K) but produced statistically significant *STRFs* ($p < 0.002$) for the 30 or 60 dB conditions.

The RSI and ASI metrics were computed for all single and multi units to compare the response rate and *STRF* energy differences for the three contrast conditions. The initial observation for the single units of Fig. 10 supports the hypothesis that ICC neurons respond more efficiently to decibel amplitude fluctuations. This hypothesis is further supported by the population data of Figs. 12–14. Histograms for the ASI and RSI metric

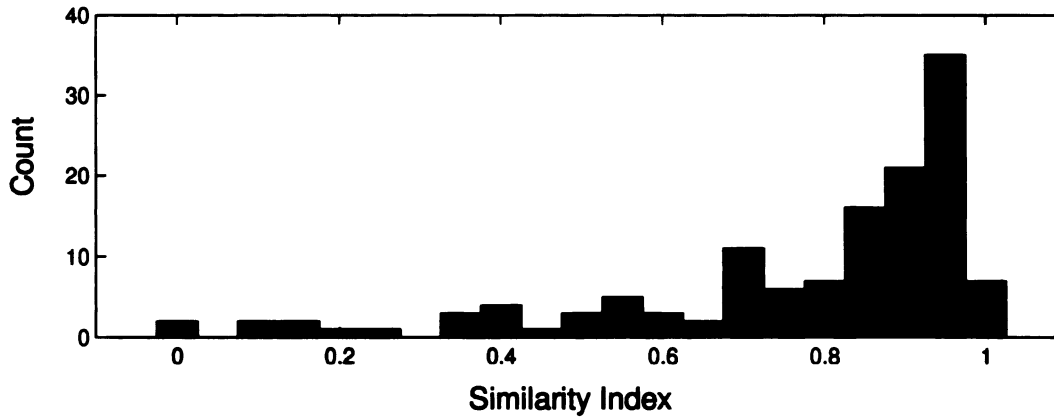


Figure 11: Population similarity index histogram. Similarity index measurements were obtained by measuring the *STRF* correlation coefficient for the 30dB versus *Lin*, 60dB versus *Lin*, and 60dB versus 30dB contrast conditions ($n=57$ single units and $n=75$ multi units). The population histogram is highly skewed towards positive one (mean=0.77, median=0.87) indicating that the obtained *STRFs* for the different contrast conditions have similar spectro-temporal patterns.

were positively skewed and had only a few negative values. On the average, a large

increase in spike rate (geometric mean: $RSI_{30, Lin}=98\%$ and $RSI_{60, Lin}=168\%$,

$RSI_{60,30}=69\%$; median: $RSI_{30, Lin}=60\%$ and $RSI_{60, Lin}=172\%$, $RSI_{60,30}=21\%$

) and *STRF* energy (geometric mean: $ASI_{30, Lin}=118\%$ and $ASI_{60, Lin}=197\%$,

$ASI_{60,30}=52\%$; median: $ASI_{30, Lin}=103\%$ and $ASI_{60, Lin}=141\%$,

$ASI_{60,30}=22\%$) was observed for the 30 or 60 dB contrast relative to the Linear contrast condition. Furthermore, both the ASI and RSI are significantly correlated (

$\rho_{30, Lin}=0.95\pm 0.05$ and $\rho_{60, Lin}=0.95\pm 0.04$, $\rho_{60,30}=0.99\pm 0.03$), indicating that

the observed increase in mean firing rate is accompanied directly by an *STRF* strength

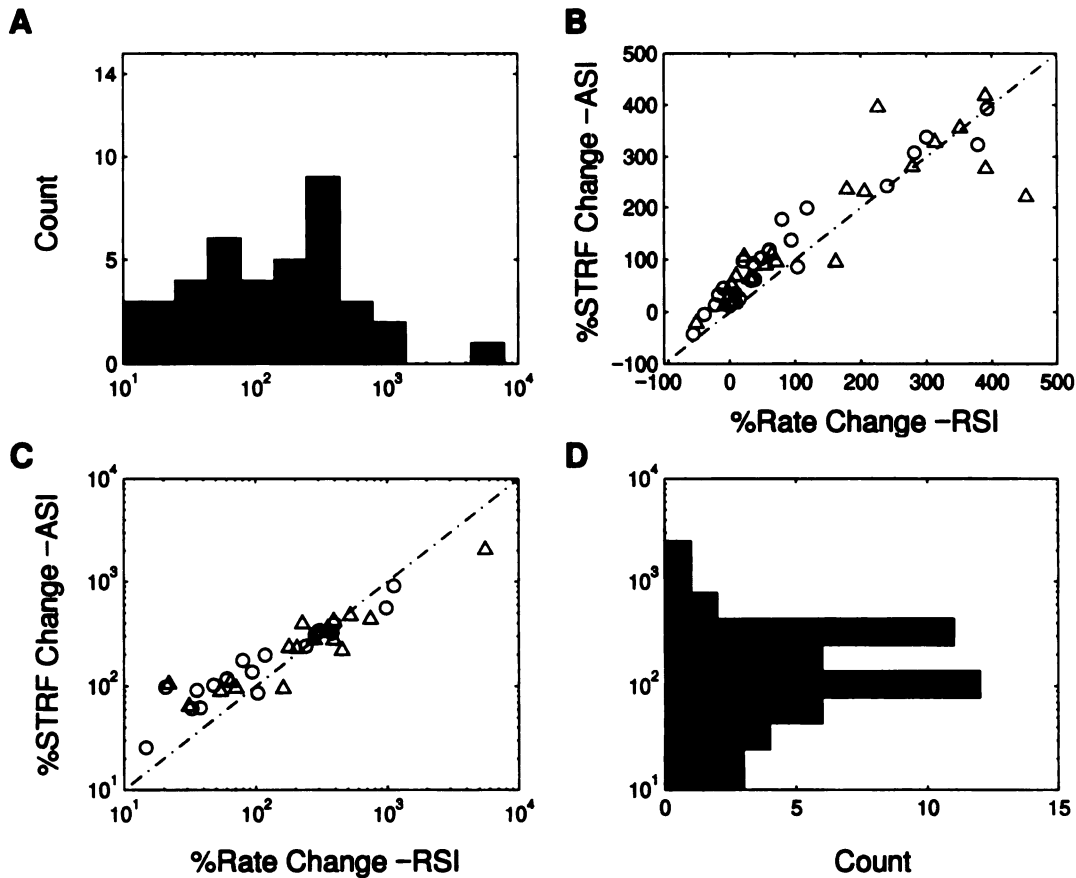


Figure 12: Population RSI and ASI statistics ($n=26$ single unit, opened circles; $n=27$ multi unit, open triangles) comparing 30dB versus *Lin* conditions. RSI (A) and ASI (D) population histogram (shown on a semilog plot for positive values only) show a large percent increase in the mean firing rate (A) and the *STRF* strength (D) for the 30dB contrast condition. Firing rate and *STRF* energy increases of more than 100% were observed for 29/53 neurons. Scatter plot of RSI and ASI (B) shows a significant correlation between the observed percent rate increase and percent *STRF* energy increase. Shown on a linear plot (B) and a log-log plot for positive values (C). Few negative values are observed for both RSI and ASI ($n=4$ single unit and $n=6$ multi unit) indicating that the responses for the 30dB were stronger on the average than for the *Lin* condition. RSI versus ASI (C) shown for positive values only on a log-log plot.

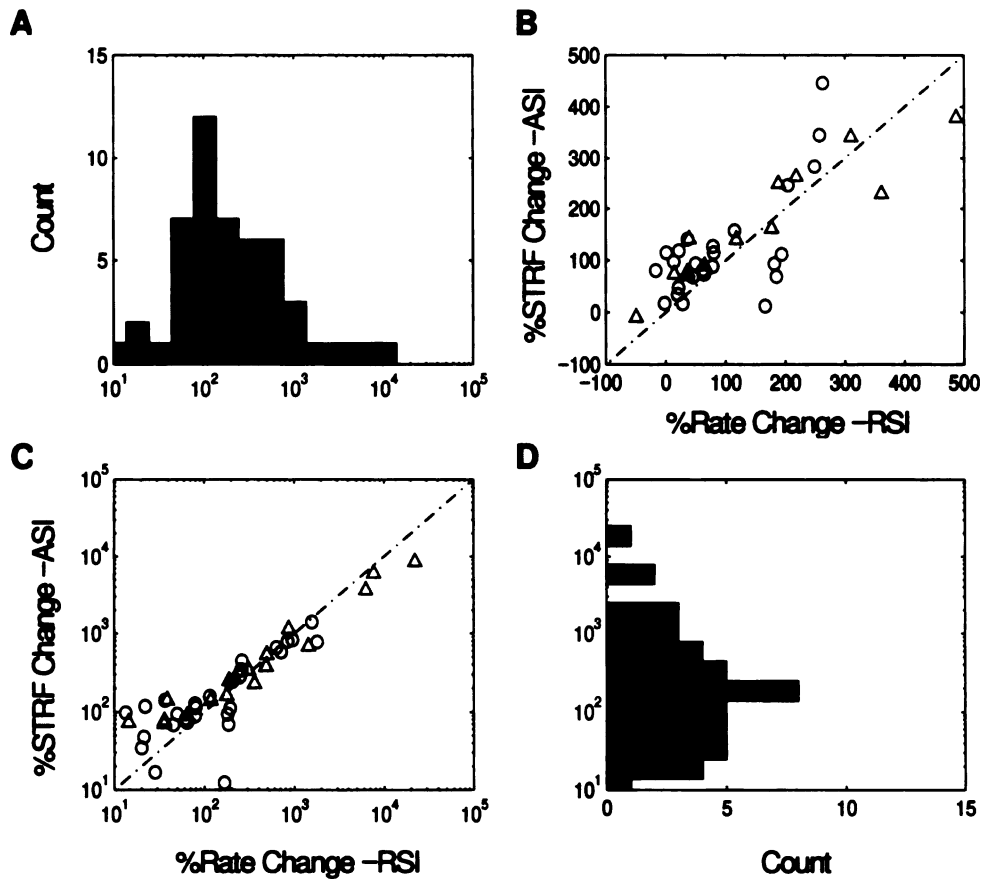


Figure 13: Population RSI and ASI statistics ($n=19$ single units, open circles; $n=31$ multi units, open triangles) comparing 60dB versus *Lin* conditions. RSI (A) and ASI (D) population histogram (shown on a semilog plot for positive values only) show a large percent increase in the mean firing rate (A) and the *STRF* strength (D) for the 60dB contrast condition. Firing rate and *STRF* energy increases of more than 100% (1000%) were observed for 34/50 (7/50) neurons tested. As for the 30dB condition (Fig. 12), a scatter plot of RSI versus ASI (B) shows a significant correlation. Shown on a linear plot (B) and a log-log plot (C). The distribution of RSI and ASI were positively biased (only 3 negative values observed) for both RSI and ASI indicating that the responses to the 30dB contrast were stronger on the average than for the *Lin* condition. RSI versus ASI (C) shown for positive values only on a log-log plot.

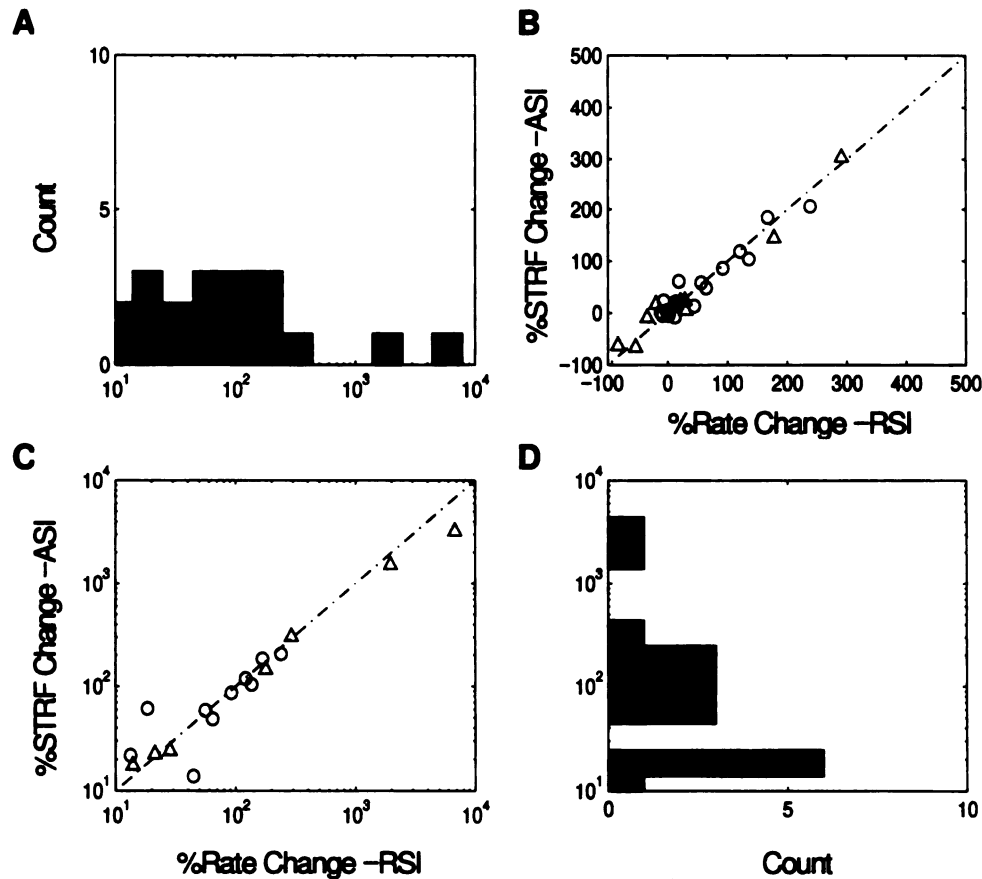


Figure 14: Population RSI and ASI statistics ($n=12$ single units, open circles; $n=17$ multi units, open triangles) comparing 60dB versus 30dB conditions. RSI (A) and ASI (D) population histogram (shown on a semilog plot for positive values only) show a moderate percent increase in the mean firing rate (A) and the *STRF* strength (D) for the 60dB contrast condition. Firing rate or *STRF* energy increases of more than 100% were observed for 8/29 neurons tested. As for the 30dB case (Fig. 12), scatter plot of RSI and ASI (B) shows a significant correlation between the observed rate increase and *STRF* energy increase. The distribution of RSI and ASI were positively biased for both RSI and ASI indicating that the responses to the 60dB contrast were stronger on the average than for the 30 dB condition. A total of 11 neurons had negative ASI or RSI ($n=4$ single and $n=7$ multi units) all of which had values greater than -100% . RSI versus ASI (C) shown for positive values on a log-log plot.

increase. Since the additional spikes for the 60 and 30 dB conditions (compared to the *Lin*) must be time-locked to the stimulus in order to produce a difference rate increase in the *STRF*, this observation suggests that the additional spikes produced during the logarithmic ripple noise encode additional spectro-temporal information. Thus, functionally, the increased firing rate provides additional spikes for which to encode spectro-temporal stimulus components.

4.6 Spike Timing Precision and Response Reproducibility

The increase in mean firing rate and *STRF* strength for the decibel contrast conditions indicate that ICC neurons have additional spikes available to encode spectro-temporal acoustic information. Given that neurons have similar *STRFs* for the *Lin* and dB conditions and the fact that the *Lin* and dB sounds have identical spectro-temporal content (since they differ only in their amplitude statistics) further suggests that ICC neurons encode information about similar acoustic features for all the conditions tested. Given that the overall spike rate and *STRF* strength is higher for the decibel contrast it is expected that the overall information rate of the neuron is higher for this condition (since the information rate is proportional to the mean firing rate). What is not presently clear, is whether the increase in firing rate for the decibel stimulus is accompanied by an increase in spike timing precision and response reproducibility (on a trial to trial basis). If so, it is expected that individual spikes convey more information about the sensory stimulus for the decibel than for the linear contrast. To determine this, we need to assess the response contribution of individual spikes.

A short sound segment of the ripple noise stimulus (5 seconds) was presented for

150 trials. Response traces were recorded for each trial and the response reproducibility was determined by measuring the mutual information (de Ruyter et. al. 1997). Each spike trace was digitized at a sampling resolution of $\Delta t = 1$ msec and the spike train entropy was determined by measuring the probability distribution, $P(W)$, of possible 10-bit words, W . A search through the whole experiment was conducted to determine the word distribution, $P(W)$. Using the distribution of 10-bit words the spike train entropy is determined as

$$S_{total} = \sum_W P(W) \log_2(P(W)) .$$

This measure provides a theoretical upper limit on the amount of information which a spike train can convey. To determine the noise inherent within the response, the noise entropy was computed by determining the trial-by-trial reproducibility of the response (e.g., the entropy in the spike train that does not convey any viable information about the stimulus). At any given time instant, t , the conditional probability distribution of obtaining a given 10-bit word was computed, $P(W|t)$. The noise entropy was then determined as

$$S_{noise} = \left\langle - \sum_W P(W|t) \log_2(P(W|t)) \right\rangle_t,$$

where $\langle \cdot \rangle_t$ is the conditional ensemble expectation computed over all time. The

information which the spike train contains about the stimulus (i.e. the mutual information) is determined by subtracting these two quantities

$$I = S_{total} - S_{noise} .$$

The spiking patterns of ICC neurons to the ripple noise stimulus are characterized by phasic response components as depicted in Fig. 15. The response rasters and peri-stimulus time histograms (PSTH) show a precisely time-locked signature down to millisecond resolution. Inspection of the response rasters and PSTH for the linear and decibel contrast immediately reveals systematic changes in firing rate and spiking precision. For the two examples shown, the increase in firing rate observed for the decibel contrast relative to the linear contrast (mean firing rates A-C: $Rate_{Lin} = 9.4$, $Rate_{30} = 10.7$, $Rate_{60} = 13.5$ spikes/sec; E-F: $Rate_{Lin} = 14.4$, $Rate_{30} = 17.9$ spikes/sec) was accompanied by an increase in peak to trough amplitude of the phasic response components (Fig. 15 D and G). Furthermore, the responses appear as more reproducible for the decibel contrast (as reflected in the response rasters: Fig. 15 A-C and E-F).

To quantify this observation, the mutual information was computed for all contrast conditions. The systematic increases in the observed reproducibility are reflected directly in the measured mutual information (A: 3.5 ± 0.1 bits/spike B: 4.63 ± 0.07 bits/spike C: 4.58 ± 0.07 bits/spike; E: 0.594 ± 0.03 bits/spike; F: 0.852 ± 0.02 bits/spike).

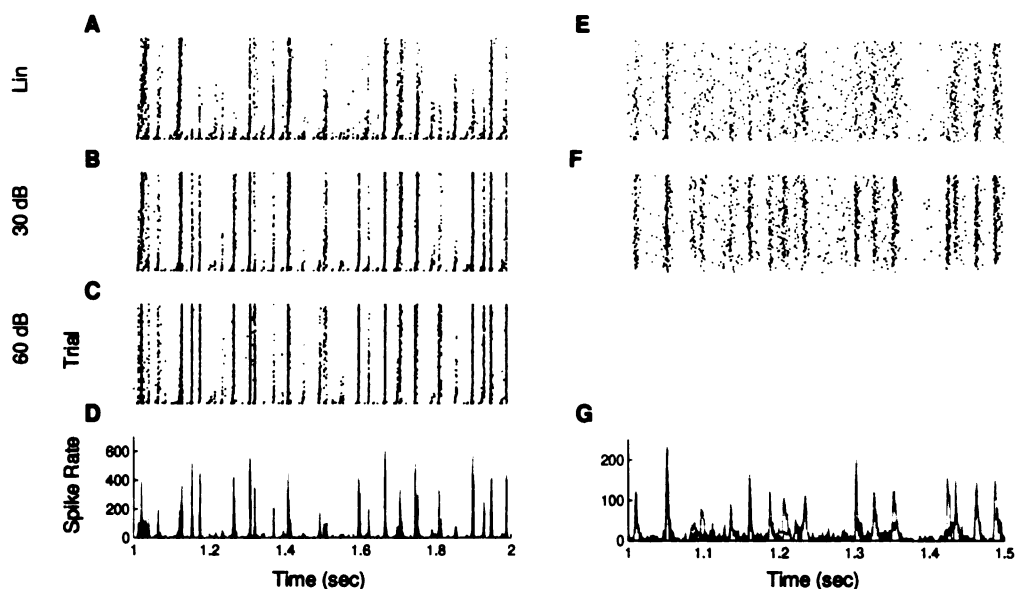


Figure 15: Spiking precision and reproducibility as a function of contrast for two single neurons. A short (5 second) segment of the *Lin*, 30 dB, and/or 60 dB ripple noise was presented. Rastergrams showing 150 response traces to the ripple noise: *Lin* (A), 30dB (B) and 60 dB (C) for neuron 1 and *Lin* (E) and 30 dB (F) for neuron 2. Each spike is shown as a single dot (bin width: 1 msec). Spike timing precision and response reproducibility is poor for the *Lin* (A and E) condition. Responses to the 30 dB and 60 dB ripple noise are significantly more reproducible than for the *Lin* contrast. In general two effects are observed: individual rasters become more precise as the contrast is changed from *Lin* to 30 and 60 dB and new responses are observed for the 30 and 60 dB ripple noise. The peri-stimulus time histogram (PSTH) depicts the instantaneous rate of the neuron as a function of time (D and G). For the *Lin* condition (blue) the response reproducibility is poor. By comparison, both neurons have stronger and more precise rate fluctuations for the 30 dB (green) and 60 dB (red) conditions.

Thus, on a per-spike basis these neurons contribute more information when driven with the decibel as opposed to the linear contrast ripple noise. Taken together, the added

reproducibility, increase in mean firing rate and the fact that the *STRF* magnitude is greater for the decibel contrast suggests that the added information carried by the spike train is used directly for more efficient spectro-temporal coding.

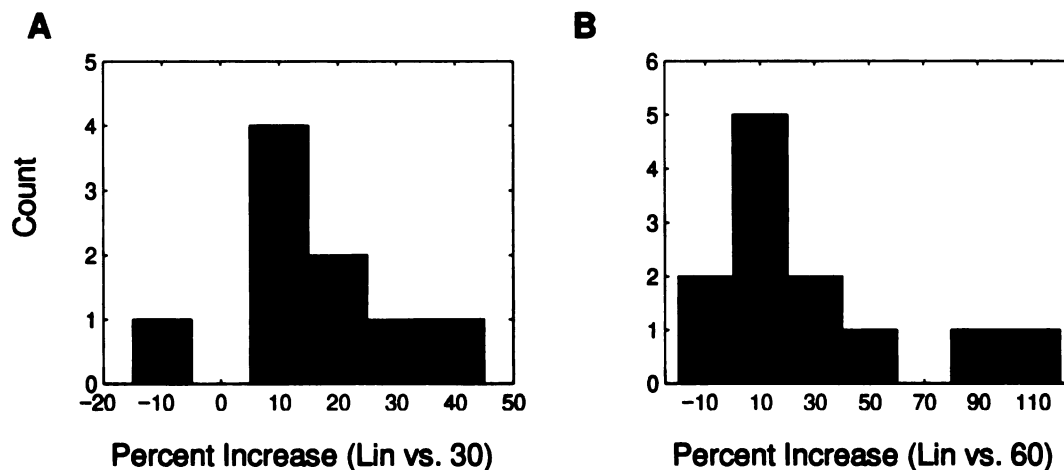


Figure 16: Percent increase in mutual information for the 30 (A) and 60dB (B) contrast conditions relative to the Linear contrast. Histograms showing the percent increase in mutual information for both conditions. Both histograms are positively skewed. Mutual information is therefore greater (on the average) for the decibel sounds.

The mutual information per spike was computed for $n=7$ single neurons and $n=14$ multi units. To assure that sufficient averaging was performed, estimates of the mutual information were computed only for neurons with mean spike rates greater than 5 spikes/sec. Histograms for the percent increase in mutual information (Fig. 16 A: Linear versus 30dB; Fig. 16 B: Linear versus 60 dB) are shown in Fig. 16. Neurons for which

$I_{30dB} > I_{Lin}$ and $I_{60dB} > I_{Lin}$ are designated by a positive percent increase. Both histogram are positively skewed indicating that the mutual information was on the

average greater for the decibel contrast condition. For the 30 dB condition a statistically significance increase in mutual information 15.2% (t-test, $p < 0.01$) was observed. For the 60 dB contrast the percent increase in mutual information was almost doubled (29.7%; t-test, $p < 0.01$). On a per-spike basis, neurons therefore carry more stimulus related information for the decibel contrast conditions.

4.7 Discussion and Conclusion

The presented study demonstrates that auditory midbrain neurons utilize their large intensity operating range for efficiently encoding spectro-temporal stimuli with similar dynamic range. Using spectro-temporal reverse correlation procedures along with an information theoretic analysis it is shown that responses of auditory neurons in the central nucleus of the inferior colliculus are strongly modulated by higher-order amplitude statistics of the spectro-temporal waveform. Neurons show increased spike rates to the decibel modulated ripple noise, stronger *STRFs*, and higher response reproducibility. Given that all sensory systems have operating ranges which span several orders of magnitude and both acoustic and visual stimuli have instantaneous spectral (spatial) and temporal gradations which are logarithmically distributed and span similar ranges (see Chapt. 1; Ruderman and Bialek 1994), we propose that the large operating range of sensory systems is utilized for spectro- (spatio-) temporal coding in addition to loudness (luminance) coding.

It is possible that the observed response differences between the linear and the logarithmic contrast conditions arise from trivial low-order stimulus parameter (such as the modulation depth or the maximum intensity) that covaries as a function of the

performed contrast alterations (i.e. changing from linear to logarithmic amplitude modulations). To guarantee that this is not so, a number pertinent stimulus statistics were closely examined. Statistics for all conditions are depicted in Table 1.

	<i>Lin</i>	15 dB	30 dB	45 dB	60 dB
β	0.965	0.822	0.965	0.995	0.999
σ_{Lin}	0.280	0.232	0.257	0.244	0.226
σ_{dB} (dB)	6.7	4.3	8.7	13	17.3
ΔI (dB)	0	1.6	0.75	1.2	1.9
skewness	0	0.59	1.12	1.57	1.96

Table 1: Low- and high-order statistics of the ripple noise sound shown for the different contrast conditions (*Lin* and 15–60 dB): modulation depth (β), linear amplitude standard deviation (σ_{Lin}), dB-amplitude standard deviation (σ_{dB}), intensity offset (ΔI), and skewness. Low-order statistics (e.g. β and σ_{Lin}) are similar for all conditions. Higher-order statistics such as the decibel standard deviation and linear amplitude skewness are largely varied and covary with the observed neuronal response changes.

Unlike other studies in the visual and auditory system which use linear amplitude gradations and vary the modulation index (i.e. the linear contrast) over a large range of values (as low as 0.05 to 1), the presented data was obtained using a relatively high modulation index for all conditions (0.822 – 0.999). Thus, the peak to trough ratio was effectively maximal for all contrast conditions. It is, therefore, unlikely that the observed effects are related to the maximum and minimum amplitude values of the stimulus

waveform. This possibility is further supported by the fact that the Linear control stimulus was designed specifically so that its modulation index, β , is identically matched to the 30 dB logarithmic sound ($\beta = 10^{-30/20} = 0.968$ for both). Despite this fact, many neurons showed significantly stronger responses for the 30 dB condition. The observed increase in firing efficacy for the 30 dB over the *Lin* condition can, therefore, not be accounted for by the modulation depth parameter of the ripple noise.

Another pertinent parameter that must be considered is the maximum sound intensities. Upon matching the RMS intensity for the different stimulus conditions the maximum intensities for each sound were no longer matched. This small but undesirable effect is a direct consequence of the fact that the spectro-temporal envelope waveforms had varying degrees of skewness for the different conditions. Given this small intensity disparity for the different conditions, it is theoretically possible that the observed results arise from the neurons' rate-level dependencies. For example, upon matching the RMS sound pressure level (SPL) for the *Lin* and 30 dB (60 dB) contrast the maximum spectro-temporal amplitude of the 30 dB (60 dB) ripple noise is 0.76 dB (1.8 dB) above the *Lin* contrast ripple noise. It is possible that this small intensity increment could modify the neuron's overall firing rate. Judging from the small values ΔI and previous data on intensity dependency of auditory neurons (Palmer and Evans 1979; Evans and Palmer 1980; Ehret and Merzenich 1988; Viemeister 1988; Eggermont 1989), however, it is unlikely that the large response disparity between the *Lin*, 30 dB, and 60 dB contrast be accounted for by this small intensity increment.

Likewise low-order statistics of the stimulus amplitude (e.g. the standard deviation) which are most often used to quantify linear contrast, both in the visual and

auditory literature, are unlikely to account for the observed disparity in the response to the *naturalistic* (30dB) and *artificial (Lin)* contrast. The linear amplitude envelope standard deviations ($\sigma_{Lin}=0.280$, $\sigma_{30}=0.257$) and the decibel amplitude standard deviations ($\sigma_{Lin}=6.70$ dB and $\sigma_{30}=8.65$), for example, are similar for these two sounds. These low-order descriptors differ only by -8.2% when computed in the linear amplitude dimension and 29.9% (30 dB relative to the *Lin*) for the decibel amplitude dimension respectively and therefore do not account for the large disparity in response strengths observed between the *artificial (Lin)* and the 30 dB *naturalistic* stimuli (average rate increase of 98%, geometric mean).

Two candidate stimulus parameters can account for many of the observed response differences, both of which are related to shape of the amplitude distribution for these sounds. As observed from Figs. 4 and 5, the contrast distribution for the artificial and naturalistic ripple noise differ depending on whether the distribution is computed in the linear or the decibel amplitude dimensions. As for natural sounds (Figs. 1 A and 2), the linear amplitude distribution of the naturalistic ripple noise has skewed values about zero. As the dynamic range of the naturalistic ripple noise sounds is increased from 15 dB to 60 dB, the measured skewness increases accordingly from 0.59 to 1.96. By comparison, the artificial ripple noise stimulus is perfectly symmetric with a skewness value of zero. Given the response efficacy increased on the average with increasing decibel dynamic range it is likely that the observed differences can be accounted by the skewness of the ripple noise stimulus. A first order descriptor which also accounts for observed response differences is the decibel standard deviation (σ_{dB}) of the sound.

Note that the skewness and σ_{dB} covary and consequently either can account for many of the observed response differences.

Studies of pure tone transients and onsets in the primary auditory cortex have demonstrated that first spike latency and response amplitude are strongly affected by the peak acceleration and peak velocity of the transient window function used (Heil 1997a 1997b). In particular, the time to first spike latency is inversely proportional to the peak acceleration (or velocity, depending on the type of transient window) of the sound pressure envelope. Furthermore, the standard deviation of the first-spike latency decreased with increasing peak acceleration (down to a few milliseconds) of the sounds envelope suggesting that the responses to more transient windows produced more precise and reproducible responses (Heil 1997a). Since comparable results are observed as early as the auditory-nerve it has been suggested that this general mechanism holds throughout the auditory pathway (Heil and Irvine 1997) .

The fact that these results are not invariant to the type of window (Heil 1997a 1997b) indicates that additional higher-order statistics of the sound pressure envelope need to be considered in order to fully describe the response characteristics for different window functions. One possibility is that the decibel amplitude dimension accounts for all or most of the necessary higher-order stimulus characteristics which give rise to the observed results. Note that as for the reported results on spike timing and response strength as a function of peak pressure velocity and peak pressure acceleration (Heil 1997a 1997b) our findings with the ripple noise sound have similar implications for sound processing. By increasing the overall dynamic range (units of dB) of the ripple noise stimulus one effectively alters the skewness of the ripple noise envelope (and vise

versa). Since the skewness of the envelope is directly related to the acceleration and velocity of the sounds spectro-temporal envelope (see Fig. 5) the observed findings are consistent with those of Heil (1997a and 997b). Note that as the dynamic range of the ripple noise envelope is increased, individual transients of the linear amplitude waveform become sharper and more pronounced. Thus the rate of change and acceleration of the envelope increase accordingly with increasing dynamic range. The fact that mean spike rates in general tend to increase with the overall dynamic range and spike timing reproducibility increases with the above parameters is consistent with previously reported findings.

Our findings largely differ from those of Heil (1997a and 1997b) since we now introduce the spectral dimension and since we ultimately consider a significantly more dynamic scenario. Thus, in addition to temporal onsets, we additionally consider temporal offsets, and spectral resonances that are produced by the ripple noise. The ripple noise represents a unique sound processing scenario where the mean intensity is held constant but yet spectral and temporal fluctuations coexist along the sensory epithelium. Consequently it can be thought of as a borage of spectral and temporal features that are superimposed on a mean level of noise. This is of interest since natural sounds are ubiquitously composed of both spectral and temporal sound components (which may or may not be independent) that ultimately produce a complex excitation pattern on the primary sensory epithelium in the cochlea. The spectrographic sound pattern, thus, highlights many of the prominent acoustic features that are relevant to the auditory neuronal network. Furthermore, natural signals generally do not occur in isolation and are often superimposed on a background of noise. Given the earlier findings

on natural sounds statistics, our findings have direct implication for spectro-temporal processing of natural sounds by the brain since these have spectro-temporal envelopes with decibel distributed amplitude fluctuations. These findings and the subsequent mechanisms provide evidence that the dynamic range and skewness of natural sound envelopes are a critical cue for spectro-temporal sound processing.

The hypothesis that the human ear is adapted to encode signals with large dynamic range is supported by human speech studies. Analysis of the spectral envelope have shown the peak to valley ratios in human speech (vowel formants) can extend over more than 20 dB (Plomp 1983). Resonances associated with vowel formants provide a critical cue for the perception of vowel sounds. Given the large amount of across subject variability inherent in natural speech, human perception must be robust to envelope alterations under many operating conditions. Production of same vowels by different speakers, for examples, shows a large amount of inter-subject variability with an average intensity standard deviation near 16 dB (Klein, Plomp, and Pols, 1970). Additional sources of envelope noise (roughly 5 dB) are introduced by the reverberant characteristics associated with environmental and room acoustics (Schroeder 1954). Yet psychoacoustic studies indicate that perception of vowel sounds is robust to such alterations (Van Veen and Houtgast 1983 1985; Ghitza and Goldstein 1983). The just noticeable peak to valley ratio of a vowel's envelope is exceptionally large (roughly 10 dB, Flanagan, 1970, 1972; Ghitza and Goldstein 1983) in comparison the JND sound intensity for broadband sounds (about 0.5 dB). Such contrast related cues likely play a critical role in speech perception since increasing the decibel contrast alters the perceived sound in such a manner that it improve inter-category discrimination of vowels sounds

(Van Veen and Houtgast 1983 1985).

A critical question instantly raised by the presented findings is whether the nervous system processes linear amplitude or decibel amplitude gradations. In essence this gets at the question of: "what is the fundamental variable for defining sensory signals?" Should one use linear spectro-temporal amplitude variable, $s(t, f)$, or the corresponding decibel amplitude variable, $s_{dB}(t, f) = 20 \log_{10}(s(t, f))$. Historically temporal modulation signals (both in auditory and visual modalities) have been defined on a linear amplitude dimension largely because of convenience and because these go along with the mathematical conventions devised for communications engineering (which came about during the advent of telephony). In general there is no a priori reason for using either linear or decibel amplitude fluctuations as the pertinent stimulus parameter. Thus, the choice of assigning the linear amplitude variable as the relevant stimulus variable has been for the most part arbitrary. Only through proper examination can one determine which of these dimensions is most suitable for defining and quantifying the response characteristics of a sensory system.

Studies on loudness coding and intensity discrimination support the notion that decibel amplitude is the relevant stimulus variable. Human perception of loudness follows linear relationship with sound pressure level (measured in dB) over most of the hearing range (Stevens 1957 1972) (except at low intensities near the threshold of hearing). Furthermore, intensity discrimination thresholds, ΔI , are constant (about 0.5 dB) throughout most of the hearing range following the well known Weber's law (Miller 1947; Harris 1963; Jesteadt *et al.* 1977; Florentine *et al.* 1987). Similar findings also hold for visual and somatosensory modalities. The fact that response level curves of neuronal

data also follow a simple monotonically increasing function of decibel intensity further supports the idea that decibel quantities are most suitable for describing physiologic data (Palmer and Evans 1979; Evans and Palmer 1980; Ehret and Merzenich 1988; Viemeister 1988; Eggermont 1989). Furthermore, since the spectro-temporal envelope is simply an extension of the intensity variable which extends over time and along the sensory epithelium at fine scales, it is not unprecedented the auditory system process fine spectro-temporal information in a similar manner as it would for intensity. The finding that the spectro-temporal fluctuations of natural sounds extend over a comparable range of differential intensities as the operating range of single auditory neurons (in the order of 30–60 dB; Evans and Palmer 1980; Viemeister 1988) suggest that the operating range of auditory neurons is physically matched to the dynamic range of natural sounds. Aside from the well accepted doctrine that sensory systems utilize their large operating range directly for level coding, our findings further suggest that the large operating range of the auditory system is utilized for efficiently processing spectro-temporal information in acoustic stimuli with a comparable dynamic range.

4.8 Methods

Animal Preparation

Cats were initially anesthetized with a mixture of Ketamine HCL (10 mg/kg) and acepromazine (0.28 mg/kg) which was injected intramuscularly. For the surgical procedure an intravenous infusion line was inserted. A surgical state of anesthesia was induced with ~30 mg/kg Nembutol and maintained throughout with supplements. Body temperature was measured with a rectal probe and maintained with a heating pad at ~

37.5°C. An incision was made in the intercartilaginous area of the trachea and a tracheotomy tube was inserted. After performing a craniotomy, the inferior colliculus was exposed by removing the overlying cerebrum and part of the bony tentorium using a dorsal approach. Upon completion of the surgery, the animal was maintained in an areflexive state of anesthesia via continuous infusion of Ketamine ($1-2 \text{ mg}\cdot\text{kg}^{-1}\cdot\text{h}^{-1}$) and Diazepam ($1-2 \text{ mg}\cdot\text{kg}^{-1}\cdot\text{h}^{-1}$) in lactated ringer solution ($1-2 \text{ ml}\cdot\text{kg}^{-1}\cdot\text{h}^{-1}$). The state of the animal was monitored (heart rate, breathing rate, temperature, and periodically checked reflexes) throughout the experiment and the infusion rate was adjusted according to these physiologic criteria. Every 12 hours the cat received an injection of dexamethasone (0.14 mg/kg s.c.) to prevent brain edema and atropine to reduce salivation (1 mg i.m.). All surgical methods and experimental procedures were approved by the committee on animal research, UCSF.

Neuronal Recording

Data was obtained from single and multi units in the central nucleus of the inferior colliculus (ICC) of four cats. One or two closely spaced parylen coated tungsten microelectrodes ($1-3 \text{ MW}$ at 1kHz) were advanced with a hydraulic microdrive (Kopf). Electrode penetration trajectories were at $\sim 45^\circ$ relative to the sagittal plane and approximately orthogonal to the isofrequency band lamina. Action potential traces were recorded onto a digital audio tape (Cygnus CDAT16) at a sampling rate of 24.0 kHz (41.7 msec resolution) for off line analysis. Off line analysis consisted of digital bandpass filtering ($0.3-10 \text{ kHz}$) all spike trains and individually spike sorting the action potential traces using a Bayesian spike sorting algorithm (Lewicki 1994).

Natural Sound Analysis

A large collection of environmental sounds, human speech, animal vocalizations were obtained from various sources of digitally recorder or remastered media. All sounds were sampled at a sampling rate of 44.1kHz and analyzed over the frequency range of 100 Hz–15 kHz. To characterize spectro–temporal stimulus features we use a detrended spectrographic representation of the stimulus (see chapter 1). We additionally considered a physiologically motivated spectro–temporal stimulus representation which produced qualitatively similar results but is not discussed here (see chapter 1). For all sounds sequences, $x[n]$, we evaluated the discrete time short–time Fourier transform (Cohen 1995; Oppenheim and Schafer 1989)

$$X[n, \omega_k] = \sum_{m=-N}^N x[n+m]w[m]e^{-j\omega_k m}$$

using a B –spline window function (Roark and Escabí), $w[m]$. The variables n designates the discrete time axis and ω_k designates the discrete time frequency variable. The spectrogram was obtained as the magnitude

$$S[n, \omega_k] = \sqrt{X[n, \omega_k] \cdot X[n, \omega_k]^*}$$

of the short–time Fourier transform. Here $X[n, \omega_k]^*$ designates the complex conjugate of the short–time Fourier transform.

A frequency dependent detrending procedure was applied to the data in order to remove $1/f$ energy dependencies observed in natural sounds (see Chapt. 1). The detrended spectrogram, $\bar{S}[n, \omega_k]$, is obtained by applying a pre-emphasis filter (Picone 1997) which magnifies the high frequency components of the signal. This is accomplished by considering a detrending function $S[\omega_k]$ which is used to divide out the spectrogram

$$\bar{S}[n, \omega_k] = \frac{S[n, \omega_k]}{S[\omega_k]} = 1 + \frac{\Delta S[n, \omega_k]}{S[\omega_k]}$$

where the quantity $\Delta S[n, \omega_k] = S[n, \omega_k] - S[\omega_k]$ is the difference spectrogram about the detrending function. The detrending function is obtained by applying a linear fit (in mean square sense) of general form $A\omega_k + B$ to the stimulus mean decibel power spectrum, $20 \log_{10}(E[S[n, \omega_k]])$ (expectation taken with respect to n). On a linear amplitude dimension the detrending function is expressed as

$$S[\omega_k] = 10^{(A\omega_k + B)/20} .$$

Combining terms decibel spectro-temporal envelope is conveniently expressed as

$$\bar{S}_{dB}[n, \omega_k] = 20 \log_{10}(S[n, \omega_k]) - A\omega_k - B .$$

Although linear trends are subtracted from the stimulus decibel spectrogram using this procedure, the detrended stimulus is not white since only a 1st order linear fit was applied to estimate the detrending function. In general strong spectral oscillations are still present. An example is shown in Fig. 1. This descriptor is useful since the perception of intensity differences is ordered on a logarithmic space (Miller 1947; Harris 1963; Jesteadt *et al.* 1977) and since temporal fluctuations of natural sounds are likewise logarithmically distributed (Attias and Schreiner 1998a).

The outlined detrending procedure is applied for several reasons. First note that natural sounds generally have very little energy at high frequencies. On a logarithmic (decibel) plot the power spectrum is usually strongly biased at low frequencies despite the fact that relevant stimulus components are also present at high frequencies. This procedure therefore removes spectral trends which are characteristic of natural sounds. Note that the auditory system effectively performs a similar detrending operation, since frequency tuning and integration bandwidths in the sensory epithelium of the cochlea are logarithmically spaced (Kiang *et al.* 1965; Evans 1972;). Because of this, similar detrending procedures are often employed for speech modeling and in speech recognition systems (Picone 1997). Unlike the spectrogram which depicts absolute energy variations of the stimulus, the defined spectro-temporal envelope depicts relative energy variations along time and frequency. This is not an unreasonable descriptor since it is arguable that relative quantities are far more important for the auditory processing than absolute quantities (for example Weber's law). Note that similar reasoning is also applied to visual processing since visual contrast is likewise defined as a relative quantity (

$$C = (I_{Max} - I_{Min}) / (I_{Max} + I_{Min}) .$$

Acoustic Stimulus

Ripple noise stimuli were designed which are compatible with reverse correlation procedure (see chapter 2). This stimulus has "white noise" like spectro-temporal properties where the range of spectro-temporal envelope fluctuations has been significantly reduced. This reduction in the spectro-temporal content of the stimulus is performed since central auditory neurons in the ICC only respond to a restricted range of spectro-temporal modulations. The ripple noise stimuli is therefore spectro-temporally bandlimited and has a flat spectro-temporal power distribution. The spectro-temporal autocorrelation function has impulse like properties up to the limits set by the envelopes modulation rate, F_m , and ripple frequency, Ω , stimulus parameters. The ripple frequency parameter, Ω , designates the number of resonances per octave along the spectral axis of the stimulus. Likewise, the temporal modulation rate, F_m , designates the number of modulation onsets and/or offsets per unit time along the temporal axis of the stimulus. These parameters were set to a maximum modulation rate, $F_{Max} = 350$ Hz and a maximal ripple frequency $\Omega_{Max} = 4$ cycles/octave since these values encompass roughly 95 % of the neurons in the ICC (Rees and Møller 1983; Møller and Rees 1986; Schreiner and Langner 1988).

A generic spectro-temporal envelope, $S_g(t, X_k)$, was used to construct the sampled acoustic waveforms. Using this generic envelope we constructed five sounds that differ only in their envelope's contrast statistics. Here t designates the discrete time

variable (sampling rate of 44.1 kHz) and $X_k = \log_2(f_k/500)$ designates a sampled octave frequency axis relative to 500 Hz lower limit for our sounds. The spacing, ΔX , between adjacent carrier components corresponded to 0.0231 octaves. Sounds are constructed using the spectro-temporal representation of chapter 2 (Eq. 2.15).

We consider five acoustic pressure waveforms each with distinct contrast distributions: $s_{Lin}(t)$, $s_{15}(t)$, $s_{30}(t)$, $s_{45}(t)$, and $s_{60}(t)$. The subscripts denote the type of contrast statistic. *Lin* designates an acoustic waveform constructed using a spectro-temporal envelope with linearly distributed amplitude statistics whereas numerical values designate the dynamic range for sounds constructed from a spectro-temporal envelope with decibel distributed contrast statistics. The later sound therefore have contrast statistics that resemble those of natural sounds. Since sounds where constructed using an identical generic envelope ($S_g(t, X_k)$) by applying a nonlinear transformation, all sounds sequences therefore have identical spectro-temporal content and differ only in their contrast (amplitude) statistics.

The generic ripple noise envelope has uniformly distributed amplitude statistics in the interval 0 to 1. Decibel distributed sounds where constructed by applying the transformation

$$S_M(t, X_k) = 10^{\frac{M \cdot S_g(t, X_k) - M}{20}}$$

where M designates the modulation depth or, equivalently, the dynamic range of the

envelope in units of dB (M assumes values of 15, 30, 45, or 60 dB). Here $S_M(t, X_k)$ corresponds to the linear amplitude spectro-temporal envelope. Note that the decibel envelope for this sound, $20 \log_{10}(S_M(t, X_k)) = M S_g(t, X_k) - M$, has uniformly distributed contrast statistics in the interval $[-M, 0]$ (see Fig. 5).

A control linearly distributed sound (*Lin*) was also designed since most sounds used in neurophysiologic experiments have linearly distributed contrast statistics. The linearly distributed spectro-temporal envelope for this sounds is designated as (see Fig. 4)

$$S_{Lin}(t, X_k) = \beta \cdot S_g(t, X_k) + (1 - \beta)$$

where the modulation index of $\beta = 1 - 10^{-30/20} = 0.968$ was chosen so that the linearly distributed sound, $S_{Lin}(t, X_k)$, has an identical modulation index as for the 30 dB decibel distributed sound, $S_{30}(t, X_k)$ (i.e. the maximum and minimum amplitude values are identical). These sounds are thus matched at their extremities and only differ in the shape of the contrast distribution. The linear distributed control stimulus has an uniform amplitude distribution on a linear amplitude dimension (skewed distribution on a decibel dimension) whereas the decibel distributed sounds follow an uniform amplitude distribution on a decibel amplitude dimension and thus have skewed amplitude distribution similar to that of natural sounds (Fig. 2) when displayed on a linear amplitude axis.

Although the amplitude distributions for the linearly distributed and the 30 dB

decibel distributed sounds are vastly different (Fig. 3 and 4), their low-order statistics (which are most often used for quantifying contrast) are very similar. The linear distributed sound has a standard deviation of $\sigma_{dB}=6.71$ when computed on a decibel axis and $\sigma_{Lin}=\beta/\sqrt{12}=0.28$ when measured using a linear amplitude axis. The 30dB distributed sounds alternately has standard deviations of $\sigma_{dB}=8.66$ dB and $\sigma_{Lin}=0.23$ when computed in the corresponding amplitude dimension.

As can be seen from Figs. 3 and 4, the amplitude distributions for these two envelopes differ largely in the skewness of the distribution. When plotted on a decibel axis the decibel distributed sound has a symmetric contrast distribution (Fig. 4 E). Likewise the contrast distribution for the linearly distributed sound is also symmetric when it is plotted on a linear amplitude axis (Fig. 3 C). However, when the contrast distributions for these envelopes are constructed in the sound's converse amplitude dimension (e.g. decibel amplitude axis for the *Lin* sound and vice versa) the contrast distributions are highly skewed (Fig. 3 E and 4 C).

Stimulus Presentation

Stimuli were presented binaurally with an independent ripple noise sound sequence for each ear. This allowed us to compute independent *STRFs* for the contralateral and ipsilateral ears. Single neurons and/or clusters of neurons were isolated audio-visually by presenting pure tones and/or white noise. After single/multi units were isolated, a pseudo-random sequence of 15 second ripple noise segments was presented at five intensities and five contrast (*Lin*, 15, 30, 45, or 60 dB). Upon inspection of the

derived contrast versus intensity response curve, an 18 minute segment of the ripple noise stimulus was then presented at key "hot spots" in the contrast intensity response curve: *Lin* and 30dB conditions, *Lin* and 60 dB conditions, or for the *Lin*, 30, and 60 dB conditions. In all instances sounds were presented at 30–70 dB above the neurons response threshold to pure tone. For ~66% of the recording sites, a five–second segment (repeated 140 times) of the dynamic ripple and ripple noise were also played at the end of the recording sessions. All experiments were conducted in an acoustically sealed sound chamber (IAC).

Contrast Intensity Response Function

The contrast–intensity response function was estimated by presenting fifteen second ripple stimulus segments binaurally. The spectro–temporal content for all segments was identical. Sounds only differ in their contrast statistics. Each sound was presented in pseudo–random order at at 5 contrast (*Lin*, 15, 30, 45, and 60 dB) and 5 intensity conditions (25 combinations). Each fifteen second segment was presented four times for a total time of 1 min at each intensity–contrast operating point. The mean firing rate was measured for each condition and the contrast–intensity response function, $R(C,SPL)$, was approximated by a 5 by 5 matrix of mean firing rates. For visualization purposes, the contrast–intensity response matrices were interpolated using the `interp2` function (cubic interpolation) in MATLAB (© Mathworks Inc.). To determine significant differences in firing rate for the different contrast conditions, the mean firing rate of the tested neurons was modeled by Poisson spiking process with counting distributions

$$p(N(T)=n) = \frac{(\hat{\lambda} T)^n \cdot e^{-\hat{\lambda} T}}{n!}$$

where $\hat{\lambda}$ is the measured firing rate taken over a total period of $T=60$ sec. The significance probability was determined numerically by finding the tail probabilities of the overlapping distributions (Zar 1999). A similar procedure was also performed using a bootstrap estimate of the firing rate distributions.

4.9 References

- D.G. Albrecht. Visual cortex neurons in monkey and cat: Effects of contrast and spatial and temporal phase transfer function. *Visual Neurosci.* **12**, 1191–1210 (1995).
- H. Attias, and C.E. Schreiner. Temporal Low Order Statistics of Natural Sounds. *Advances in Neural Information Processing Systems* **10**, 27–33 (1998a).
- H. Attias, and C.E. Schreiner. Coding of Naturalistic Stimuli by Auditory Neurons. *Advances in Neural Information Processing Systems* **10**, 103–109 (1998b).
- H.B. Barlow. Summation and inhibition in the frog retina, *J. Physiol. (London)* **119**, 69–78 (1953).
- H.B. Barlow. The coding of the sensory message, In: *Current problems in animal behavior* (W.H. Thorpe O.L. Zangwill, eds.). Cambridge, MA, MIT Press, 217–234 (1961).
- L. Cohen. *Time Frequency Analysis*. Prentice Hall, New Jersey, 1995.
- G.C. DeAngelis, G.M. Ghose, I. Ohzawa, R.D. Freeman. Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *J. Neurosci.* **19** (10), 4046–4064, 1999.
- D.W. Dong and J.J. Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems* **6** (3), 345–58 (1995).
- Y. Dan, J.J. Atick, and R.C. Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *J. of Neurosci.* **16** (10), 3351–3362 (1996).
- J.J. Eggermont. Coding of free field intensity in the auditory midbrain of the leopard frog. I. Results for tonal stimuli. *Hearing Research* **40**, 147–166 (1989).
- G. Ehret and M.M. Merzenich. Neural Discharge Rate is Unsuitable for Encoding Sound Intensity at the Inferior Colliculus Level. *Hear. Res.* **35**, 1–8, 1988.
- E.F. Evans. The frequency response and other properties of single fibers in the guinea-pig cochlear nerve. *J. Physiol.* **226**, 263–287 (1972).
- E.F. Evans and A.R. Palmer. Relationship Between the Dynamic Range of Cochlear Nerve Fibers and Their Spontaneous Activity. *Experimental Brain Research* **40**, 115–118, 1980.
- J.L. Flanagan. Digital representation of speech signals. BTL symposium on digital

techniques in communication, 1970.

J.L. Flanagan. *Speech Analysis, Synthesis, and Perception*. Springer Verlag, New York, 1972.

M. Florentine, S. Buus, and C.R. Mason. Level discrimination as a function of level for tones from from 0.25–16 kHz. *J. Acoust. Soc. Am.* 81, 1528–1541 (1987).

O. Ghitza and J.L. Goldstein. JNDs for the spectral envelope parameters in natural speech. *Hearing Physiological Bases and Psychophysics* (R. Klinke and R. Hartmann Eds.), Springer Verlag, New York, 352–358 (1983).

J.D. Harris. Loudness discrimination. *J. Speech Hear. Disord. Monographs, Supplement* 11, 1–63 (1963).

P. Heil. Auditory cortical response revisited. I. First spike timing. *J. Neurophysiol.* 77, 2616–2641, 1997a.

P. Heil. Auditory cortical response revisited. II. Response strength. *J. Neurophysiol.* 77, 2642–2660, 1997b.

P. Heil and D.R.F. Irvine. First–spike timing of auditory–nerve fibers and comparison with auditory cortex. *J. Neurophysiol.* 78, 2438–2454, 1997.

W. Jesteadt, C.C. Wier, and D.M. Green. Intensity discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.* 61, 169–177 (1977).

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

G.A. Miller. Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *J. Acoust. Soc. Am.* 191, 609–619 (1947).

A.R. Møller and A. Rees. Dynamic properties of the responses of single neurons in the inferior colliculus of the rat. *Hear. Res.* 24, 203–215, 1986.

I. Nelken, Y. Rotman, and O.B. Yosef. Responses of auditory–cortex neurons to structural features of natural sounds, *Nature* 37, pp. 154–157, 1999.

J.P. Nordmann, R.D. Freeman, and C. Casanova. Contrast sensitivity in Amblyopia: Masking Effects of Noise. *Investigative Ophthalmology and Visual Science* 33, 2975–2985 (1992).

A.V. Oppenheim and R. W. Schaffer. *Discrete Time Signal Processing*. Prentice Hall, New Jersey. 1989.

- A.R. Palmer and E.F. Evans. On the peripheral coding of level of individual frequency components of complex sounds at high sound levels. in *Hearing Mechanisms and Speech*, edited by O. Creutzfeldt, H. Scheich and C.E. Schreiner, Springer Verlag, Berlin, 1979.
- J.W. Picone. Signal Modeling Techniques in Speech Recognition. *Proc. IEEE* **8** (9), 1215–1247 (1997).
- R. Plomp. The Role of Modulations in Hearing. *Hearing Physiological Bases and Psychophysics* (R. Klinke and R. Hartmann Eds.), Springer Verlag, New York, 270–276 (1983).
- R. Plomp. Timbre as a multidimensional attribute of complex tones. *Frequency analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G.F. Smoorenburg, Sijthoff Linden (1970).
- A. Rees and A.R. Møller. Response of neurons in the inferior colliculus of the rat to AM and FM tones. *Hearing Research* **10**, 301–330 (1983).
- D.S. Reich F. Mechler, K.P. Purpura, J.D. Victor. Interspike intervals, receptive fields, and information encoding in primary visual cortex. *J. Neurosci.* **20** (5), 1964–1974, 2000.
- F. Rieke, D.A. Bodnar, and W. Bialek. Naturalistic Stimuli Increase the Rate and Efficiency of Information Transmission by Primary Auditory Fibers. *Proc. R. Soc. Lond.* **262**, 259–265 (1995).
- R.M. Roark and M.A. Escabí. B-spline design of maximally flat and prolate spheroidal-type FIR filters. *IEEE Trans. on Signal Processing* **47** (3), 701–716, 1998.
- D.L. Ruderman. Origins of scaling in natural images. *Vision Research* **37**, n. 23, 3385–3398 (1997).
- D.L. Ruderman and W. Bialek. Statistics of Natural Images: Scaling in the Woods. *Physical Review Letter* **73**, No. 6, 814–817 (1994).
- M.R. Schroder, D. Gottlob, and K.F. Siebrasse. Comparative study of European concert halls: correlation of subjective preferences with geometric and acoustic parameters. *J. Acoust. Soc. Am.* **56**, 1192–1201 (1974).
- C.E. Schreiner and G. Langner. Periodicity coding in the inferior colliculus of the cat. II. Topographic Organization. *J. Neurophysiol.* **60**, 1799–1822, 1988.
- J.W.T. Smolders, A.M.H.J. Aertsten, and P.I.M. Johannesma. Neural representation of

the acoustic biotope: A comparison of the response of auditory neurons to tonal and natural stimuli in the cat. *Biological Cybernetics* **35**, 11–20 (1979).

G.B. Stanley, F.F. Li, and Y. Dan. Reconstruction of natural scenes from ensemble response in the lateral geniculate nucleus. *J. of Neurosci.* **19** (18), 8036–8042 (1999).

S.S. Stevens. On the psychophysical laws, *Psychol. Rev.* **64**, 153–181, 1957.

S.S. Stevens. Perceived level of noise by Mark VII and decibels (E), *J. Acoust. Soc. Am.* **93**, 425–434, 1972.

G. Strang. *Linear algebra and its applications*. Harcourt Brace College Publishers (1988).

T.W. Troy, A.E. Krukowski, N.J. Priebe, and K.D. Miller. Contrast-invariant orientation tuning in cat visual cortex: Thalamocortical input tuning and correlation-based intracortical connectivity. *J. Neurosci.* **18** (15), 5908–5927 (1998);

T.M. Van Veen and T. Houtgast. On the Perception of the Spectral Envelope. *Hearing Physiological Bases and Psychophysics* (R. Klinke and R. Hartmann Eds.), Springer Verlag, New York, 277–281 (1983).

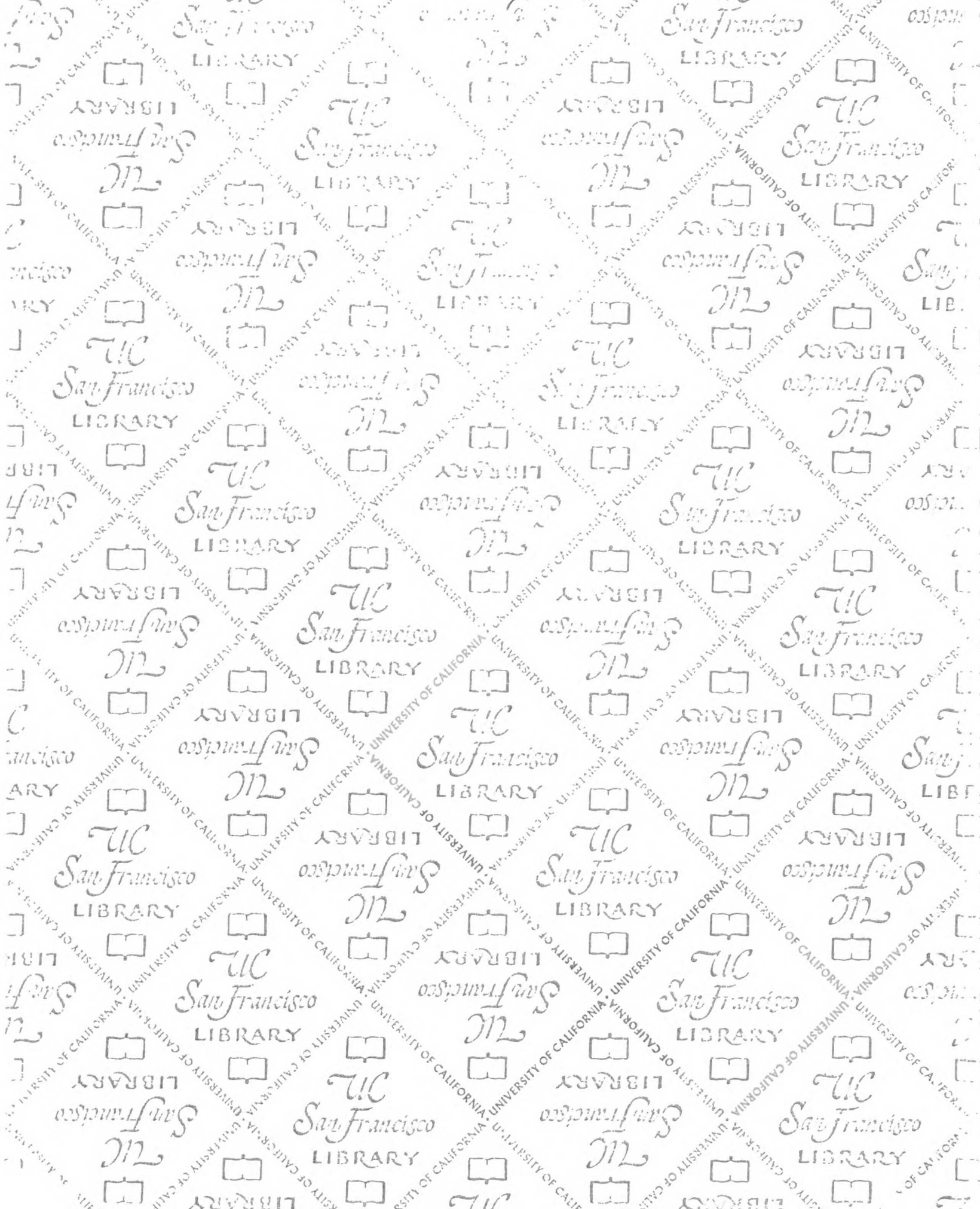
T.M. Van Veen and T. Houtgast. Spectral Sharpness and Vowel Dissimilarity, *J. Acoust. Soc. Am.* **77** (2), 628–634 (1985).

N.F. Viemeister. Intensity coding and the Dynamic Range Problem. *Hearing Research* **34**, 267–274, 1988.

R.V. Voss and J. Clarke. 1/f noise in music: Music from 1/f noise. *J. Acoust. Soc. Am.* **63**, 258–263 (1978).

R.V. Voss and J. Clarke. 1/f noise in music and speech. *Nature* **258**, 317–318 (1975).

J.H. Zar. *Biostatistical Analysis*. Prentice Hall, New Jersey (1999).



For reference

Not to be taken
from the room.

LIBRARY

7063822



3 1378 00706 3822

1977
1977