UCONN LIBRARY

University of Connecticut
OpenCommons@UConn

Doctoral Dissertations | University of Connecticut Graduate School

7-31-2017

Neural Coding and Models for Natural Sounds Recognition: Effects of Temporal and Spectral Features

Seyedeh Fatemeh Khatami Firoozabadi
University of Connecticut - Storrs, khatami@engr.uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
Khatami Firoozabadi, Seyedeh Fatemeh, "Neural Coding and Models for Natural Sounds Recognition: Effects of Temporal and Spectral Features" (2017). *Doctoral Dissertations*. 1520.
https://opencommons.uconn.edu/dissertations/1520

Neural Coding and Models for Natural Sounds Recognition: Effects of Temporal and Spectral Features

Seyedeh Fatemeh Khatami Firoozabadi, PhD

University of Connecticut, 2017

The mammalian brain is able to recognize natural sounds in the presence of acoustic uncertainties such as background noise. A prevailing theory of neural coding suggest that neural systems are optimized for natural environment signals and sensory inputs that are biologically relevant. The optimal coding hypothesis thus suggests that neural populations should encode sensory information so as to maximize efficient utilization of environmental inputs. In the first part of my thesis, I will explore the origins of scale invariance phenomena which has been previously described for natural sounds and has been observed in a variety of natural sensory signals including natural scenes. In the second part, I will explore the ability of the brain to utilize high-level statistical regularities in natural sounds to perform sound identification tasks. Using a catalog of natural sounds, texture synthesis procedures to manipulate sounds statistics from various sound categories, and neural recordings from the auditory midbrain of awake rabbits, I will show that neural population response statistics can be used to identify discrete sound categories. In the last part of the thesis, I will explore the role of hierarchical organization in the auditory pathway for sound recognition and optimal coding in the presence of challenging background noise. Using neural responses from auditory nerve, midbrain, and auditory cortex, I developed optimal computational neural network model for word recognition in presence of speech babble noise. I demonstrate that the optimal computational strategy for word recognition in noise predicts various transformations performed by the ascending auditory pathway, including a sequential loss of temporal and spectral resolution, increasing sparseness and selectivity.

Neural Coding and Models for Natural Sounds Recognition: Effects of Temporal and Spectral Features

Seyedeh Fatemeh Khatami Firoozabadi

B.S., Azad University of Tehran (Central Branch), 2006

M.S., Iran University of Science and Technology, 2011

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2017

Copyright by

Seyedeh Fatemeh Khatami Firoozabadi

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

Neural Coding and Models for Natural Sounds Recognition: Effects of Temporal and Spectral Features

Presented by

Seyedeh Fatemeh Khatami Firoozabadi, B.S., M.S.

Major Advisor _____
Monty A. Escabi

Major Co Advisor _____
Heather L. Read

Associate Advisor _____
Ian Stevenson

Associate Advisor _____
Sabato Santaniello

University of Connecticut
2017

Dedication

My deep gratitude goes first to my advisor, Dr. Monty Escabi who expertly guided me thorough my graduate education. His enthusiasm about the science make me to have a joyful time during my research.

My appreciation extends to my Major Co advisor, Dr. Heather Read, whose mentoring and guidance during this 4 years are very valuable.

Also, I want to thank Dr. Ian Stevenson and Dr. Sabato Santaniello who were guided me as committee members to improve quality of my work.

Above ground I am indebted to my family, specially my Mom and Dad, Rezvan and Saeed, whose value to me only grows with age. Finally, I acknowledge my husband, Hamed, who is my champion and who blessed me with the light of joy when the lab lights were off.

Contents

APPROVAL PAGE iii
Dedication iv
Chapter 1: Introduction 1
1.1 Optimal Coding 2
1.2 Spectral and Temporal Modulation 4
1.3 Auditory System 5
1.4 Chapters 7
Chapter 2: The role of temporal cues in 1/f scaling phenomenon 11
2.1 Introduction 12
2.2 Material and Methods 13
2.2.1 Sound Analysis 14
2.2.2 Analytical Model Based on Modulation Power Spectrum 15
2.2.4 Analytical Cutoff Frequency 18
2.3. Results 19
2.3.1 Model Fitting 19
2.3.2 Effect of Acoustic Features 21
2.3.3 Effect of Distribution of Vocalization Duration 22
2.3.4 Statistical Structure of Model Parameters 23
2.3.5 Analytical vs. Experimental Model 25
2.3.6. Conceptual Model 29
2.4. Discussion 30

2.5. Funding Source 33
Chapter 3: The Role of Statistical Regularities for Sound Category Identification 34
3.1 Introduction 35
3.2 Materials and Methods 35
3.2.1 Sound Dataset 35
3.2.2. Data Recording 36
3.2.3. Stimuli dataset analysis and statistics selection 39
3.2.4 Stimuli Sound Synthesis 43
3.2.5 Frequency Response Area (FRA) 44
3.2.6 Shuffled Autocorrelogram 46
3.3 Results 47
3.3.1 Neural Correlation Statistics and Classifier 47
3.3.2 Effect of Sound Duration and Neural Population Size on Classification Results 50
3.4 Discussion 51
3.5 Funding sources 52
Chapter 4: The Role of Auditory System Hierarchy for Optimal Coding and Sound Recognition 53
4.1 Introduction 54
4.2 Materials and Methods 56
4.2.1 Speech Corpus 56
4.2.2 Auditory System Data 56
4.2.3 Auditory Model and Hierarchical spiking neural Network (HSNN) 57
4.2.4 Decision Model 60
4.2.5 Network Constraints and Optimization 62
4.2.6 Receptive Field and Mutual Information Calculation 64
4.2.7 Generalized Linear Model (GLM) Networks 64
4.3 Results 66
4.3.1 Hierarchical spiking neural network for Word Recognition 66
4.3.3 Hierarchical Organization of Auditory System versus Optimal HSNN 71
4.3.5 Acoustic Transformation Between Consecutive Network Layers (Optimal vs. High Resolution) 78
4.3.6 Human Performance and Network Hierarchy and Nonlinearity Properties 81
4.3.7 Optimal Spiking Timing Resolution 84
4.4 Discussion 85

4.5 Funding Sources..... 89
References: 90

Chapter 1: Introduction

Hearing ability is essential for everyday survival among all animals and humans. Understanding how the auditory system works by developing models can help us to understand how different parts of the auditory system collaborate with each other. Furthermore, such models have the potential to be used as new artificial systems for sound recognition that can optimally process and recognize sounds using biological principles.

To model a biological system, such as the brain, it is critical to first understand the underlying anatomy and physiology and their relationship to the overall system function. To have an optimal coding system, it is essential to know the input and output properties of the system which the system is designed to transfer optimal information of them. The input, output, and functionality of the biological system tightly depends on each other. Auditory system is the complicated system, which developed to process natural sound, such as vocalization, environmental sounds, and spoken words. The necessity of optimal coding in different stages of

auditory system from auditory nerve to auditory cortex, is an interesting topic which sounds may process in such a way that most optimal information reaches a cortex for further processing, such as decision or memorizing. Specifically, the optimal coding is critical at the level of inferior colliculus in the auditory system, and it has been shown that cochlea is optimized for preserving sound waveform information (Lewicki 2002). Thus, to transfer maximum information through the auditory pathway, the optimal coding seems necessary.

The optimal coding hypothesis was introduced by Horace Barlow in 1961 as a model of sensory coding in the brain (Barlow 1961). Neurons in different part of the brain, communicate with each other by using electrical impulses, i.e. action potentials or spikes. It is important to find out the meaning of these spikes to understand how the brain processes information about the outside world. Barlow hypothesized that the spikes in sensory systems efficiently represent sensory information. That is, neurons ought to encode sensory information in a manner that maximizes the stimulus information content while at the same time minimizing neural circuitry and metabolic resources. Based on Barlow's model, the brain should employ a code that is specifically adapted for representing visual and audio information representative of an organism's natural environment.

1.1 Optimal Coding

In 1961 Barlow proposed optimal coding hypothesis which states that the brain encode sensory information in an optimal manner. Following ideas from evolutionary biology (Edelman 1987) and Information theory (Shannon 1958) Barlow hypothesized that the brain evolved to optimally take advantage of environmental stimuli in order to enable species survival. Based on this model the brain is optimized to code environmental visual and auditory information that is prevalent in natural environments. Optimal coding theories also suggest that networks of neurons should encode information in such a way to efficiently encode environmental input, meaning that stimulus

information conveyed by different neurons should be minimally redundant (Simoncelli 2003, Ulanovsky, Las et al. 2003). Such a code would guarantee that neurons work largely independently. Such a scheme can also potentially minimize the number of neurons and the required neural circuitry.

The optimal coding hypothesis states that adaptation of sensory processing in the brain to natural stimuli is necessary. In other words, neurons in auditory or visual system should be optimized for sounds or images represented in the environment. One example of applying this theory in auditory systems leads to filters which is similar to a cochlear filter (Lewicki 2002, Smith, Lewicki 2006, Rodriguez, Chen et al. 2010). According to optimal coding theory, for better understanding of brain functionality, it is important to understand the nature of environmental signals.

One of the simplest statistical feature of environmental sounds is temporal cues of sound statistics. 1/f scaling is one of the universal phenomenon which has been observed in a variety of natural sensory signals. The power spectrum of natural visual scenes, for instance, exhibits scale invariance and drops off as a power law of the spatial frequency (Ruderman, Bialek 1994, Field 1987). Although this phenomenon has been observed for many years, the origin of this phenomenon and sounds' statistics parameters which contribute to this phenomenon is still unknown. To understand these parameters we developed models based on temporal features of input sounds.

Spectral cues are also critical for identifying and discriminating natural sounds. Recently, McDermott and Simoncelli demonstrated that high-order statistics related to the spectral structure of sounds are used by human listeners during sound recognition (McDermott, Simoncelli 2011a). Currently, its poorly understood which and if how such sound statistics are encoded and utilized

by neurons in different region of auditory system. In my thesis, I looked at high order statistic information which are transferred to the inferior colliculus, and investigate if we can use this information for discrimination environmental sounds.

After looking at the role of some inputs and some output features in optimal and efficiency coding, it is worthwhile to look at role of hierarchical organization of auditory system to investigate and get a sense how optimal coding works along with auditory pathway. In last part of the thesis, I designed a hierarchical mathematical model with spiking neurons which is optimized for word recognition. The first layer of this network is designed based on some biological parameters which have been seen in auditory nerve, and by using inhibitory and exhibitory connection weights which connect consecutive layers, we designed a biologically inspired network.

1.2 Spectral and Temporal Modulation

According to optimal coding theory, it is first essential to understand the statistical characteristics of the environmental signals. Natural auditory signals contain temporal and spectral features. Carrier frequency (frequency modulation, faster changes) and amplitude envelope (slower structures) are two main components of temporal structure of audio signals. Both of them carry perceptual cues to the listeners.

For hearing and analyzing sounds, the auditory system is able to detect frequency properties or spectral shape of stimulus which includes carrier frequency range (tones (Hz)), carrier bandwidth, and modulation frequencies, as well as temporal changes such as amplitude, onset and offset of stimulus. On one hand, rate modulation transfer functions (rMTF) capture neural response magnitude across modulation frequencies and on the other hand temporal modulation transfer functions (tMTF) have information about synchronization.

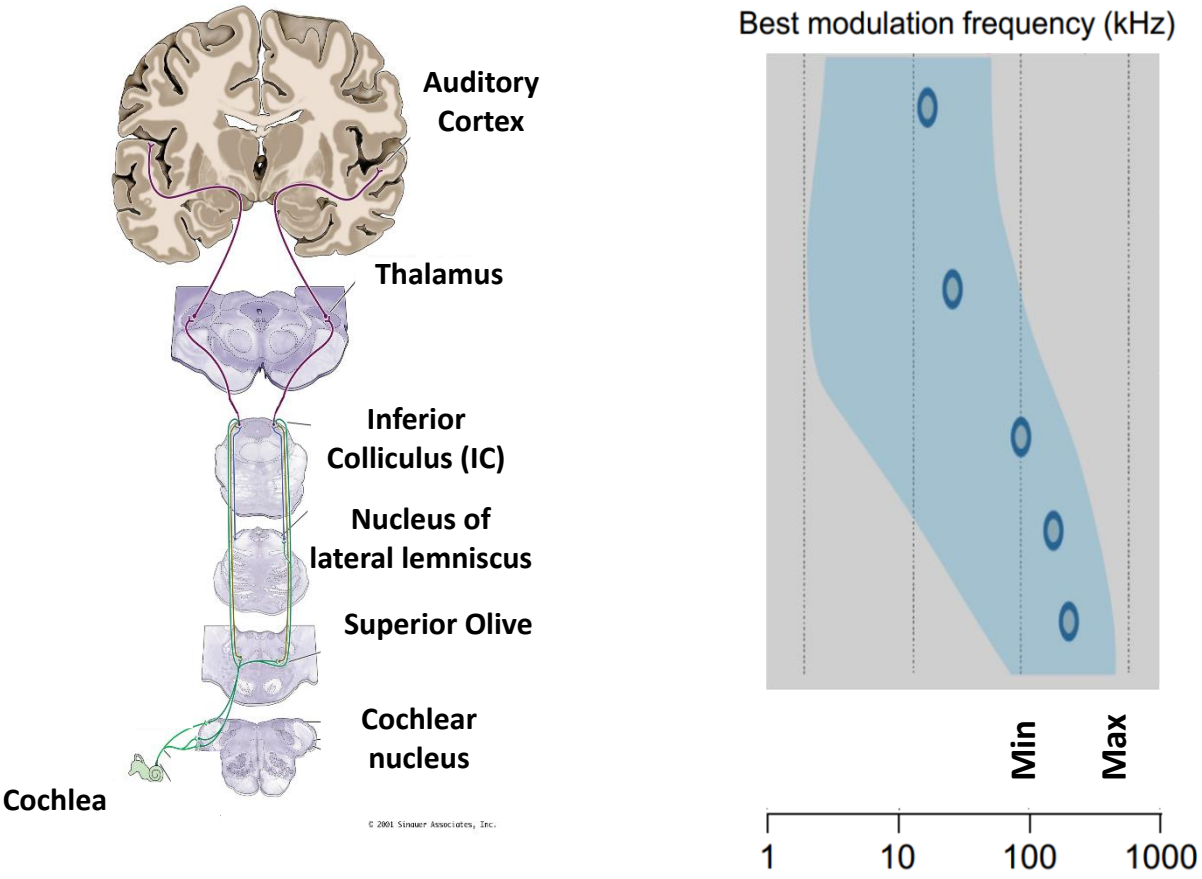
It has been hypothesized that sensory pathway has been adapted to have an optimal coding of the statistics of environment (Schwartz, Simoncelli 2001, McDermott, Schemitsch et al. 2013). Neurons in auditory pathway for different animals are adapting to the structural regulatory (Escabi, Miller et al. 2003, Rodriguez, Chen et al. 2010, Ter-Mikaelian, Semple et al. 2013) and it has been shown that environmental statistics improve discrimination of natural sounds (Woolley, Fremouw et al. 2005). In general, it has been seen that not only the temporal and spectral modulation are important for sound recognition, but also statistics of these properties can increase discrimination ability.

1.3 Auditory System

The lemniscal auditory pathway consists of sequentially connected auditory nuclei that are ultimately responsible for encoding and perceptual phenomenon. Mammalian audition is resilient to acoustic variability, such as background noise and multiple talkers, yet how the brain accomplishes this seemingly simple feat is unknown. One hypothesis is that the auditory pathway is organized into hierarchical processing stages with sequentially changing feature extraction capabilities that culminate in an invariant noise robust representation.

Several hierarchical changes in spectral and temporal selectivity have been documented and are consistently observed in the ascending auditory system of mammals. Temporal selectivity and resolution change dramatically over more than an order of magnitude, from a high-resolution representation in the cochlea, where auditory nerve fibers synchronize to temporal features of up to ~1000 Hz, to progressively slower (limited to ~25 Hz) and coarser resolution representation as observed in auditory cortex (Joris, Schreiner et al. 2004). Although changes in spectral selectivity can be described across different stages of the auditory pathway, and temporal resolution is

somewhat coarser in central levels, changes in frequency resolution are somewhat more homogeneous and less dramatic (Rodriguez, Chen et al. 2010, Miller, Escabi et al. 2002, Mc Laughlin, Van de Sande et al. 2007). It's plausible that such hierarchical transforms across auditory nuclei are essential for feature extraction and ultimately high-level auditory tasks such as acoustic object recognition.



(a) (b)

Figure 1. Auditory pathway stages and functional properties. (a) different stages of auditory pathway. (b) Modulation sensitivity in different stages of auditory pathway (Joris, Schreiner et al. 2004, von Trapp, Buran et al. 2016, Malone, Schreiner 2010).

Measures of neural coding in auditory nerve represents both envelope and fine structures

(elements with high frequency). It has been observed that rate modulation transfer function is flat across modulation frequencies, while temporal modulation is finely tuned and has a cut off frequency around 1-2kHz, Figure 1.b (Joris, Schreiner et al. 2004, Malone, Schreiner 2010). In other words, neurons in auditory nerve are more sensitive to high frequency components. When we are moving from the auditory nerve to the auditory cortex neurons become more sensitive to lower frequency components, and it seems modulation frequency and rate properties change systematically across the auditory pathway. In the next stage, the inferior colliculus, best modulation frequency decreased to 100Hz, and finally at the auditory cortex best modulation frequency is in range of 7-15Hz, which it means neurons in auditory cortex are more sensitive to lower component frequencies (Joris, Schreiner et al. 2004, Malone, Schreiner 2010). Similar systematic trends have been observed for spike rate. Rate of modulation transfer function has been shown to gradually decrease from auditory nerve to auditory cortex. Although these transformations have been observed in auditory pathway in the past, the goal of these sequential transformation and its role in optimal computational strategy is still unclear.

1.4 Chapters

In the following chapters, various topics, discussed above, will be explained in more detail by designing different models based on previously data collection from cat and by collecting data from awake rabbit. I generate new model for discrimination of natural sounds based on their statistical information. In next chapters I will explore optimal coding theory by both investigating different natural sounds, investigating how sound statistics contribute to discrimination ability in inferior colliculus, and finally by designing a biologically inspired computational auditory pathway model try to find out how optimal coding cause almost stable recognition in presence of challenging noise.

In Chapter 2, I explore the origins of scale invariance phenomena in the envelope of natural sounds. Scale invariance, whereby the power of a signal is inversely proportional to the signal frequency, is likely a universal phenomenon since it has been previously observed in a variety of natural sensory signals. The power spectrum of natural visual scenes, for instance, exhibits scale invariance where the luminance power drops off as a power law of the spatial frequency (Ruderman, Bialek 1994, Field 1987). Although the mere presence of spatial edges contributes and is partly responsible for 1/f scaling phenomenon, detailed examination of image statistics suggests that high-order statistical regularities in the spatial arrangement of edge boundaries are necessary to explain the observed 1/f scaling (Zylberberg, Pfau et al. 2012, Ruderman 1997). Neurons in the central visual system appear to be optimized for visual edge detection (Hubel, Wiesel 1959, Field 1987), consistent with the hypothesis that brain may be specialized for statistical regularities prevalent in natural environments. Like for visual scenes, the envelope of natural sounds also exhibits scale invariance (Rodriguez, Chen et al. 2010, Geffen, Gervain et al. 2011) and neurons in the auditory midbrain are believed to encode such acoustic properties in an optimal fashion (Rodriguez, Chen et al. 2010). However, the physical features that contribute to this phenomenon are unknown. In this chapter, I will explore sounds physical cues that cause this phenomenon.

In Chapter 3, I will explore the hypothesis that auditory pathway neurons respond reliably and that pair-wise neuron-to-neuron correlated activity is modulated by to higher-order statistical regularities in natural sounds. Further, I hypothesize that neural populations can utilize the response properties for sound discrimination and categorization. It is known that neural responses in the central auditory system can be modulated by a variety of high-order sound statistics, including modulation and correlation statistics (Escabi, Miller et al. 2003, Attias, Schreiner 1998,

Lesica, Grothe 2008) . I am going to test the hypothesis that neural responses in the inferior colliculus are modulated by high-order statistical regularities in sounds and that statistical patters of the neural activity can ultimately be used to discriminate and categorize sounds. To test the above hypothesis, I will perform neural recordings in inferior colliculus of unanesthetized rabbits in response to an ensemble of synthetic texture sounds. Texture sounds, are stochastic sounds that can be modeled by high-order statistical regularities that can be selectively perturbed (McDermott, Simoncelli 2011a). Many natural sounds, such as rain, fire, and water stream are considered as texture sounds and I will use some of these along with perturbed variants in my study to determine how sound statistics contribute to neural coding in the auditory midbrain.

In Chapter 4, I explore the role of auditory system hierarchy and its relationship to optimal coding of speech sounds in background noise. Being able to identify sounds in the presence of background noise is essential for everyday audition and vital for survival. Although current knowledge of auditory physiology is comparatively advanced, the neural transformations responsible for the robustness of the mammalian auditory pathway remain poorly understood. Furthermore, although several cortical mechanisms have been proposed to facilitate robust coding of sounds (Mesgarani, David et al. 2014, Schneider, Woolley 2013), it is still unclear how the sequential organization and resulting transformations performed by the auditory pathway contribute to robust sound recognition. In Chapter 4, I will investigate network mechanisms that contribute to robust coding of sounds in the presence of competing variability, such as background noise and multiple talkers. Although previous studies identified mechanisms that contribute to robustness at the single cell level (Mesgarani, David et al. 2014, Schneider, Woolley 2013), there are no prior studies that have evaluated how and if the auditory pathway's ascending organization contributes to robust coding. Specifically, I will test the hypothesis that the auditory pathway is

organized into hierarchical processing stages with sequentially changing feature extraction capabilities (spectro temporal features) that culminate in a noise robust representation.

Chapter 2: The role of temporal cues in 1/f scaling phenomenon

2.1 Introduction

1/f scaling is likely a universal phenomenon since it has also been observed in a variety of natural sensory signals. The power spectrum of natural visual scenes, for instance, exhibits scale invariance and drops off as a power law of the spatial frequency (Ruderman, Bialek 1994, Field 1987). Although the mere presence of spatial edges contributes and is partly responsible for 1/f scaling phenomenon, detailed examination of image statistics suggests that high-order statistical regularities in the spatial arrangement of edge boundaries are necessary to explain the observed 1/f scaling (Zylberberg, Pfau et al. 2012, Ruderman 1997). Neurons in the central visual system appear to be optimized for visual edge detection (Hubel, Wiesel 1959, Field 1987), consistent with the hypothesis that brain may be specialized for statistical regularities prevalent in natural environments. Moreover, the amplitude fluctuations of natural sounds exhibit scale invariance or 1/f scaling, such that the power of temporal amplitude fluctuations are inversely related to the modulation frequency. This phenomenon was discovered for natural sounds over 40 years ago (Geffen, Gervain et al. 2011, Voss, Clarke 1978), however, the origins and implications of this universal phenomenon remain unknown.

As for the visual scenes, I hypothesize that temporal edges may contribute to 1/f scaling for sounds. Temporal amplitude fluctuations of vastly different sounds including speech, animal vocalizations, environmental sounds, and music all exhibit 1/f scaling (Attias, Schreinert 1997a, Voss, Clarke 1978, Rodriguez, Chen et al. 2010, Singh, Theunissen 2003, Geffen, Gervain et al. 2011). Upon decomposing a natural sound into a carrier component, which accounts for the fine structure or frequency content of the sound, and a temporal envelope $x\,(t)$, which accounts for the temporal amplitude fluctuations (Cohen 1995a), the modulation power spectrum (MPS) is defined as the Fourier transform magnitude of the envelope signal and represents the power in the temporal

fluctuations as a function of modulation frequency (Houtgast, Steeneken 1985). For many natural sounds the MPS is well described by a power law function of the form $S_{xx}(f) \propto f^{-\alpha}$, such that the power in the envelope signal drops off with increasing frequency with exponent of $\alpha \approx 2$. With the exception of water sounds, where 1/f scaling can be accounted for by the distribution of self-similar acoustic "droplets" (Geffen, Gervain et al. 2011), the physical acoustic features that contribute to 1/f scaling across broader categories of natural and man-made sounds remain unknown. Furthermore, it remains a mystery as to whether there are universal features that determine the 1/f MPS in natural sounds. Answering this question has an important implication as neurons in the mammalian auditory system efficiently encode 1/f structure in the sound envelope (Rodriguez, Chen et al. 2010) suggesting it is a critical driver of brain pathway function and perception abilities.

In Chapter 2, I will evaluate the temporal cues that contribute to 1/f scaling phenomenon for natural sounds. I will parameterize multiple acoustic features in vocalization of several species and speech sequences to demonstrate how that temporal acoustic features accounts for the 1/f scaling phenomenon.

2.2 Material and Methods

Physically, vocalization production in many species entails a source generator (e.g., vocal folds) that produces quasi-periodic envelope signal and articulatory gestures such as the opening and closing of the mouth and postural adjustments of the lips and tongue that dynamically shape the sound envelope. Envelope fluctuations created by vocal fold vibration typically lie outside the modulation frequency range where 1/f scaling is observed (Rodriguez, Chen et al. 2010) (i.e., >100 Hz) and thus should not contribute to 1/f scaling directly. However, ubiquitous to all vocalization

sequences across species is the presence of abrupt temporal onsets and offsets that mark the beginning and end of individual isolated vocalizations (Fig. 1a, b, c, d). In human speech, for instance, these pronounced features are generated by the time-dependent opening and closing of the oral cavity and related articulatory gestures. Here we tested the hypothesis that the resulting time-dependent temporal edges account for 1/f scaling in vocalization sounds.

2.2.1 Sound Analysis

Sound Database: Sequences of vocalized sounds were obtained from a variety of sources and species, including rat pup (Wöhr, Schwarting 2008), mouse pup (Wöhr, Dahlhoff et al. 2008), bird (Bradbury, Budney), primate (Hice 2000), crying infant (Green, Gustafson et al. 1998), and speech (Branagh, Dearman). Sound sequences varied in length between five to nine minutes and were sampled at a sampling rate to preserve the frequency content of each species (see Table 2.1).

Envelope Extraction and Modulation Power Spectrum: For each vocalization sequence, we compute the modulation power spectrum by extracting the temporal envelope of each sound sequence. Sounds were first bandpass filtered for frequencies f_{low} and f_{high} in order to isolate the

Table 2.1: Parameters used for envelope extraction and model fitting.

	Sound Duration (min)	**Fs, Sampling Rate (kHz)**	**Envelope down sample factor**	**Envelope Threshold**	f_{low} **(kHz)**	f_{high} **(kHz)**	f_m **(Hz)**
Rat pup	6	250	100	100	30	100	30
Mouse pup	5	300	100	100	30	100	30
Bird	5	44.1	10	10	0.5	20	10
Primate	9.66	44.1	10	10	0.5	20	10
Infant	7	44.1	10	5	0.5	20	10
Speech	5	44.1	10	5	0.5	20	10

dominant vocalization component:

$$s_{band}(t) = s(t) * h_{band}(t)$$

where $h_{band}(t)$ is the bandpass filter impulse response and * is the convolution operator. Since the vocalizations for each species has dominant energy over a unique frequency range, the frequencies f_{low} and f_{high} were individually selected based on visual inspection of the sound spectrum (Table 2.1). The temporal envelope for the chosen frequency band ($x(t)$) was obtained by computed the Hilbert transform to compute the analytic signal magnitude:

$$x(t) = |s_{band}(t) + iH\{s_{band}(t)\}|.$$

Each envelope was normalized for unit standard deviation and the modulation power spectrum of each species was obtained by computing the power spectral density of $x(t)$ using a Kaiser ($\beta = 3.39$) windowed Welch averaged Periodigram (50% overlap; nominal frequency resolution of 0.02 Hz).

2.2.2 Analytical Model Based on Modulation Power Spectrum

The modulation [i]power spectrum of the vocalization model is obtained by taking the long-term expectation of the Fourier Transform Magnitude:

$$S_{xx}(f) = \lim_{T\to\infty} \frac{1}{T} E[X(f)X(f)^*]$$

where $E[\cdot]$ is the expectation operator taken across the three random variables (onset time, duration and amplitude) and

$$X(f) = \Im\{x(t)\} = \Im\left\{\sum_{n=1}^{N} A_n \cdot rect\left(\frac{t-t_n}{D_n}\right)\right\} = \sum_{n=1}^{N} A_n \cdot \frac{\sin(\pi D_n f)}{\pi f} e^{-j2\pi f t_n}$$

is the envelope Fourier transform ($\Im\{\cdot\}$). The model MPS is then obtained as

$$S_{xx}(f) = \lim_{T\to\infty} \frac{1}{T} E[X(f)X(f)^*]$$

$$= \lim_{T\to\infty} \frac{1}{T} E\left[\left(\sum_{n=1}^{N} A_n \cdot \frac{\sin(\pi D_n f)}{\pi f} e^{-j2\pi f t_n}\right)\left(\sum_{k=1}^{N} A_k \cdot \frac{\sin(\pi D_k f)}{\pi f} e^{+j2\pi f t_k}\right)\right]$$

$$= \lim_{T\to\infty} \frac{1}{T} E\left[\sum_{n=1}^{N} A_n^2 \cdot \frac{\sin^2(\pi D_n f)}{\pi^2 f^2} + \sum_{n=1}^{N}\sum_{k\neq n} A_k \cdot A_n \cdot \frac{\sin(\pi D_n f)}{\pi f}\frac{\sin(\pi D_k f)}{\pi f} e^{+j2\pi f(t_k - t_n)}\right].$$

Table 2.2: Estimated model parameters for each vocalization sequence. Mean and standard deviation values are provided for the vocalization amplitude (μ_A, σ_A) (normalized amplitude), duration (μ_D, σ_D) and inter-vocalization interval (μ_I, σ_I). λ is the vocalization rate (units of vocalizations/sec).

	μ_A	σ_A	$\mu_D(s)$	$\sigma_D(s)$	$\mu_I(s)$	σ_I(s)	λ (Hz)
Rat pup	4.45	5.72	0.05	0.03	2.12	2.96	0.47
Mouse pup	1.65	5.49	0.02	0.008	0.74	1.82	1.34
Bird	1.79	2.26	0.41	1.80	0.63	2.02	1.57
Primate	3.21	1.64	1.13	3.14	3.23	5.27	0.31
Infant	2.44	1.85	0.28	0.20	0.96	0.77	1.03
Speech	3.13	1.63	0.29	0.30	0.70	0.53	1.41

Given that the model parameters are largely independent and onset times are serially uncorrelated (Fig. 2.2; Table 2.3), we assume independence so that the second term inside the expectation approaches zero so that

$$S_{xx}(f) = \lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{N} E\left[A_n^2 \cdot \frac{\sin^2(\pi D_n f)}{\pi^2 f^2}\right] = \lim_{T\to\infty} \frac{N}{T} E\left[A_n^2 \cdot \frac{\sin^2(\pi D_n f)}{\pi^2 f^2}\right].$$

Since in the limiting case $\lambda \simeq N/T$ and the random variables are approximately independent the MPS simplifies as follows

$$S_{xx}(f) = \lambda \cdot E[A_n^2] \cdot E\left[\frac{\sin^2(\pi D_n f)}{\pi^2 f^2}\right] = \lambda \cdot (\mu_A^2 + \sigma_A^2) \cdot E\left[\frac{\sin^2(\pi D_n f)}{\pi^2 f^2}\right]$$

Finally, under the assumption that the vocalization durations are uniformly distributed with in the interval $[T_1,T_2]$

$$E[sin^2(\pi D_n f)] = \int p(\gamma)\, sin^2(\pi\gamma f) d\gamma$$

$$= \frac{1}{T_2 - T_1} \int_{T_1}^{T_2} sin^2(\pi\gamma f) d\gamma = \frac{1}{(T_2 - T_1) \cdot 2} \int_{T_1}^{T_2} 1 - \cos(2\pi\gamma f) d\gamma$$

$$= \frac{1}{2} \cdot \frac{1}{T_2 - T_1} \cdot \left[T_2 - T_1 - \frac{\sin(2\pi T_2 f) - \sin(2\pi T_1 f)}{2\pi f}\right]$$

so that MPS is

$$S_{xx}(f) = \frac{\lambda \cdot (\sigma_A^2 + \mu_A^2)}{2 \cdot \pi^2 f^2} \cdot \left[1 - \frac{\sin(2\pi T_2 f) - \sin(2\pi T_1 f)}{(T_2 - T_1) \cdot 2\pi f}\right]$$

$$= \frac{\lambda \cdot (\sigma_A^2 + \mu_A^2)}{2 \cdot \pi^2 f^2} \cdot \left[1 - \frac{T_2}{T_2 - T_1} \cdot \mathrm{sinc}(2\pi T_2 f) + \frac{T_1}{T_2 - T_1} \cdot \mathrm{sinc}(2\pi T_1 f)\right].$$

2.2.4 Analytical Cutoff Frequency

The vocalization model MPS cutoff frequency (f_c) is defined as the frequency where MPS achieves half power (- 3dB) relative to the MPS at zero frequency

$$S_{xx}(f_c) = \frac{1}{2} \cdot S_{xx}(0),$$

which for the model requires that the following equation be satisfied

$$\frac{\lambda \cdot (\sigma_A^2 + \mu_A^2)}{2 \cdot \pi^2 f_c^2} \cdot \left[1 - \frac{\sin(2\pi T_2 f_c) - \sin(2\pi T_1 f_c)}{(T_2 - T_1) \cdot 2\pi f_c}\right] = \frac{\lambda \cdot (\sigma_A^2 + \mu_A^2)}{6(T_2 - T_1)} [T_2^3 - T_1^3].$$

An approximate solution is obtained by noting that for large $f_c > 1/2\pi(T_2 - T_1)$

$$\frac{\sin(2\pi T_2 f_c) - \sin(2\pi T_1 f_c)}{(T_2 - T_1) \cdot 2\pi f_c} < \frac{1}{(T_2 - T_1) \cdot 2\pi f_c} < 1.$$

Considering this upper bound, the above equation is approximated as

$$\frac{1}{2 \cdot \pi^2 f_c^2} \approx \frac{1}{6(T_2 - T_1)} [T_2^3 - T_1^3]$$

and solving for the cutoff frequency yields

$$f_c \approx \frac{1}{\pi}\sqrt{\frac{3\cdot(T_2-T_1)}{[T_2^3-T_1^3]}}.$$

Finally, since $\mu_D = (T_1+T_2)/2$ and $\sigma_D^2 = (T_2-T_1)^2/12$ for a uniform distribution the cutoff can be expressed as

$$f_c \approx \frac{1}{\pi}\frac{1}{\sqrt{\mu_D^2+\sigma_D^2}} = \frac{1}{\pi}\frac{1}{\sqrt{E[D_n^2]}}.$$

2.3. Results

2.3.1 Model Fitting

To test the hypothesis that temporal acoustic edges are responsible for 1/f modulation spectrum of vocalized sounds we consider a simplified model of vocalization envelopes, x(t), where temporal edges are used to model the beginning and end of isolated vocalizations. The envelope of a vocalization sequence can be approximated as a superposition of rectangular pulses, $p_n(t)$, each which marks the beginning and end of each isolated vocalization

$$x(t) = \sum_{n=1}^{N} p_n(t) = \sum_{n=1}^{N} A_n \cdot \mathrm{rect}\left(\frac{t-t_n}{D_n}\right) \quad (2.1)$$

as seen in Fig 2.1, where n is the pulse number and $rect(\cdot)$ is a unit amplitude rectangular pulse with start time zero and duration of 1 s. The number of isolated vocalizations within the T second interval is $N \simeq \lambda T$ where λ is the average vocalization rate in units of vocalizations/sec. To account for the vocalization-to-vocalization variability in the sequence, pulse amplitudes (A_n), onset times (t_n) and durations (D_n) are modeled as independent random variables.

The envelopes from each vocalization sequences were fitted to the model of Eq. 2.1 to assess how temporal sequence parameters (vocalization peak amplitudes, durations and onset times) contribute to 1/f structure. The fitting procedure consisted of 1) identifying isolated vocalizations in the sequences that stood out above the background noise level and 2) fitting the isolated vocalization to rectangular pulses.

In the first iteration of the fitting procedure, I detected isolated vocalizations that exceeded a baseline level above the recording noise. Since isolated vocalizations occur at relatively low rates (Greenberg 1999, Liu, Miller et al. 2003) the envelopes of each vocalization sequence, $x(t)$, were first filtered to a maximum frequency f_m (Table 2.1) in order to detect the main vocalization onset and offset components and remove undesirable high frequency envelope noise. Rat and mouse pup vocalizations tended to occur at a faster rate and thus the envelopes were filtered 30 Hz whereas for speech, infant cry, bird and primate vocalization rates were slower and the envelopes were filtered at 10 Hz. The onset and offset vocalizations transients were identified if the envelope exceeded a designated threshold level (5 to 100 SD) above the envelope of sequence segments containing background noise. For each detected vocalization segment, $x_n(t)$, we next fitted a rectangular pulse of variable start time (t_n), duration (D_n), and peak amplitude (A_n). The start time for each detected vocalization segment was defined by the onset time where the envelope crossed 50% level relative to the peak amplitude while the duration was defined by the width of the envelope at 50% of the peak amplitude. Finally, the fitted model envelope was normalized for unit variance.

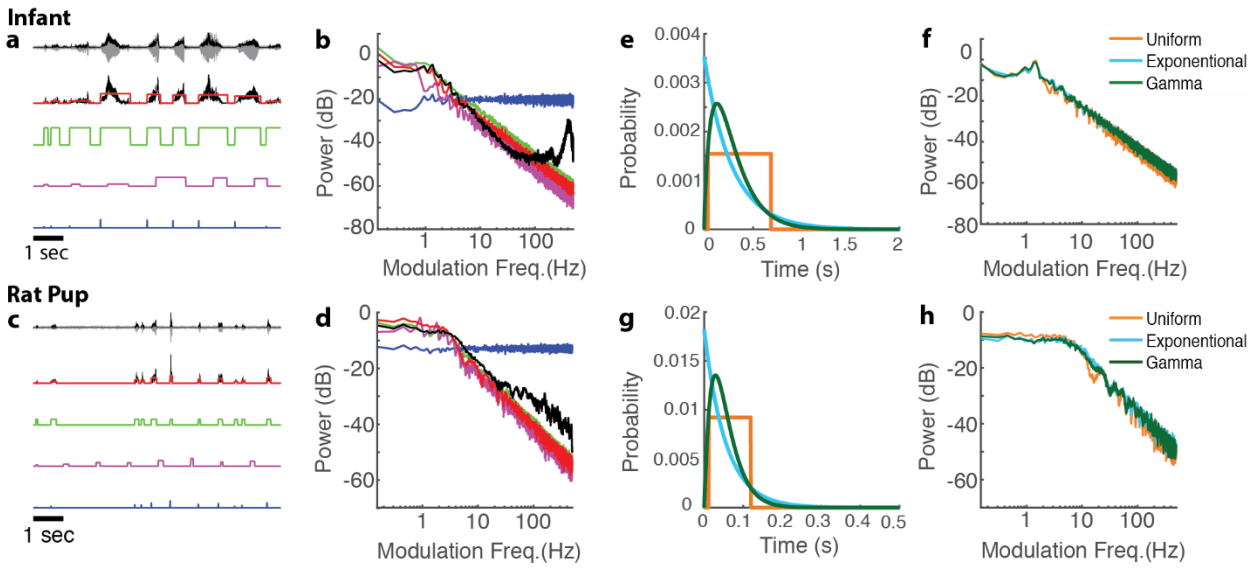


FIG. 2.1. Relationship between vocalization envelope and the modulation power spectrum illustrated for a crying baby (a and b) and rat pup (c and d) vocalization sequences. (a and c) The original sound waveforms (gray line) and envelopes (black line) are shown along with the pulsed vocalization model (red line). Three models are also shown where one of the three parameters (inter-vocalization interval, duration, or amplitude) was perturbed. The perturbed pulse sequences have either a constant pulse amplitudes (green), constant inter-vocalization interval (magenta line), or zero duration (blue line). (b and d) Power spectrum for original vocalization envelope and corresponding models (same color convention). (e and g) Vocalization were also perturbed by synthetically modifying the envelopes so that the infant (e) or rat (g) vocalization durations follow either a uniform, exponential, or gamma distribution (with matched mean and variance as the of original vocalization). (h) The corresponding MPS for the tested duration distributions.

2.3.2 Effect of Acoustic Features

Fig. 2.1 illustrates how each of the acoustic features accounted for by the model parameters affect the MPS of natural vocalization sequences from an infant (a and b) and a rat pup (c and d), respectively. Vocalization amplitudes, onset times, and duration parameters were obtained for each vocalization in the sequence by fitting the model (a and c; red curve) to the original sound envelope (a and c; black curve) and the MPS of the model envelope was computed (Fig. 2.1c and d).

As seen in Fig. 2.1 (b and d), the model MPS (red) has a lowpass shape and 1/f scaling similar to the original vocalization sequence MPS (black; ~1-100 Hz for the infant and for frequencies greater ~4 Hz for the rat). Although the model follows the actual MPS for low and intermediate modulation frequencies, it tends to deviate from the actual MPS at high modulation frequencies (Fig. 2.1b, >100 Hz for infant; Fig.2.1d, >40 Hz for rat pup). In humans, these higher modulation frequencies correspond to the periodicity pitch which is generated by vocal fold vibration (Rodriguez, Chen et al. 2010) and though critical for identifying speech source attributes such as gender, they are not essential for vocalization recognition (Elliott, Theunissen 2009). To assess the contribution of each acoustic feature to the MPS, we synthetically altered the model envelopes by perturbing one of the three parameters and examined the resulting model MPS. Removing the amplitude variability by fixing the envelope amplitudes (Fig.2. 1a and c, green) has minimal effect on the model MPS (Fig.2.1b and d, green). Similarly, when the onset time variability is removed by imposing a constant inter-vocalization interval of 1 sec, which is defined as a time difference between adjacent vocalization onsets (Fig.2. 1a and c, magenta) there is little effect on the resulting model MPS (Fig. 2.1b and d, magenta). In contrast, constraining the vocalization duration by replacing the square pulses with a Dirac impulse while maintaining all other parameters (Fig. 2.1a and c, blue; pulse duration of zero) results in flat MPS (Fig. 2.1b and d, blue), which differs dramatically from the full model and the original sound MPS. These simulation results indicate that vocalization durations are critical determinants of the 1/f MPS shape including its 1/f trend; whereas variations in vocalization onset timing and amplitudes are not.

2.3.3 Effect of Distribution of Vocalization Duration

As mentioned in previous section, changing the overall pulse durations can have a dramatic impact on the overall shape of the MPS. We thus asked whether and to what extent the distribution of

vocalization durations has an effect on the observed 1/f modulation spectrum. In addition to considering the actual distribution of vocalization durations in our simulation (as in Fig. 2.1b and d), we also generated synthetic envelopes with perturbed duration distributions (Fig. 2.1e and g), which include uniform (orange), exponential (light blue), and gamma (dark green) distributed durations with matched mean and variance to the original data from rat and pup vocalizations. As can be seen, the MPS is largely unaffected by the model distributions as long as they have a matched mean and variance (Fig. 2.1f, Fig2.1h). Thus, the MPS shape is largely independent of the type of distribution used to model the vocalization durations. However, as will be demonstrated subsequently, the first and second duration moments (mean and variance) have a substantial impact on the overall MPS shape.

2.3.4 Statistical Structure of Model Parameters

To further assess the role of acoustic features to the temporal structure of the vocalization sequences, we examined the statistical structure of the three model parameters. Figure 2.2 shows the joint distribution of pulse durations and amplitudes measured for the infant (Fig. 2.2a) and rat pup vocalizations. The distributions are tightly distributed with relatively compact support. Durations and amplitudes exhibit a significant but weak correlation (infant, r=0.16 ± 0.05; rat, r=0.26±0.08; mean±SE; t-test, p<0.01; see Table2.3 for additional vocalization statistics), consistent with the independence approximation of the model. The autocovariances for the measured duration and amplitude time series have impulsive structure, indicating minimal serial correlation for the infant vocalization sequence (Fig. 2.2c and d). Furthermore, the inter-vocalization intervals defined as the time difference between consecutive vocalization onset times, $\Delta t_n = t_n - t_{n-1}$, follow an approximately exponential distribution as expected for a Poisson point process (Fig. 2.2b), although there is a short latent period (~150 ms, infant; ~30 ms, rat pup) in the

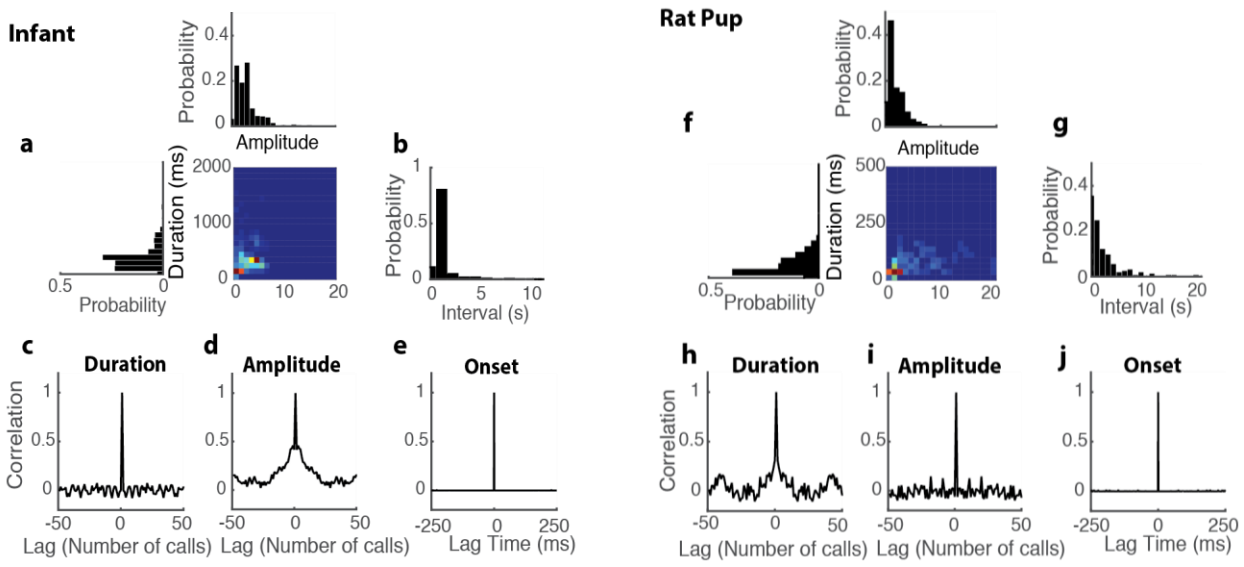


FIG. 2.2. Vocalization parameters and serial statistics for a crying infant (a-e) and rat pup call (f-j). (a and f) Joint distribution of vocalization duration and amplitude is tightly distributed. The duration and amplitude marginal distributions are shown to the left and above the joint distribution. Inter-vocalization interval distributions (b and g) exhibit long exponential-like tails and a refractory region at short intervals. Serial statistics of the vocalization parameters exhibit weak temporal correlation (c-e for a crying infant and h-j for rat pup call). Duration (c and h) and amplitude (d and i) parameters are largely serially uncorrelated. (e and j) Normalized autocovariance for a point process consisting of onset times for each vocalization exhibits an impulsive correlation.

interval distribution indicating a brief silent period between consecutive vocalizations. Inter-vocalization intervals are weakly correlated with the vocalization duration and amplitude parameters (Table2.2). Finally, upon treating the vocalization onset times as a renewal point process, we find that these are uncorrelated as evident from the impulse structure of the point process autocovariance (Fig. 2.2e). These analyses indicate that vocalization durations, amplitudes, and inter-vocalization intervals are distributed in a largely independent and serially uncorrelated fashion within a vocalization sequence.

Table 2.3: Joint correlation statistics between the measured model parameters for each vocalization sequence. Correlation statistics between the vocalization amplitudes (*A*), durations (*D*), and inter-vocalization intervals (*I*) are quantified using the Pearson correlation coefficient (mean±SEM). A significant correlation is noted by a * (bootstrap t-test, p<0.01).

	r_{AD}	r_{AI}	r_{DI}
Rat pup	0.26 ± 0.08*	-0.20 ± 0.04*	-0.12 ± 0.06
Mouse pup	0.12 ± 0.03*	-0.07 ± 0.01*	-0.04 ± 0.02
Bird	0.18 ± 0.03*	0.04 ± 0.05	0.29 ± 0.20
Primate	0.19 ± 0.08	0.16 ± 0.10	0.26 ± 0.09 *
Infant	0.16 ± 0.05*	-0.20 ± 0.03*	-0.002 ± 0.04
Speech	0.32 ± 0.04*	-0.01 ± 0.04	0.06 ± 0.04

2.3.5 Analytical vs. Experimental Model

To gain further insight on how each envelope parameters contributes to the 1/f MPS structure, we derived the model MPS in closed form by computing the power spectral density of the model envelope. Assuming independence of the model parameters (as in Fig. 2.2, Table 2.2) the model MPS is :

$$\begin{aligned} S_{xx}(f) &= \lambda \cdot E\left[A_n^2\right] \cdot E\left[\frac{\sin^2(\pi D_n f)}{\pi^2 f^2}\right] \\ &= \lambda \cdot \left(\mu_A^2 + \sigma_A^2\right) \cdot E\left[\frac{\sin^2(\pi D_n f)}{\pi^2 f^2}\right]. \end{aligned} \qquad (2)$$

Here $E[\cdot]$ is the expectation operator, μ_A^2 and σ_A^2 are the peak amplitude mean-squared and variance, and $E[A_n^2] = \mu_A^2 + \sigma_A^2$ is the second-order moment of A_n. This result demonstrates that although the rate of vocalizations (λ) and peak amplitude statistics ($\mu_A^2 + \sigma_A^2$) can both affect the overall MPS by a multiplicative gain factor, they do not affect the shape of the MPS. Instead, the shape of the model MPS, and presumably its 1/f structure, is dictate by the distribution of pulse

durations. To explore this possibility, we note that the distributions of pulse durations is relatively compact (65-671 ms for infant and 12-135 ms for rat vocalizations; 90th percentile range) and thus we assume that the distributions can be approximated as uniform distributions with matched mean and variance to the actual measurements. The MPS is then evaluated in closed form as (Section 2.2.2)

$$S_{xx}(f) = \frac{\lambda\left(\mu_A^2 + \sigma_A^2\right)}{2\pi^2 f^2} \cdot \left(1 - \frac{T_2}{T_2 - T_1}\sin c\left(2\pi T_2 f\right) + \frac{T_1}{T_2 - T_1}\mathrm{sinc}\left(2\pi T_1 f\right)\right). \quad (3)$$

The closed form solution of the model MPS (Fig. 2.3, dotted blue line) is shown alongside the simulated (Fig. 2.3, red line) and actual (Fig. 2.3, black line) MPS for the infant and the rat pup vocalization sequences. Despite the simplifying assumption of uniformly distributed vocalization durations, the model captures the general structure of the MPS. Not only is the 1/f structure well explained by the model, but as seen in the actual MPS, the low frequency region is flat and lacks 1/f structure. These two-regimes are well captured by the model as seen in Fig. 2.3. As $f \to 0$, it can be shown by applying L'Hospital's rule that:

$$S_{xx}(0) = \frac{\lambda\left(\mu_A^2 + \sigma_A^2\right)}{3\left(T_2 - T_1\right)}\left(T_2^3 - T_1^3\right) \quad (4)$$

which is the limiting value in the flat region in the MPS as observed in the model and actual data. By comparison, in the limiting case where the modulation frequency is large (i.e, $f \to \infty$) it is easy to verify that:

$$S_{xx}(f) = \frac{\lambda\left(\mu_A^2 + \sigma_A^2\right)}{2\pi^2} \cdot \frac{1}{f^2} \quad (5)$$

so that the MPS behaves as a power law for high modulation frequencies with a power law exponent of $\alpha = 2$. We have observed this dual regime lowpass filter structure in a variety of additional vocalization sequences (Fig. 2.3; mouse pup, primate, bird, speech). In all cases, the model captures the general lowpass structure with 1/f trend. As can be seen the vocalization MPS can deviate from the model, largely as a result of vocalization production mechanisms not directly responsible for the acoustic onsets and offsets (e.g., vocal fold vibration). Furthermore, differences may also arise due to the simplifying assumptions (independence between the model parameters and uniformly distributed durations). Despite this, the residual error spectrum (Fig. 2.2) lacks 1/f structure indicating that temporal edges are the main acoustic components accounting for the general 1/f behavior.

Given the dual-regime lowpass structure of the MPS, it is of interest to identify how the transition point between the flat and 1/f regimes is related to the vocalization model and its statistical parameters. This was approached by finding the half power or cutoff frequency (f_c) of the MPS, which for the model yields (Section 2.2.4)

$$f_c \approx \frac{1}{\pi} \cdot \frac{1}{\sqrt{E\left[D_n^2\right]}} = \frac{1}{\pi} \cdot \frac{1}{\sqrt{\left(\mu_A^2 + \sigma_A^2\right)}} \qquad (6)$$

where μ_D and σ_D^2 are the duration mean and variance, respectively. This result suggests that the

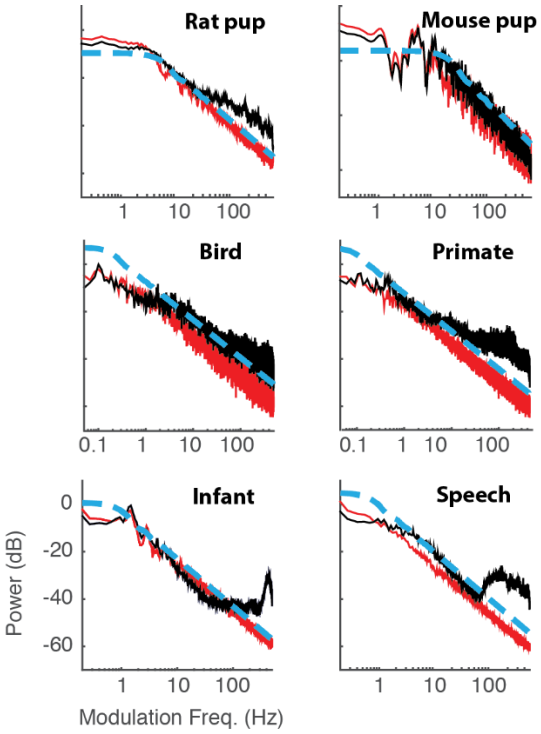


FIG. 2.3. Comparison of MPS from different species with the model simulation and the theoretical solutions. MPS are shown for rat and mouse pup, birds, primate, crying infant and speech (black). The simulated pulse vocalization model (red) obtained by fitting each isolated vocalization with a rectangular pulse (Fig. 2.1) has lowpass structure and 1/f trend at high frequencies that mirrors the 1/f decrease in the actual MPS. The closed form theoretical solution (Eq. 3; dotted blue) likewise exhibits a lowpass structure with 1/f trend at high frequencies.

vocalization duration statistics (D_n) alone account for the cutoff frequency. Specifically, the cutoff frequency is inversely related to the square root of the second order moment of the vocalization duration distribution. This result is a statistical variant of the uncertainty principle for a vocalization ensemble, which requires that the signal duration in the time-domain be inversely related to its bandwidth in the frequency-domain (Cohen 1995b).

This approximate solution is on average within 8.0 % of the actual cutoff frequency of the model for the vocalizations tested. Furthermore, the theoretical solution accurately predicts the empirical data. As seen in Fig. 2.4,the measured cutoff frequencies for the different species closely match the model predictions (Pearson r=0.89±0.15, mean±SEM; t-test, p<0.01). Furthermore, measured cutoff frequencies for different species are inversely related to the measured second-order duration moment (log(f_c) vs. log (√ (E [D_n^2])), Pearson r=-0.89±0.15, mean±SEM; t-test, p<0.01) as predicted by Eq. 6 (dotted line)

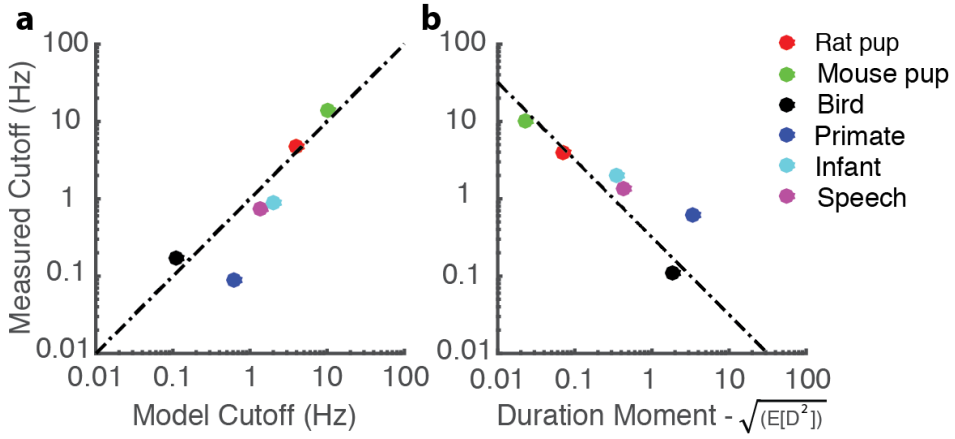


FIG. 2.4. Empirically measured and predicted cutoff frequencies for different species. The predicted cutoff frequencies obtained from the vocalization duration statistics (Eq. 6) for the different species closely match the actual measurements. Empirically measured cutoff and duration second moment are inversely related as predicted by the model (Eq. 6; dashed dot line).

2.3.6. Conceptual Model

Finally, we explored the mechanism by which temporal edges lead to the observed MPS to gain insight on how onsets and offsets in vocalizations produce the lowpass structure with 1/f trend. Conceptually, the MPS of a vocalization sequence can be derived by averaging the pulse spectrum across an ensemble of rectangular pulses, where the power spectrum of each isolated rectangular pulse is:

$$S_{p_n p_n}(f) = A_n^2 \cdot D_n^2 \cdot \mathrm{sinc}^2(\pi D_n f) = A_n^2 \cdot \frac{\sin^2(\pi D_n f)}{\pi^2 f^2}. \quad (7)$$

The power spectrum (Fig. 2.5b) for three rectangular pulses (Fig. 2.5a) taken from the speech ensemble are shown along with the average ensemble MPS (Fig. 2.5c). Note that the individual pulse spectrum consists of a sinc^2 function with a main lobe bandwidth that is inversely related to

the pulse duration and spectrum amplitude proportional to the squared of the pulse amplitude. The spectrum side lobes (secondary peaks in the spectrum) have a peak power that decreases proportional to f^{-2} as noted by the blue dotted line depicting the maximum side lobe power. The vocalization sequence MPS is obtained by averaging across the ensemble of pulses. Since pulses in the speech ensemble have different durations and amplitudes, which produce different side lobe and notch configurations, interferences between the side lobes and the adjacent notches cancel each other. On average the resulting MPS has a flat regime with bandwidth that is inversely related to the pulse duration second moment and a decreasing 1/f regime created by interference of the isolated vocalization side lobes.

2.4. Discussion

Although the presence of 1/f scaling behavior in the amplitude fluctuations of natural sounds was first observed more than 40 years ago (Voss, Clarke 1978), the physical origins of this phenomenon have remained elusive. Our results identify for the first time a single physical cue in vocalization sounds and speech that is responsible for this phenomenon. The findings demonstrate that temporal edges at the beginning and end of isolated vocalizations are the principal determinants of 1/f scaling. The findings have a number of implications for models of auditory perception and coding since rising and decaying temporal edges are perceptually salient and such physical cues serve as temporal boundaries for acoustic objects.

Although we have not extended the acoustic analysis employed here to broader categories of sounds, other natural sounds (Geffen, Gervain et al. 2011, Attias, Schreinert 1997b, Rodriguez, Chen et al. 2010, Singh, Theunissen 2003) and music (Voss, Clarke 1978) also exhibit 1/f scaling. Since sounds and music, in general, are composed of transient and time-varying acoustic elements

that can be coarsely modeled as onsets and offsets future studies will need to explore how these findings generalize into a theoretical framework that applies to an even broader range of natural and man-made sounds.

The results have a number of implications for theories of coding by the brain since neurons throughout the auditory pathway are exquisitely sensitive to temporal transitions with millisecond precision. Furthermore, central auditory neurons have been shown to produce an efficient neural representation in which they equalize the modulation power and thus cancel the 1/f trend in natural sounds (Rodriguez, Chen et al. 2010). Similar efficient coding strategies have been proposed in the visual system where visual cortex neurons through edge detection equalize or "whiten" the spectrum of natural images, which enables a more equitable use of neural resources (Field 1987). Mechanistically, two distinct temporal coding mechanisms could contribute to such efficient representation. First, central auditory neurons have temporal responses with leading excitation and lagged inhibition, which produce an on-off temporal integration pattern akin to taking a smooth temporal derivative operation of the sound envelope (Bregman 1994, Lee, Osman et al. 2016, Heil 1997, Zheng, Escabi 2008). In the time domain, such an operation can accentuate and facilitate detection of temporal boundaries for important information bearing acoustic segments, analogous to edge detection in vision (Ruderman 1997, Zylberberg, Pfau et al. 2012). Alternately, in the frequency domain a temporal derivative operation has a transfer function magnitude $H(f)^2 = 4\pi^2 f^2$ that opposes and precisely cancels the 1/f trend seen in vocalization envelopes. Secondly, power equalization could be achieved through modulation filter bandwidth scaling as previously described for auditory midbrain neurons (Rodriguez, Chen et al. 2010) and perceptually (Depireux, Simon et al. 2001) measured modulation filters. In both cases, neural and perceptually modulation filter bandwidths scale proportional to the modulation frequency of sounds, in such a way that

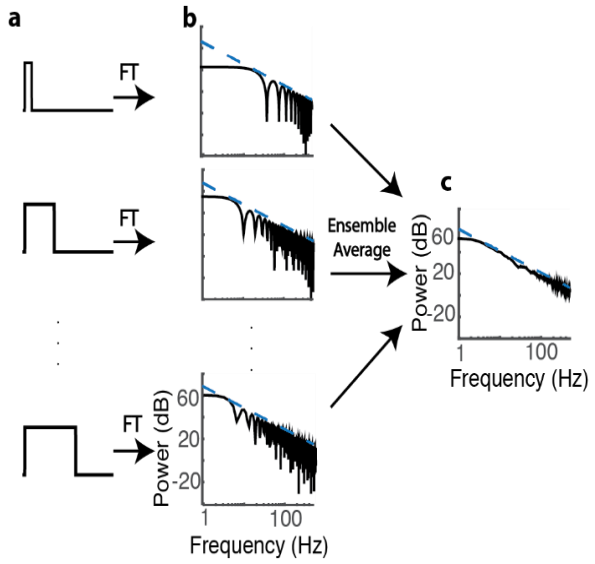


FIG.2.5. Relationship between vocalization pulse approximation and the observed 1/f modulation power spectrum. (a) Three example pulses from the speech ensemble and the corresponding power spectrum obtained by taking the Fourier Transform (FT) of each pulse (b). The spectrum for each pulse consists of a sinc2 function with side lobe peaks and notch locations that depend on the vocalization duration and the notch peak amplitudes that drop off proportional to $1/f^2$ (blue dotted lines). The model MPS is obtained as the ensemble average across all durations, which produces an average spectrum with lowpass structure and $1/f^2$ trend at high frequencies.

the power output of each filter increases proportional to the f^2, canceling the $1/f^2$ trend observed for natural sounds. Thus, temporal edge detection and bandwidth scaling may serve as general mechanisms to whiten and thus equalize the modulation power in vocalizations allowing for efficient information transfer and coding by the brain, analogous to principles in vision.

Finally, these findings can be applied to sound coding and hearing technologies. Specifically, the statistical framework could be used to improve coding, compression, and sound recognition algorithms. The statistical framework and the resulting observations could further be used to improve filtering algorithms to enhance detection of transient sound elements (Miller, Escabi et al. 2002) and to facilitate recognition in hearing aid and cochlear implant technologies by equalizing the modulation signal power.

2.5. Funding Source

This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award R01DC015138 and the National Science Foundation under award IOS 1355065.

Chapter 3: The Role of Statistical Regularities for Sound Category Identification

3.1 Introduction

Environmental sounds, both man-made and natural, vary on multiple time and frequency scales and cause a large range of temporal, spectral and amplitude modulations that are obvious in the high order statistics of the sound spectrogram. Healthy hearing humans use this information to categorize and discriminate sounds. Although it has been shown that neural responses can be affected by high-order sound statistics, the neural mechanisms under auditory pathways that confer these abilities are unknown. Prior work demonstrated that neurons can respond selectively to sound statistics, there is a motivation to explore how statistics of sounds play a role in neural representations (Escabi, Miller et al. 2003, Attias, Schreiner 1998, Lesica, Grothe 2008).

In this chapter, I will explore the hypothesis that auditory pathway neurons respond reliably to higher-order statistical regularities in natural sounds. Further, we hypothesize that neural populations can utilize these responses for sound discrimination and categorization. It is known that neural responses in the central auditory system can be modulated by a variety of high-order sound statistics, including modulation and correlation statistics (Escabi, Miller et al. 2003, Attias, Schreiner 1998, Lesica, Grothe 2008) . I am going to test the hypothesis that neural responses in the inferior colliculus are modulated by high-order statistical regularities in sounds and that statistical features of the neural response can ultimately be used to discriminate and categorize sounds. To test the above hypothesis, we will perform neural recordings in inferior colliculus of unanesthetized rabbits in response to an ensemble of texture sounds which are produced by the concurrence of many similar acoustic events that overlap in time (McDermott, Simoncelli 2011a). Many natural sounds, such as rain, fire, and water stream are considered as texture sounds and I will use some of them in my study.

3.2 Materials and Methods

3.2.1 Sound Dataset

Five different natural sounds (No additional sounds in background), fire (Compact Fire), bird (Golden_winged_Parakeet), crowd (Africa-Crowd-Outdoor), water (Mountain Stream), and snake

(Western-Diomondback Rattele snake) (Bradbury, Budney). Sounds were sampled at 44.1kHz sampling rate. From 5 second of each sound, 3 second duration of it has been selected which is synthetically manipulated by adding high order statistics, and stimuli sounds are delivered sequentially and randomly with 0.1 second in between split and played to rabbits in 2-hour experiment sessions. A short 0.5 second segment of the original sounds is shown in Figure 3.1.

3.2.2. Data Recording

Data is collected from Dutch rabbits. Rabbits are selected because they can sit still during recording, so we can record neural data while they are awake. Also, the range of rabbit hearing (100 Hz to 30 kHz) overlaps and is comparable to human range (20 Hz to 20 kHz). All animals are housed and handled according to a protocol approved by the Institutional Animal Care and Use Committee (IACUC) of the University of Connecticut. Both surgery and recording are conducted in the vivarium at University of Connecticut.

Before each surgery, rabbits were trained to sit still for two hours. Two surgeries are performed on rabbits to prepare rabbits for recording. The aim of the first surgery is to put a fixation bar which help us to fix the rabbit's head during recording. In a second surgery craniotomies are performed over the right temporal cortex. Anesthesia is induced and maintained with a cocktail of ketamine throughout the surgery. Reprosil cap (Hydrophilic Vinyl Polysiloxano Impression

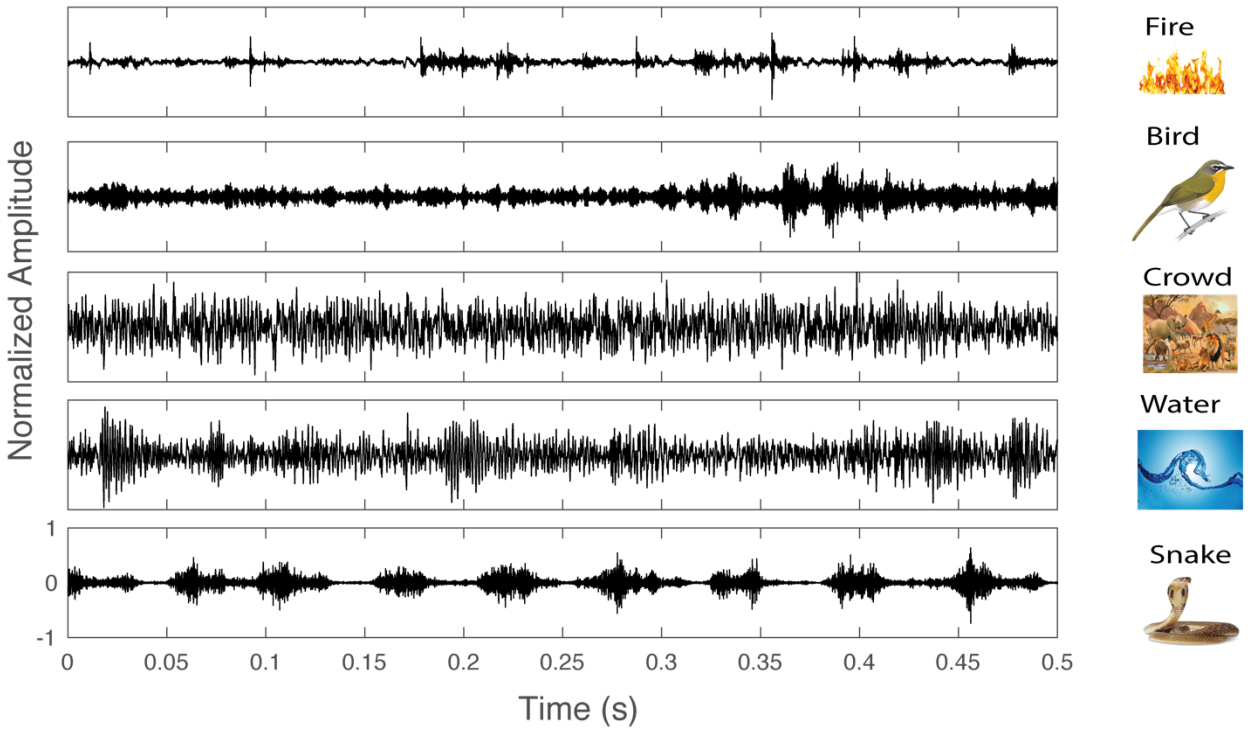


Figure3.1. Natural sound stimulus: 0.5 seconds of each sound category stimulus in time domain.

materials) are used after surgery to cover the cortex. Also after surgery, using Reprosil material, a special ear molds are produced for each rabbit for both ears in order to have clean sound delivery. After two weeks of surgery and rabbits' recovery electrophysiological recording procedures have been done without using any anesthesia under the approved protocol by IACUC.

NeuroNexus probes are batch fabricated in the state-of-the-art clean room. NeuroNexus probes are packaged in PCBs for easy connection to standard data acquisition systems ((http://www.neuronexus.com) 16 array NeuroNexus probes are used for recording electrophysiological

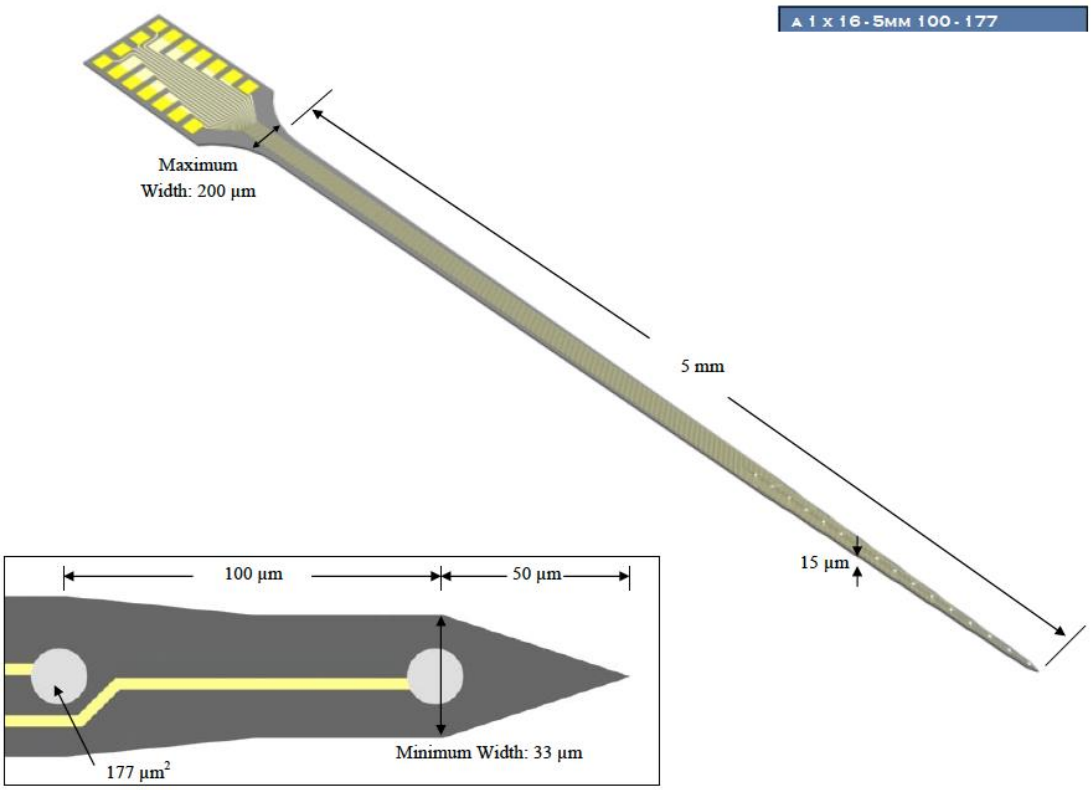
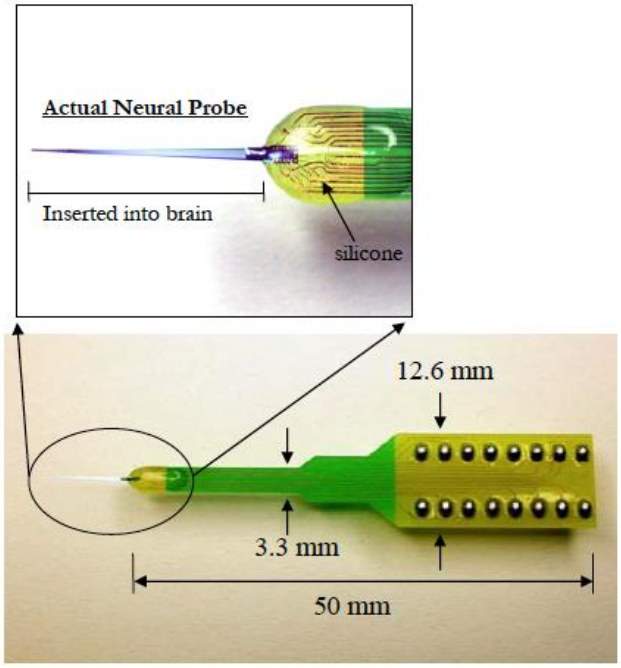


Figure3. 2. Schematic of 16 array NeuroNexus probe that is used for electrophysiological recording (http://www.neuronexus.com).

responses from Inferior colliculus of anaesthetized rabbit. This probe (Figure 3.2) has one shank with the length of 5mm which the shank has 16 sites (channels) for recording. Each site area is $177\mu m^2$ and distance between each site (channel) is 100 μm. This probe is selected because frequency responses in central inferior colliculus increasing tonotopically, and we record across this region, so that we can observe this tonitopical changes.

The layout of the16 array NeuroNexus probe is demonstrated in Figure3.3(a) with details in its dimension and the shank which is inserted into the brain for extracellular recording. This probe is fixed and connected to the amplifier before start recording (Figure3.3(b)). All recording is done with TDT equipment, RZ2, RS4, and OpenEx software from Tucker-Davis Technologies.



(a) (b)

Figure3. 3. Sixteen array NeuroNexus probe and paradigm used for electrophysiological recording (http://www.neuronexus.com).

3.2.3. Stimuli dataset analysis and statistics selection

High-order statistical regularities in the cochlear representation of sounds are critical for recognition and identification. Using texture sounds and synthesis procedure, we test whether neural responses in the auditory midbrain (inferior colliculus, IC) of awake rabbits are sensitive to sound statistics and whether neural response statistics can be used to discriminate sounds. Five different sounds including fire, bird, water, crowd, and snake were used to generate synthetic variants. The statistics considered include of 1) Spectrum (S1), 2) Envelope Marginals (Mean and

Var) (S2),3) Modulation spectrum (S3), and 4) Correlation structure (S4). These tested statistics are chosen because: 1) They are perceptually salient and have been shown to contribute to perception and discrimination of sounds and 2) Neurons in the inferior colliculus are sensitive to the same statistics.

At first, all sounds are decomposed with auditory filterbanks and the envelope is extracted from spectrogram and its average and its variance are considered as one of the statistics. Then for extracting other sound statistics and adding them later to the white noise, sounds were decomposed using a cochlear model representation (Figure3.4) and statistics from the sound spectrogram were extracted and used to generate synthetic variants of each sound (McDermott, Simoncelli 2011a) . By using 60 gamma tone filters (Fig3.4, Escabi 2003), and modulation frequency of 500Hz, cochleogram representation of each sound is extracted. From the cochleogram, the correlation matrix and modulation spectrum are extracted for further analysis.

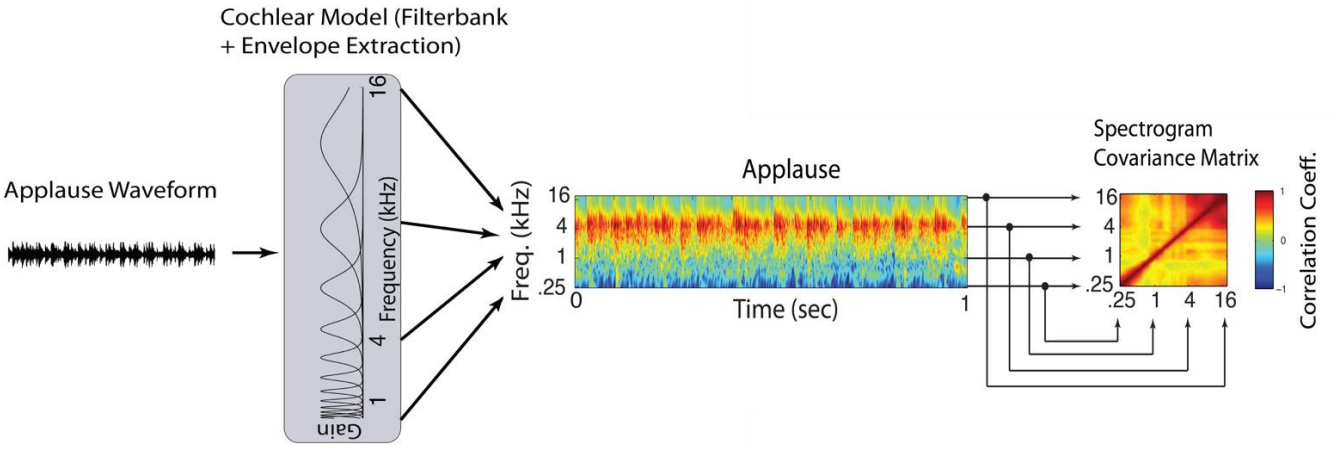


Figure3.4. Decomposing sounds to spectrotemporal pattern and covariance calculation from resulted cochleogram (Rodriguez, Chen et al. 2010).

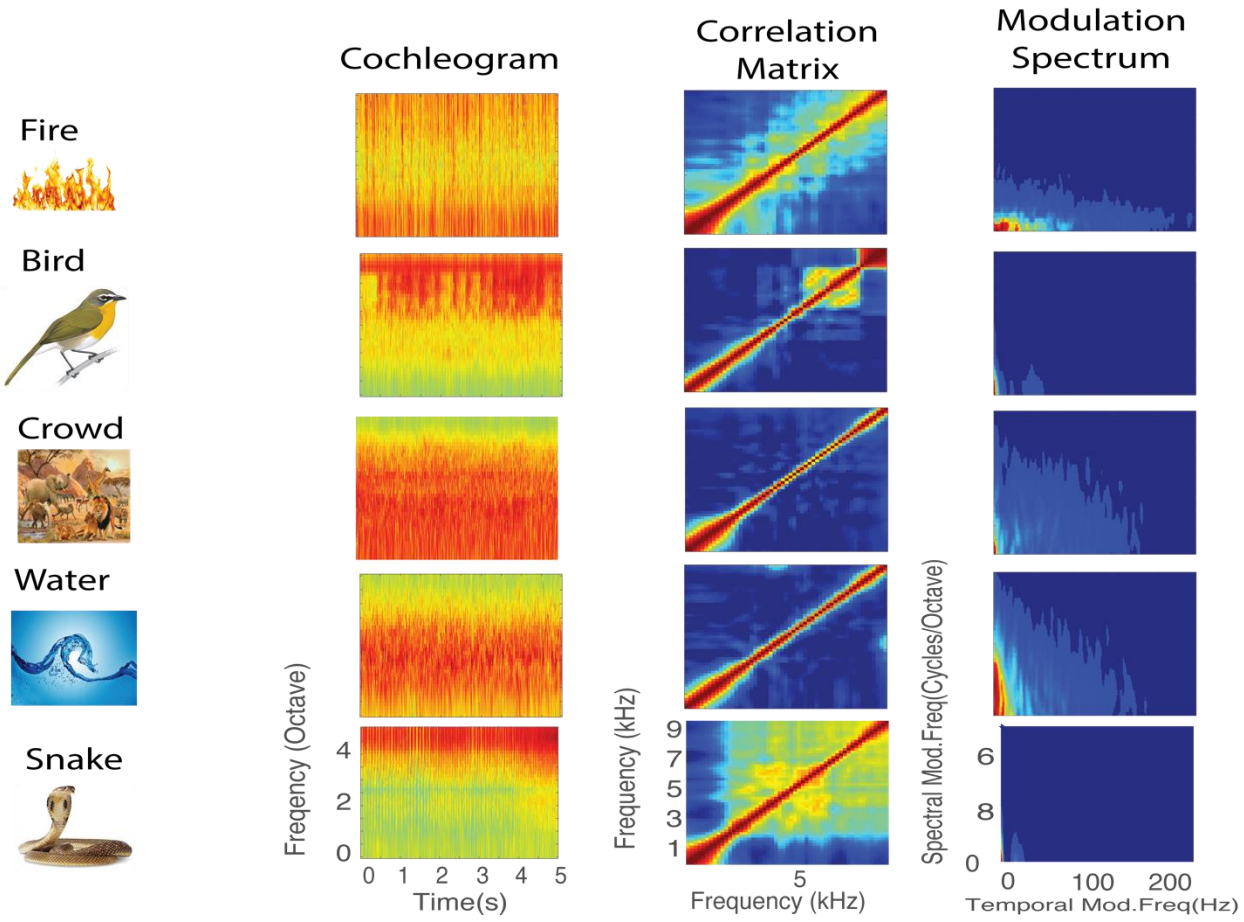


Figure3.5. Extracted features from cochleogram for different natural sounds.

The cochleogram, correlation matrix, and modulation spectrum are shown in Figure 3.5. The Correlation matrix is obtained by measuring correlation between the envelopes of all pairwise selected frequency subbands in the cochleogram. The normalized covariance between subband i and j is:

$$C_{ij} = C(x_i, x_j) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_i - \mu_i}{\sigma_i}\right)\left(\frac{x_j - \mu_j}{\sigma_j}\right)$$

where x_i is the temporal envelope in subband i and x_j is temporal envelope in subband j, N is number of time points, μ and σ are average and standard deviation of the envelopes in each sub

band. C_{ij} is the correlation between each two bands. And the correlation matrix can be defined based on the correlation between each frequency subband by below definition (Kendall 1979):

$$R = \begin{pmatrix} C_{1,1} & C_{1,2} \dots & C_{1,j} \\ C_{2,1} & C_{2,2} \dots & C_{2,j} \\ & \vdots & \\ C_{j,1} & C_{j,2} \dots & C_{j,j} \end{pmatrix}$$

The correlation matrix (R) of spectrogram has been plotted for different sounds in Figure 3.5. Although the cochleogram of some sounds, for example bird and snake, shared high frequency components and they are similar to each other, correlation matrix of them are completely different. Although by using correlation matrix we can differentiate these two sound categories, discriminating crowd and water category is still not easy based on spectrogram and correlation matrix.

After sound decomposing into the spectrogram, modulation power spectrum (MPS) is computed (Rodriguez, Chen et al. 2010, Singh, Theunissen 2003). MPS is a function of sound's temporal and spectral modulation. MPS is calculating by segmenting spectrotemporal envelop and then averaging over each half second block. If $s_n(t, x_k)$ is a nth segment of spectrotemporal envelop, MPS of each block is calculated as below (Rodriguez 2010):

$$P(f_m, \Omega) = \frac{1}{N} \sum_{n=1}^{N} |\Im\{s_n(t, x_k).w(t, x_k)\}|^2$$

where f_m is temporal modulation frequency (Hz), Ω is spectral modulation frequency (cycles/octave), N is number of blocks, $\Im$ is 2D Fourier transform, $w(t, x_k)$ is two dimensional Kaiser window. MPS of the stimulus dataset is shown in figure 3.5. Although the correlation

matrix of crowd and water are similar to each other, the MPS of them are different from each other, which may help to discriminate them.

Based on stimuli discriminability, we selected four different statistics, power spectrum, envelope marginal, correlation matrix, and MPS for further analysis and generated synthesized sound for neural recording experiment.

3.2.4 Stimuli Sound Synthesis

Each stimuli sound is synthetically manipulated by selectively adding the high-order statistics noted in previous section using the synthesis algorithm developed by McDermott and Simoncelli (McDermott, Simoncelli 2011b).

The statistics measured in the previous section using the cochlear model is used synthesis perturbed variants of the texture sounds. Gaussian white noise is used to initialize the synthetic sounds, which are then modified iteratively until it shared added statistics match. Each cycle of the iterative process has following steps: 1- decomposing the synthetic sound into cochlear subbands, 2- computing subband envelopes using the Hilbert transform, 3- dividing out envelops to the subbands to catch subband fine structure, 4-downsampling envelopes for computational efficiency, 5- measuring envelope statistics to compare to the original sound and generating error signal (squared error between synthetic signal's and original signals' statistics), 6- modifying downsampled envelopes using gradient descent, causing statistics to be closer to the original sound, 7- upsampling modified envelops and combining with fine structure, and 8- finally, combining subbands to have a new synthesized sound. By computing the error for each statistic, convergence was checked for each iteration, and SNR was measuring as the squared error ratio for each statistic and summed up over all statistics.

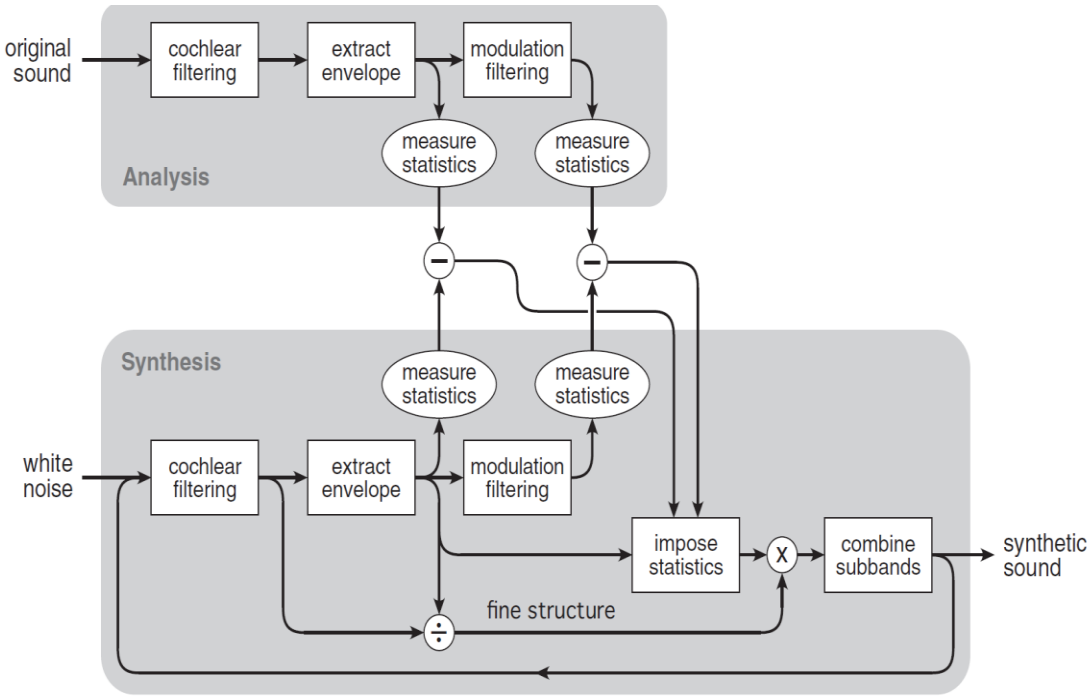
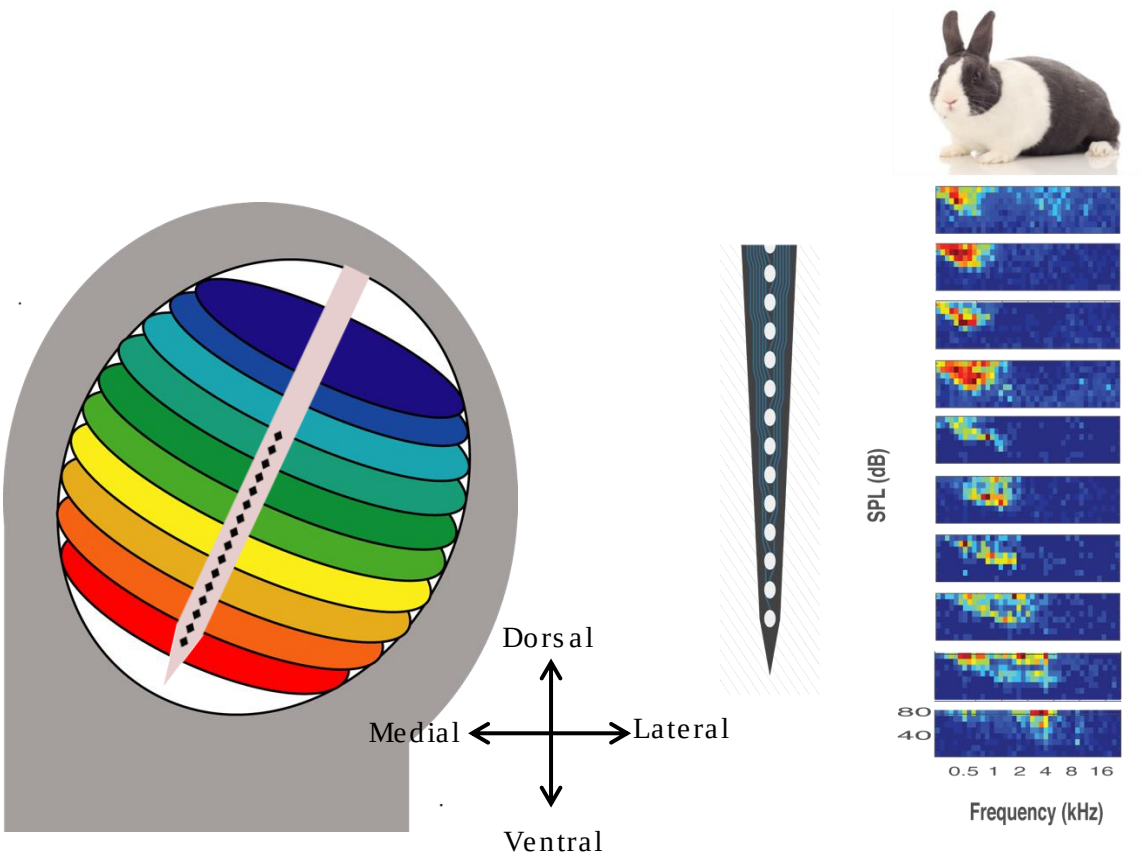


Figure 3.6. Schematic of synthesis procedure introduced by McDermott and Simoncel(McDermott, Simoncelli 2011b).

The iterative synthesis procedure was stopped when either the signal-to-noise ratio (SNR) was higher than 30dB or when 60 iterations were reached (McDermott, Simoncelli 2011b).

3.2.5 Frequency Response Area (FRA)

To identify the central nucleus of the inferior colliculus (ICC), and to characterize the frequency response areas of the selected recording locations, a pseudo random sequence of tone pips (50 ms duration) were delivered to the animals. Tones spanned 0.1-30 kHz in 1/8 octave steps and intensities between 5-85 dB SPL (10 dB steps). After presenting pure tones, the frequency response area of each tone and sound pressure level response is measured by averaging their firing rate. Sites with a clearly identified tonotopic gradient are presumed to lie within the central nucleus of the inferior colliculus (Merzenich, Reid 1974).



(a) (b)

Figure 3.7. (a) Schematic of tonotopic frequency organization in central nucleus of the inferior colliculus. Blue indicates low frequency and red indicates high frequency. (b) Experimentally measured tonotopically organized frequency response area (FRA) from a single penetration.

The frequency organization of ICC is schematized in Fig. 3.7a along with an example recording location showing a clearly defined tonotopic gradient (Fig. 3.7b). In some cases, dynamic ripple (DMR) noise was presented to measure spectro-temporal characteristics of ICC neurons. DMR sequences was presented at 80dB sound pressure level and has total of 20 minutes (two ten minutes) duration which covered frequency range of 0.1-30kHz (Escabi, Schreiner 2002a).

3.2.6 Shuffled Autocorrelogram

The shuffled autocorrelogram is a metric of neural responses that measures spike timing reliability, precision and dependency on the shape (Zheng, Escabi 2008, Chen, Read et al. 2012, Zheng, Escabi 2013). Shuffled autocorrelogram (SAC) is measured by calculating the spike train correlation between all trials for each different stimulus response. The goal of calculating SAC is to see how each statistic can add information across trials, and then investigate how and if the firing reliability, spike timing precision, and strength of responses are systematically related to different sound statistics. For calculating SAC, I used 1s neural responses of each trial to calculate SAC and create the histogram. A maximum delay of 25 ms was used. Pair wise SAC was computed between all trials in each neural recording channel in the Neuronexus recording array (Figure 3.8). After calculating SAC, zero delay values of SAC are used in the covariance matrix for further analysis. Finally, the response correlation in Figure 3.8 is used as a measure of the population response statistics to each sound.

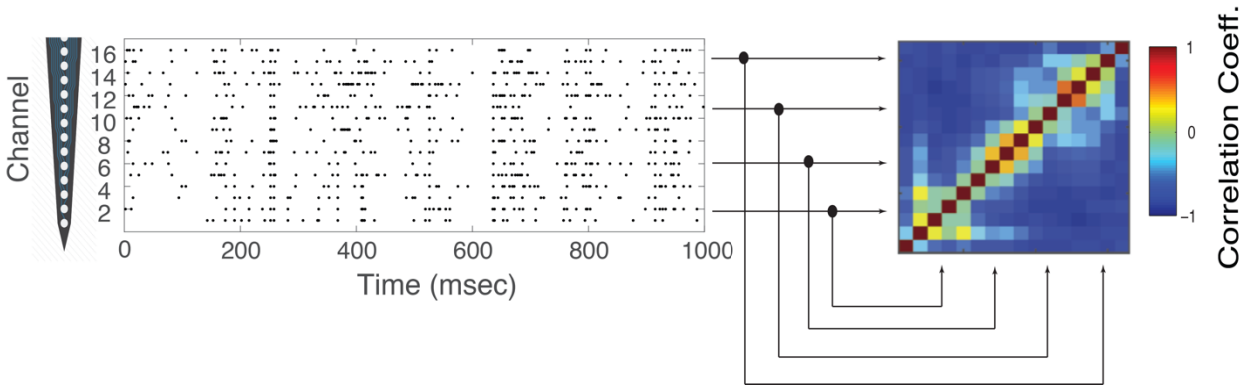


Figure 3.9. Sample response correlation which is used for a metric to compute population response statistics of original and synthesized sound.

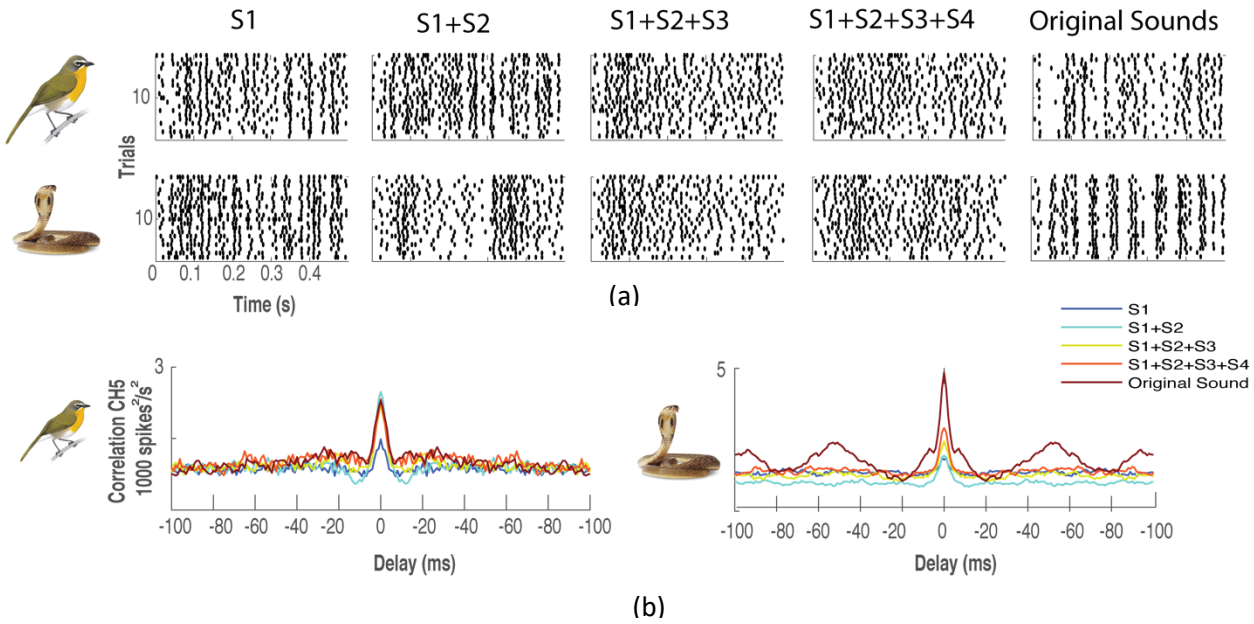


Figure 3.9. (a) sample neural responses of bird vs. snake sounds. (b) Shuffled autocorrelation of same channel (CH5) for different synthesized and original sounds.

3.3 Results

3.3.1 Neural Correlation Statistics and Classifier

Neural response for a single recording site are shown in Figure 3.9 (a) for bird and a snake sound. The corresponding shuffled-autocorrelogram for the same sounds are also shown for the different conditions. As can be seen, the neural response statistics and pattern changes as the sound statistics are varied. Shuffled autocorologram of neural responses of channel five for different statistics is shown in Figure 3.9 (b). It is interesting that in most the cases neural responses to the original sound have the higher correlation, and by adding more statistics, neural responses became

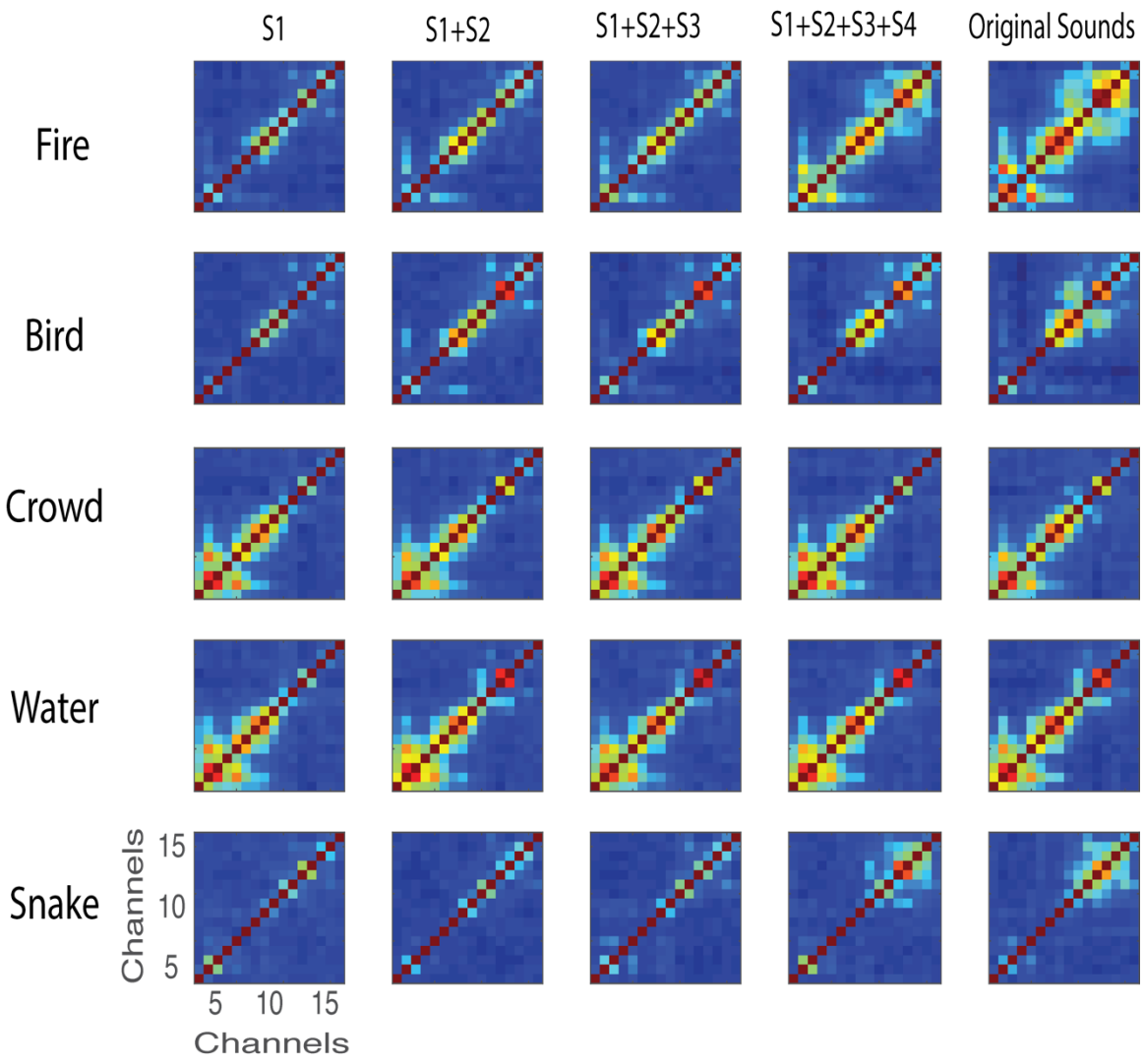


Figure 3.10. Population neural responses using across-channel correlations.

progressively closer to original sound correlation. In the snake sound, a periodic correlation pattern is clear in the neural responses. This periodic response pattern arises because snake rattle and provides information that can potentially be used for differentiating between bird and snake sounds.

Zero delay of correlation responses are considered for measuring population responses of neural coding. Neural correlation statistics were measured between the different channels in a recording site. As can be seen in Figure 3.10, across-channel correlations vary with the particular sound and the statistics that are included in the synthetic variants.

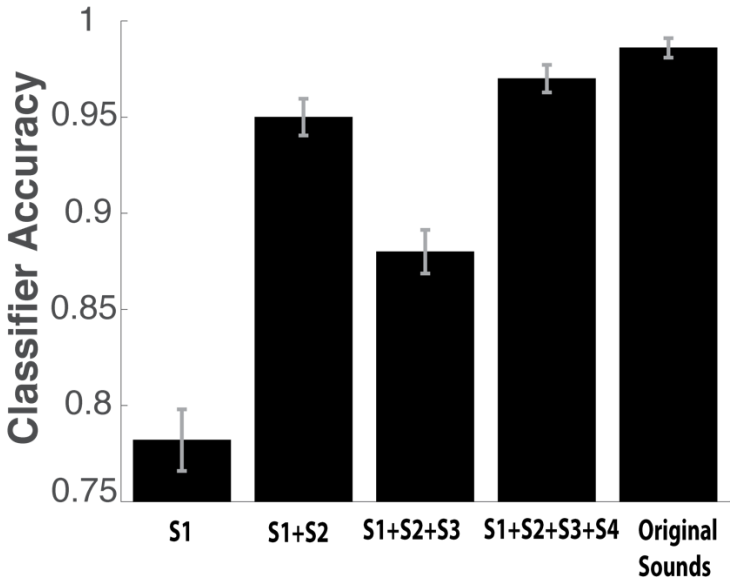


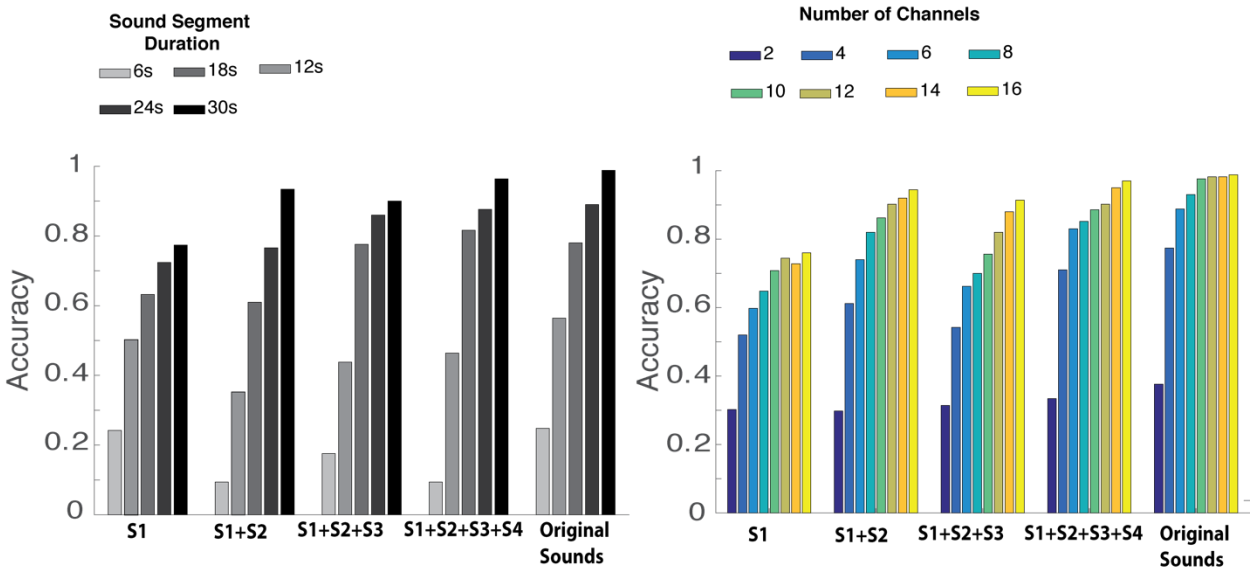
Figure 3.11. Classifier response using cross validation

We developed a neural classifier to read out the population activity and to identify the potential contribution of sound and neural response statistics to sound recognition. The classifier was trained based on the measured correlation structure between recorded channels. A minimum distance between trained and validation set (template match), maximum classification rate was observed for original sound (98.60% whereas chance is 20%). Moreover, adding sound statistics tends to improve classification performance.

3.3.2 Effect of Sound Duration and Neural Population Size on Classification Results

The performance of the neural classifier was tested by changing the number of electrode recording locations and the sound duration. These manipulations allow us to determine how the neural population size and the neural response duration affect recognition performance.

By changing segment duration from 6 second to 30 second, we can assess how discrimination ability is affected by neural recording length. As seen in figure 3.12 (a), the classifier performance improves sequentially with increasing sound duration.



(a) (b)

Figure 3.12. Effect of sound duration and neural population size on classification results. (a) Increasing sound duration can improve classifier performance. (b) Classifier performance improves with increasing population size.

Similarly, increasing the number of recording channels significantly improves the classifier performance, Figure 3.12 (b). The procedure was bootstrapped across channels in order to avoid potential bias of the selected channels. Finally, we note that adding statistics to the sound further improves the classifier performance indicating that sound statistics can potentially play a role in the sound recognition process.

3.4 Discussion

Previously it has been shown (by McDermott) that sound textures and statistics which are common to many natural environments play a crucial role for sound recognition(McDermott, Simoncelli 2011a). By using psychoacoustic experiments on human subjects, they found that synthetic sounds can be recognized by subjects and adding more statistics help subjects to recognize synthetic sounds. Using just spectral information, like spectrum, for synthetic sound is not sufficient for natural sound recognition, causes poor classification(McDermott, Simoncelli 2011a). However, how neurons response to natural sound's statistics and how these responses can use for categorizing sounds in auditory pathway stages including inferior colliculus, is still unclear.

In this study we looked at responses of neurons in central auditory pathway to synthesized sounds, and investigate how high order statistics, such as temporal correlations of the cochleogram can affect neural responses and sound categorization. First of all, by using neural responses recorded from unanesthesized rabbits we show that higher order sound statistics contribute and modulate neural responses in the central inferior colliculus and can be used for further analyzing and ability to recognize sounds. Interestingly neural population statistics (i.e., correlation statistics) in the inferior colliculus can be used to identify /recognize sounds with accuracy between 76% to 98.6% for one of the experiment (using all 16 array channels). At last we have been looking at some parameters such as number of channels and duration which are used for training and

validating of neural responses. We also find that categorization performance improves with increasing population size and sound duration for our experiments.

3.5 Funding sources

Research reported in this chapter was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award Number R01DC015138.

Chapter 4: The Role of Auditory System Hierarchy for Optimal Coding and Sound Recognition

4.1 Introduction

Being able to identify sounds in the presence of background noise is essential for everyday audition and vital for survival. Although current knowledge of auditory physiology is comparatively advanced, the neural transformations responsible for the robustness of the mammalian auditory pathway remain poorly understood. Furthermore, although several cortical mechanisms have been proposed to facilitate robust coding of sounds (Mesgarani, David et al. 2014, Schneider, Woolley 2013), it is still unclear how the sequential organization and resulting transformations performed by the auditory pathway contribute to robust sound recognition.

Mammalian audition is resilient to acoustic variability, such as background noise and multiple talkers, yet how the brain accomplishes this seemingly simple feat is unknown. Our general hypothesis is that the hierarchical processing evident in biological systems and captured

in our model provides a mechanism to sequentially change feature extraction capabilities and ultimately provide an invariant noise robust representation.

Several hierarchical changes in spectral and temporal selectivity have been documented and are consistently observed in the ascending auditory system of mammals. Temporal selectivity and resolution change dramatically over more than an order of magnitude, from a high-resolution representation in the cochlea, where auditory nerve fibers synchronize to temporal features of up to ~1000 Hz, to progressively slower (limited to ~25 Hz) and coarser resolution representation as observed in auditory cortex (Joris, Schreiner et al. 2004). Although changes in spectral selectivity can be described across different stages of the auditory pathway, and spectral resolution is somewhat coarser in central levels, changes in frequency resolution are somewhat more homogeneous and less dramatic (Rodriguez, Read et al. 2010, Miller, Escabi et al. 2002, Mc Laughlin, Van de Sande et al. 2007). It's plausible that such hierarchical transforms across auditory nuclei are essential for feature extraction and ultimately high-level auditory tasks such as acoustic object recognition. Yet, it is unclear whether these sequential transformations comprise an optimal computational strategy for noise robust sound encoding. For instance, from a computational standpoint, it could be postulated that a high-resolution auditory network in which temporal and spectral resolution do not degrade across layers may be better suited for sound recognition as it would faithfully preserve detailed sound information across processing levels.

In this study, I will investigate network mechanisms that contribute to robust coding of sounds in the presence of competing variability, such as background noise and multiple talkers. Although previous studies identified mechanisms that contribute to robustness at the single cell level (Mesgarani, David et al. 2014, Schneider, Woolley 2013), there are no prior studies that have evaluated how and if the auditory pathways ascending organization contributes to robust coding.

Specifically, I will test the hypothesis that the auditory pathway is organized into hierarchical processing stages with sequentially changing feature extraction capabilities (spectro temporal features) that culminated a noise robust representation.

4.2 Materials and Methods

4.2.1 Speech Corpus

Sounds in the experimental dataset consist of isolated digits (*zero* to *nine*) from eight male talkers from LDC TI46 corpus (Liberman). Ten utterances for each digit are used for a total of 800 data samples (8 talkers x 10 digits/subject x 10 utterances/digit). Words are temporally aligned based on the waveform onset (first upward crossing that exceeds 2 SD of the background noise level) and speech babble noise (generated by adding 7 randomly selected speech segments(Branagh, Dearman)) is added at multiple signal-to-noise ratios (SNR=-5, 0, 5, 10, 15 and 20 dB).

4.2.2 Auditory System Data

Previously published data from single neurons in the auditory nerve (*n*=214)(Kim, Young 1994), auditory midbrain (Central Nucleus of the Inferior Colliculus, *n*=125)(Rodriguez, Chen et al. 2010), thalamus (Medial Geniculate Body, *n*=88) and primary auditory cortex (*n*=83) (Miller, Escabi et al. 2002) is used to quantify transformations in spectral and temporal selectivity between successive auditory nuclei. Using the measured spectro-temporal receptive fields of each neuron (Fig. 3), the spectral and temporal selectivity are quantified by computing integration times, response latencies, and bandwidths as described previously (Rodriguez, Chen et al. 2010). Sequential changes in selectivity across ascending auditory nuclei are summarized by comparing the neural integration parameters of each auditory structure.

4.2.3 Auditory Model and Hierarchical spiking neural Network (HSNN)

The auditory network consists of a cochlear model stage containing gamma tone filters and envelope extraction (Rodriguez, Chen et al. 2010) followed by a HSNN as illustrated in Fig. 4.1. Several architectural and functional constraints are imposed on the spiking neural network to mirror auditory circuitry and physiology. First, the network contains six layers as there six principal nuclei between the cochlea and cortex. Second, connections between consecutive layers contain both excitatory and inhibitory projections since long-range inhibitory projections between nuclei are pervasive in the ascending auditory system (Oliver 2000, Winer, Saint Marie et al. 1996). Each layer in the network contains 53 excitatory and 53 inhibitory frequency organized neurons per layer. Furthermore, since ascending projections in the central auditory pathway are spatially localized and frequency specific (Levy, Reyes 2012, Oliver 2000, Read, Miller et al. 2008), excitatory and inhibitory connection weights are modeled by co-tuned Gaussian profiles of unspecified connectivity width (Fig. 4.1**e**):

$$w^{E}_{l,m,n} = 1/\sqrt{2\pi\sigma_E^2} \cdot e^{-\left(x_{l,m}-x_{l+1,n}\right)^2/2\sigma_E^2}$$

$$w^{I}_{l,m,n} = 1/\sqrt{2\pi\sigma_I^2} \cdot e^{-\left(x_{l,m}-x_{l+1,n}\right)^2/2\sigma_I^2}$$

where $w^{I}_{l,m,n}$ and $w^{E}_{l,m,n}$ are the inhibitory and excitatory connection weights between the m-th and n-th neuron from layer l and l+1, $x_{l,m}$ and $x_{l+1,n}$ are the normalized spatial positions (0-1) along the frequency axis of the m-th and n-th neurons in layers l and l+1, and σ_I and σ_E are the inhibitory and excitatory connectivity widths (i.e., SD of Gaussian connection profiles), which determine the spatial spread and ultimately the frequency resolution of the ascending connections.

Each neuron in the network consists of a modified leaky integrate-and-fire (LIF) neuron (Trappenberg 2009) receiving excitatory and inhibitory presynaptic inputs. Given a presynaptic spike train from the *m*-th neurons in network layer-*l* ($s_{l,m}(t)$) the desired intracellular voltage of the *n*-th neuron in network layer *l*+1 is obtained as

$$v_{l+1,n}(t) = \sum_m w^E_{l,m,n} \cdot h_{EPSP}(t) * s_{l,m}(t) - \beta \sum_m w^I_{l,m,n} \cdot h_{IPSP}(t) * s_{l,m}(t)$$

where β is a weighting ratio between the injected excitatory and inhibitory currents, $h_{EPSP}(t)$ and $h_{IPSP}(t)$ are temporal kernels that model excitatory and inhibitory post synaptic potentials generated for each incoming spike as an alpha function (Fig. 1**e**, red and blue curves) (Trappenberg 2009). Since central auditory receptive fields often have extensive lateral inhibition beyond the central excitatory tuning area and inhibition is longer lasting and weaker (Miller, Escabi et al. 2002, Rodriguez, Chen et al. 2010) we require that $\sigma_I = 1.5 \cdot \sigma_E$, $\tau_I = 1.5 \cdot \tau_E$, and $\beta = 2/3$, as this produced realistic receptive field measurements (Fig. 4). For simplicity, we use σ and τ interchangeably with σ_E and τ_E, since these determine the overall spectral and temporal resolution of each neuron.

Because the input to an LIF neuron is a current injection, we derived the injected current by deconvolving the IF neuron time-constant from the desired membrane voltage

$$i_{l+1,n}(t) = v_{l+1,n}(t) * h^{-1}(t) + z(t).$$

As demonstrated previously(Escabi, Nassiri et al. 2005), this procedure removes the influence of the cell membrane integration prior to injecting the current in the IF neuron compartment. Above $h(t) = e^{-t/\tau}/C$ is the impulse response of the cell membrane, C is the membrane capacitance, τ, is the membrane time-constant and $h^{-1}(t)$ is the inverse kerenel (i.e., $h(t) * h^{-1}(t) = \delta(t)$ where $\delta(t)$ is the Diract function). Because the EPSP time constant and the resulting temporal resolution of the intracellular voltage are largely influenced by the cell membrane integration, we require that $\tau = \tau_E$. Finally, Gaussian white noise, $z(t)$, is added to the injected current in order to generate realistic spike timing variability (signal-to-noise ratio=15 dB). Upon injecting the current, the resulting intracellular voltage follows $v_{l+1,n}(t) + z(t) * h(t)$ and spikes are generated by the IF model whenever the intracellular voltage exceeds a normalized threshold value (Escabi, Nassiri et al. 2005) which is specified uniquely for each network layer (*l*) as

$$N_l = (V_T - V_r)/\sigma_{V,l}$$

where $V_T = -45$ mV is the threshold voltage, $V_r = -65$ is the membrane resting potentials, and $\sigma_{V,l}$ is the standard deviation of the intracellular voltages for the population of neurons in layer *l*. As demonstrated previously, this normalized threshold represents the number of standard deviations the intracellular activity is away from the threshold activation and serves as a way of controlling the output sensitivity of each network layer. Upon generating a spike, the voltage was reset to the resting potential, a 1 ms refractory period is imposed, and the membrane temporal integration continues.

4.2.4 Decision Model

The neural outputs of the network consist of a spatio-temporal spiking pattern (e.g., Fig. 4.2), represents as a *N*x*M* matrix **R** with elements $r_{n,i}$ where *N*=53 is the number of frequency organized output neurons and *M* is the number of time bins. The number of time bins is dependent on the temporal resolution for each bin, Δt, which is varied between 0.5 – 100 ms. Each response ($r_{n,i}$; $n-\text{th neuron and } i-\text{th time bin}$) is assigned a 1 or 0 value indicating the presence or absence of spikes, respectively.

A Bernoulli Naïve Bayes classifier (McCallum, Nigam 1998) is used to read out the network spike trains and categorize individual speech words. The classified digit (*y*) is the one that maximizes posterior probability for a particular response according to

$$y = \underset{d=\{0\ldots9\}}{\operatorname{argmax}} \prod_{n,i} p_{d,n,i}^{r_{n,i}} \cdot \left(1 - p_{d,n,i}\right)^{1-r_{n,i}}$$

where *d*=0 … 9 are the digits to be identified, $p_{d,n,i}$ is the probability that a particular digit, *d*, generates a spike (1) in a particular spatio-temporal bin (*n*-th neuron and *i*-th time bin).

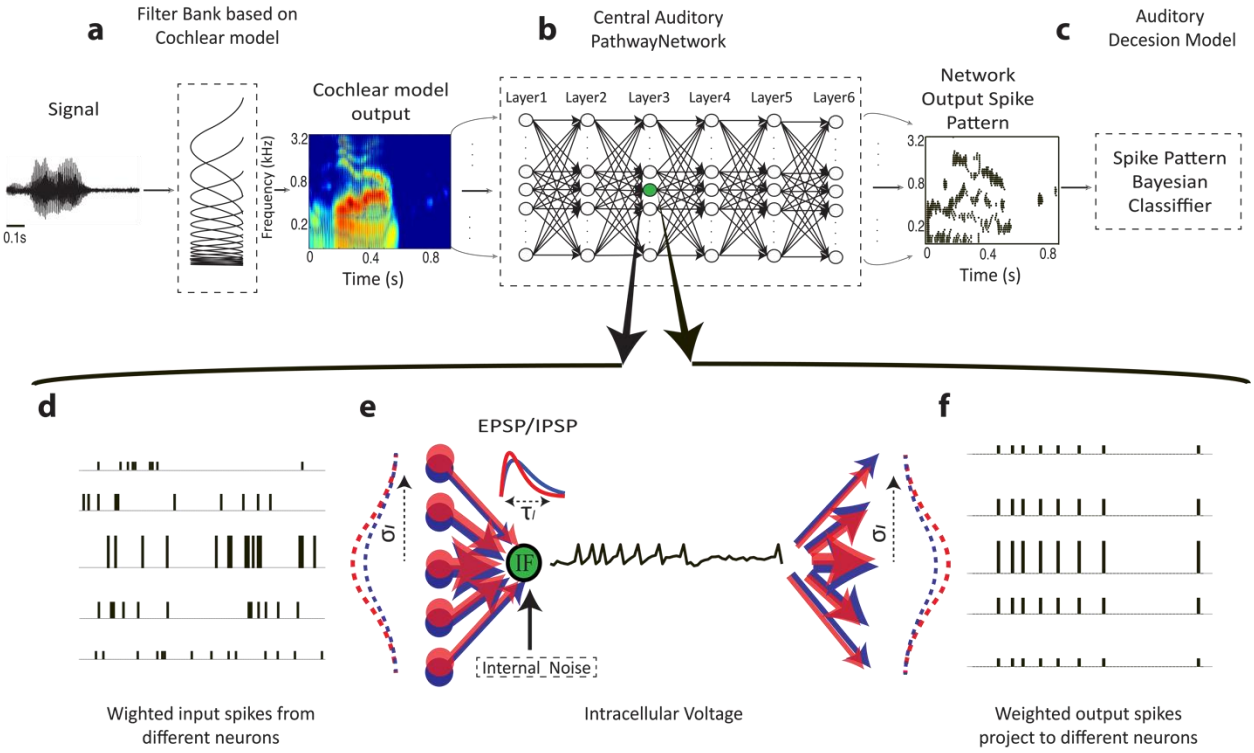


Figure 4.1. Auditory pathway spiking network model. The model consists of (**a**) a cochlear model stage that transforms the sound waveform into a spectrogram (time vs. frequency) output, (**b)** a central network of frequency specifies spiking neurons and a (**c**) Bayesian classifier is used to read the spatio-temporal spike train outputs of the network. Each dot represents a single spike at a particular time-frequency bin. (**d-f**) Zoomed in view illustrating the pattern of convergent and divergent connections between network layers for a single leaky integrate-and-fire (LIF) neuron. (**d-e**) Input spike trains from the preceding network layer are weighted by the Gaussian connectivity weights and project onto each IF neuron. These ascending inputs are spatio-temporally integrated to produce a weighted spiking output for the IF neurons in the subsequent layer (**f**). Excitatory (red) and inhibitory (blue) connectivity weights between inputs and outputs are spatially localized and modeled by Gaussian functions where the divergence and convergence between consecutive layers is controlled by the connectivity width (SD of the Gaussian model, σ_l). Each incoming spike generates and excitatory and inhibitory post-synaptic potentials (EPSP and IPSP, red and blue kernels in **e**). The integration time constant (τ_l) of the EPSP and IPSP kernels can be adjusted to control the temporal integration or smearing between consecutive network layers. Finally, the spike threshold level (N_l) can be independently adjusted to control output firing rates and the overall neuron layer sensitivity.

4.2.5 Network Constraints and Optimization

The primary objective is to determine the spectral and temporal resolution of the network connections as well as the network sensitivity necessary for robust speech recognition. Specifically, we hypothesize that the temporal and spectral resolution of each network layer, as well as the sensitivity, need to be hierarchically organized across network layers in order to maximize speech recognition performance in the presence of noise. We thus optimize three key parameters, the time constant (τ_l), connectivity widths (σ_l), and normalized threshold (N_l) that separately control these functional attributes of the network, where the index l designates the network layer (1-6). Given that spectro-temporal selectivity changes systematically and gradually between auditory nuclei, we constrain the parameters to vary smoothly from layer-to-layer according to the power law rules of Eq. 1. The initial parameters for the first network layer, $\tau_1 =$ 0.4 ms, $\sigma_1 = 0.0269$, and $N_1 = 0.5$, are selected to allow for high-temporal and spectral (~ 1/3 octave) resolution and high firing rates, analogous to physiological characteristics of auditory nerve fibers (Joris, Schreiner et al. 2004, Mc Laughlin, Van de Sande et al. 2007, Kim, Young 1994). We optimize for the three scaling parameters α, λ, and γ, which determine the direction and magnitude of layer-to-layer changes and ultimately the network organization rules for temporal and spectral resolution and network sensitivity.

The optimization is carried using a cross validation global search procedure in which we maximize word accuracy rates (WAR). Initial tests are performed to determine a suitable search range for the scaling parameters and a final global search is performed over the resulting search space ($\alpha = 0.9 - 2.3$, $\lambda = 0.5 - 1.6$ and $\gamma = 0.8 - 1.5$; 0.1 step size for all parameters). For each parameter combination, the network is required to identify the digits in the speech corpus with a ten-alternative forced choice task. For each iteration, we select one utterance from the speech

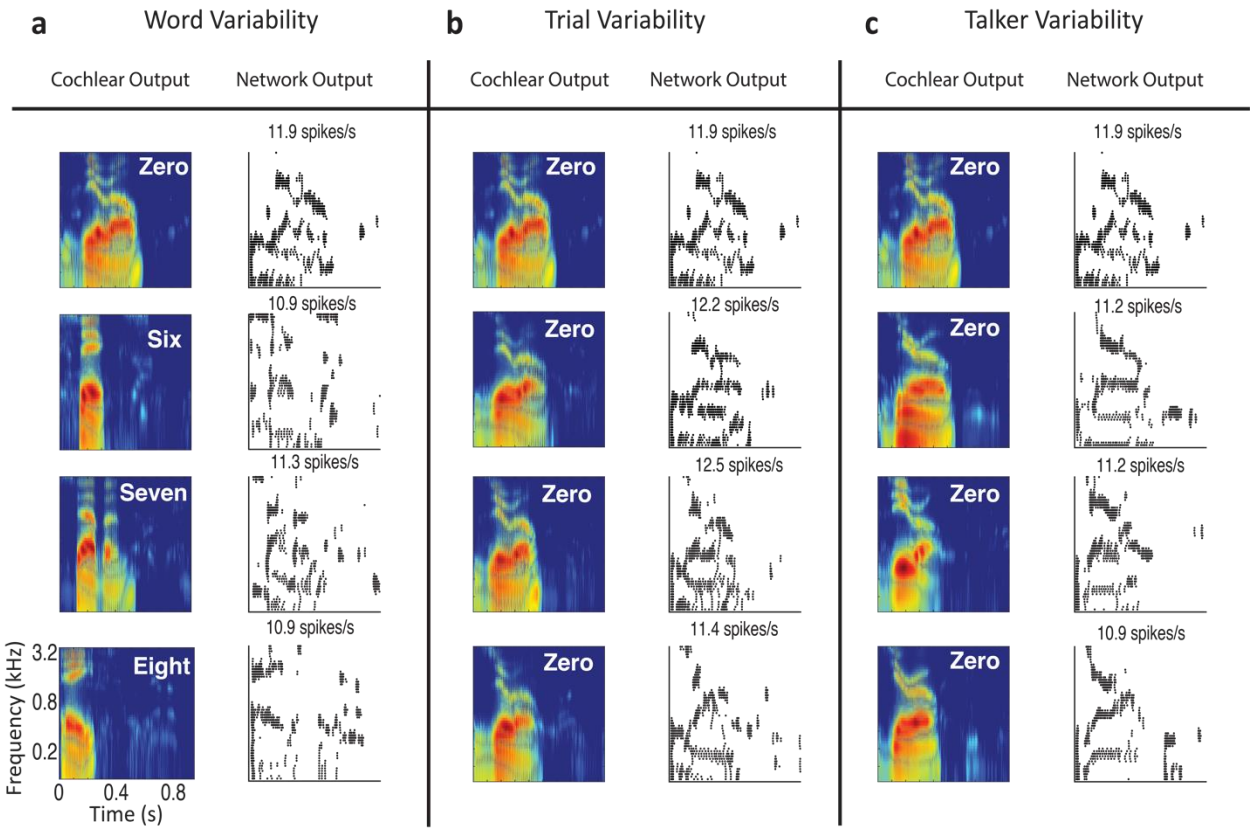


Figure 4.2. Sample cochlear model output and the corresponding spike train outputs of the network illustrate three distinct forms of speech variability under the influence of speech babble noise (**a**, illustrated at 20 dB SNR). Network response pattern for one sample of the words *zero*, *six*, *seven*, and *eight* illustrate output pattern variability that can be used to differentiate words. Each dot corresponds to single action potential. (**b**) By comparison, the network output pattern for multiple trials of the word *zero* are relatively stable when generated by a single talker. (**c**) A somewhat larger amount of output variability is observed when the word *zero* outputs are compared across talkers. In general, average firing rates are quite similar across words, talkers, and trials so that the spatio-temporal spike train pattern is necessary for word classification, analogous to auditory cortex activity (Engineer, Perez et al. 2008a).

corpus (1 of 800 available utterances) for validation and use the remaining utterances (799) to train the model by deriving the Bayesian likelihood functions. The Bayesian classifier is then used to identify the validation utterances and compute WAR for that iteration (either 0 or 100% for each iteration). This procedure is iteratively repeated 800 times over all of the available utterances and the overall WAR is computed as the average over all iterations. This procedure is also repeated for

five distinct signal-to-noise ratios (SNR=-5, 0, 5, 10, 20 dB). Sample curves showing the WAR as a function of scaling parameters and SNR are shown in Fig. 3 (**a** and **b**, shown for 5 and 20dB). The global optimal solution for the scaling parameters is obtained by averaging WAR across all SNRs and selecting the scaling parameter combinations that maximize the global WAR (Fig. 4.3**c**).

4.2.6 Receptive Field and Mutual Information Calculation

To characterized the sequential layer-to-layer transformations performed by the network, we compute spectro-temporal receptive fields (STRFs) and measure the mutual information conveyed by each neuron in the network. First, STRFs is obtained by delivering dynamic moving ripple sounds (DMR), which are statistically unbiased, and cross-correlating the output spike trains of each neuron with the DMR spectro-temporal envelope (Escabi, Schreiner 2002b). For each STRF, we estimate the temporal and spectral resolution by computing the integration time and bandwidths, as described previously (Rodriguez, Chen et al. 2010). Mutual information is calculated by delivering a sequential string of digits (0 to 9) at 5 dB SNR to the network. The procedure is repeated 50 trials with different noise seeds and the spike trains from each neuron are converted into a dot-raster sampled at 2 ms temporal resolution. The mutual information is calculated for each neuron in the network using the procedure of Strong et al.(Strong, Van Steveninck, RR De Ruyter et al. 1998) as described previously (Escabi, Nassiri et al. 2005).

4.2.7 Generalized Linear Model (GLM) Networks

To identify the role of linear and nonlinear receptive field transformations for noise robust coding, we developed two single-layers networks containing GLM neurons(Simoncelli, Paninski et al. 2004) (Fig. 6**a**) that are designed to capture linear and nonlinear transformations of the HSNN.

First, we developed a single-layer LP (linear Poisson) network consisting of model neurons with linear spectro-temporal receptive fields followed by a Poisson spike train generator (Fig. 6**a**). For each output of the optimal network (*m*-th output) we measured the STRF and fitted it to a Gabor model ($STRF_l(t, f_k)$)(Qiu, Schreiner et al. 2003). On average the fitted Gabor model accurately replicated the structure in the measured STRFs and on average accounted for 99% of the STRF variance (range 94-99.9%). The output firing rate of the *m*-th LP model neuron is obtained as

$$\lambda_m(t) = \lambda_0 + G \cdot \sum_{k=1}^{N} S(t, f_k) * STRF_m(t, f_k)$$

where $S(t, f_k)$ is the cochlear model output, * is the convolution operator, *G* is a gain term, and λ_0 is required to assure that the spike rates are strictly positive and the firing maintains a linear relationship with the sound. *G* and λ_0 are chosen so that the average firing rate taken across all output neurons and sounds matches the average firing rate of the optimal network and are strictly greater than zero. The firing rate functions for each channel, $\lambda_m(t)$, are then passed through a nonhomogenous Poisson point process in order to generate the spike trains for each output channel.

Next, we explored the role of nonlinear rectification by incorporating a rectification stage in the LP model. The firing of the m-th neuron in the LNP (linear nonlinear Poisson) network is

$$\lambda_m(t) = G \cdot max\left[0, \sum_{k=1}^{N} S(t, f_k) * STRF_m(t, f_k)\right]$$

where the gain term, *G*, was chosen so that the average firing rate taken across all output neurons and all words matches the average firing rate of the optimal HSNN.

4.3 Results

4.3.1 Hierarchical spiking neural network for Word Recognition

We developed a physiologically motivated hierarchical spiking neural network (HSNN) and trained it on a behaviorally relevant word recognition task in the presence of background noise and multiple talkers. Like the auditory pathway, the HSNN receives frequency organized input from a cochlear stage (Fig. 4.1**a**) and maintains its tonotopic organization through a network of frequency organized integrate-and-fire spiking neurons (Fig. 4.1**b**). For each sound, such as the word "zero", the network produces a dynamic spatio-temporal pattern of spiking activity (Fig. 4.1**b**, right) as observed for peripheral and central auditory structures (Winer, Saint Marie et al. 1996, Loftus, Bishop et al. 2004, Oswald, Schiff et al. 2006). Each neuron is highly interconnected containing frequency specific and co-tuned excitatory and inhibitory connections (Wehr, Zador 2003, Engineer, Perez et al. 2008b, Sachs, Voigt et al. 1983, Delgutte, Kiang 1984) that project across six network layers (Fig. 4.1**b**). Converging spikes from neurons in a given layer (Fig 4.1**d**) are weighted by frequency-localized excitatory and inhibitory connectivity functions and the resulting excitatory and inhibitory post-synaptic potentials are integrated by the recipient neuron (Fig. 4.1**d** and **e**, note the variable spike amplitudes). Output spike trains from each neuron are then weighted by connectivity function, providing the excitatory and inhibitory inputs to the next layer (Fig. 4.1**e, f**). The overall multi-neuron spiking output of the network (Fig. 4.1**b**, right) is then treated as a response feature vector and fed to a Bayesian classifier in order to identify the original sound delivered (Fig. 4.1**c**; see Methods).

Given that key elements of speech such as formants and phonemes have unique spectral and temporal composition that are critical for word identification (Dahl, Yu et al. 2012, Hinton,

Deng et al. 2012), we first test how the spectro-temporal resolution and sensitivity of each network layer contribute to word recognition performance in background noise. We optimize the HSNN to maximize word recognition accuracy in the presence of noise and to identify the network organization of three key parameters that separately control the temporal and spectral resolution and the overall sensitivity of each network layer (l=1 … 6). The neuron time-constant (τ_l), controls the temporal dynamics of each neuron element in layer l and the resulting temporal resolution of the output spiking patterns. The connectivity width (σ_l) controls the convergence and divergence of synaptic connections between consecutive layers and therefore affects the spectral resolution of each layer. Since synaptic connections in the auditory system are frequency specific and localized (Delgutte, Kiang 1984, Elliott, Theunissen 2009, Chi, Gao et al. 1999) connectivity profiles between consecutive layers are modeled by a Gaussian profile of unknown connectivity width parameter (Tan, Zhang et al. 2004) (Fig. 1**e**; specified by the SD, σ_l). Finally, the sensitivity and firing rates of each layer are controlled by adjusting the spike threshold level (N_l) of each IF neuron (Xie, Gittelman et al. 2007). This parameter controls the firing pattern from a high firing rate dense code as proposed for the auditory periphery to a sparse code as has been proposed for auditory cortex (Levy, Reyes 2012, Schneider, Woolley 2013). Because temporal and spectral selectivity vary systematically and gradually across auditory nuclei (Joris, Schreiner et al. 2004, Miller, Escabi et al. 2002, Escabi, Nassiri et al. 2005), we required that the network parameters vary hierarchically and smoothly from layer-to-layer according to (see Methods)

$$\tau_l = \tau_1 \cdot \alpha^{l-1}$$

$$\sigma_l = \sigma_1 \cdot \gamma^{l-1} \qquad \text{(Eqn. 1)}$$

$$N_l = N_1 \cdot \lambda^{l-1}$$

where τ_1, σ_1, and N_1 are the parameters for the first network layer, and the scaling parameters α, λ, and γ determine the direction and magnitude of layer-to-layer changes for each of the three neuron parameters. Scaling values greater than one indicate that the neuron parameter increases systematically across layers, a value of one indicates that the parameter is constant, while a value less than one indicates that the parameter value decreases systematically across layers.

Example speech sounds and the network outputs are shown in Fig. 4.2. Sounds in the optimization and validation corpus consist of digits from zero to nine from eight talkers (TI46 LDC Corpus (Liberman), see Methods). As a task, we require that the network identify the digit that is delivered. Analogous to auditory cortex responses for speech (Engineer, Perez et al. 2008b), the spatio-temporal spiking pattern produced by the network reflect spectro-temporal features from the sound spectrogram and each word produces a distinguishable network output as seen for the words *zero*, *six*, *seven* and *eight* (Fig. 4.2**a**). The network output spiking patterns are relatively consistent from trial-to-trial when words are generated by a single talker as illustrated for the word zero (Fig. 4.2**b**). A somewhat larger amount of spatio-temporal variability is observed in the network output when the same word is generated by multiple talkers as illustrated for the word *zero* (Fig. 4.2**c**). However, various attributes of the response, such as the response timing and neuron channels that are active, remain relatively consistent when examined across multiple talkers. For example, for both multiple or a single talker a lack of activity is observed for neurons between ~2-4 kHz within the first ~100-200 ms of the sound for the word *zero* (Fig. 4.2**b** and **c**).

4.3.2 Optimizing HSNN

To determine the network architecture required for optimal speech recognition in noise and to identify whether such a configuration is essential for noise robust performance, we do grid search for the network scaling parameters (α, λ, and γ) that maximize the network's recognition accuracy

in a ten-alternative forced choice task for multiple talkers (8) and under the influence of speech babble noise at multiple signal-to-noise ratios (SNR=-5, 0, 5, 10, 15, 20 dB). For each sound input, the network spike train outputs are used as response feature vectors and a Bayesian classifier (see Methods) is used to read the network outputs and report the identified digit (*zero* to *nine*). The network word recognition accuracy is shown in Fig. 4.3 as a function of each of the network parameters and SNR. As can be seen, the network is quite sensitive to particular scaling parameter combinations. Although the optimal parameters vary slightly when the network is tested with different SNRs (Fig. 4.3**d**; tested between -5 to 20 dB), the optimal configuration is relatively invariant and performance is comparable for the different optimization SNRs (Fig. 4.3**a**-**c**; **a**=5 dB SNR, **b**=20 dB SNR, **c**=average across all SNRs). Intriguingly, several functional characteristics of the optimal network mirror those seen in the auditory pathway. Like the ascending auditory system (Joris, Schreiner et al. 2004), time constants are scaled in the optimal network (global optimal $\alpha = 1.8$) over more than an order of magnitude between the first and last layer ($1.8^5 = 18.9$ fold increase between the first and last layer) indicating that temporal resolution becomes progressively coarser in the deep network layers. This sequential loss of temporal resolution is accompanied by a subtle change in layer-to-layer network connectivity. The spike thresholds are scaled only slightly across the networks layers (global optimal $\lambda = 1.1;\ 1.1^5 = 1.6$ fold increase between the first and last layer), which is consistent with the idea that neural responses become progressively sparser and more selective along the ascending auditory pathway. By comparison, the optimal connectivity bandwidths do not change across layers ($\gamma = 1.0$), and it is consistent with the idea that spectral resolution remains relatively constant and analogous to changes in

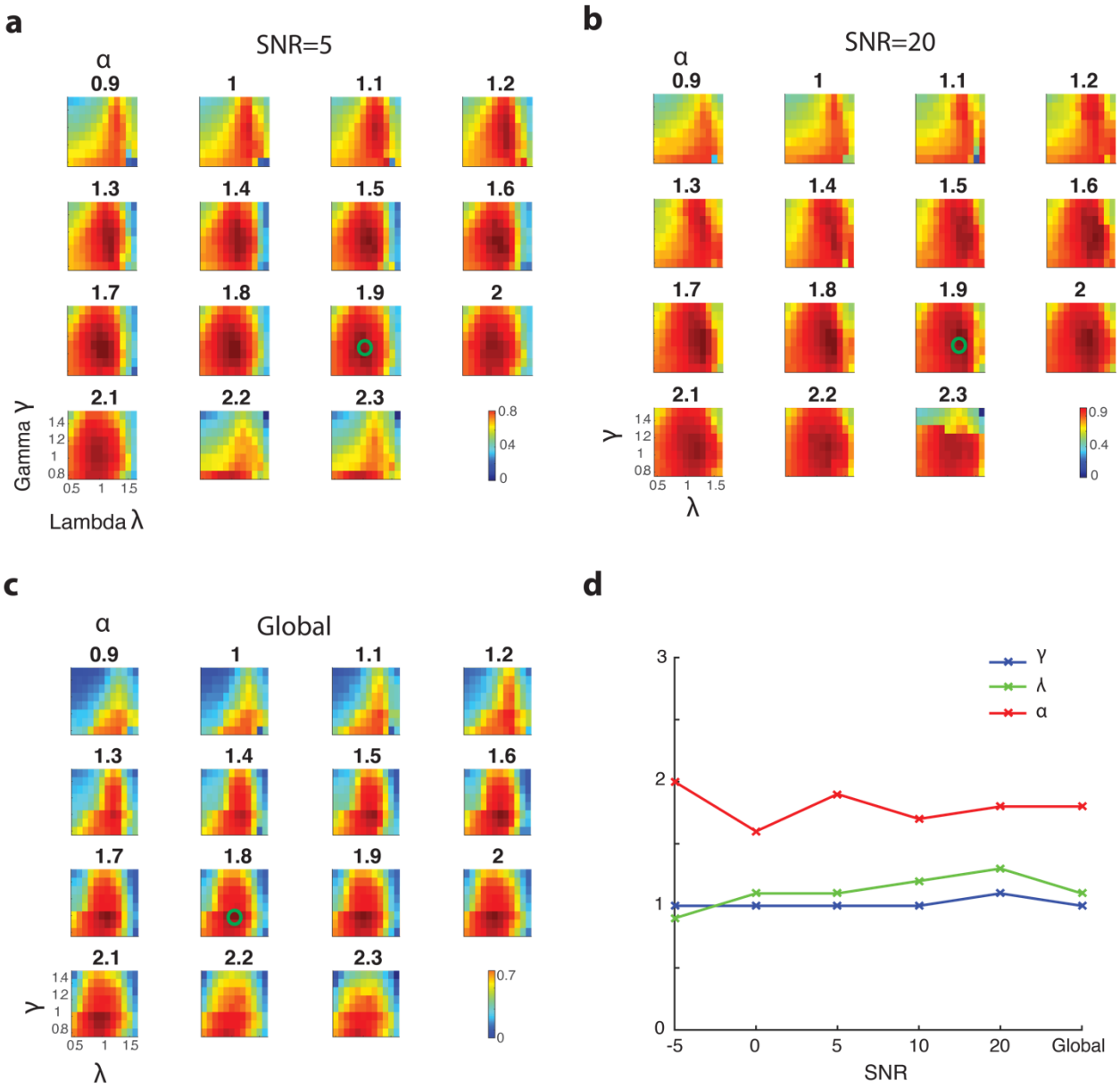


Figure 4.3. Hierarchical scaling is predicted by a global optimal solution that maximizes word recognition accuracy in the presence of background noise (-5, 0, 5, 10, 15 and 20 dB SNR). Cross-validated word recognition accuracy (see Methods) is measured using the network outputs as a function of the three scaling parameters (σ_l, τ_l, and N_l). Word recognition accuracy curves are shown at 5 and 20 dB SNR (**a** and **b**, respectively) as well as for the global solution (**c**, average accuracy between -5 and 20 dB SNR). In all cases shown, word recognition accuracy curves are tuned for the different scaling parameters and exhibit a similar optimal solution (green circles). (**d**) The resulting optimal parameters are relatively stable across SNRs (global optimal $\alpha = 1.8$, $\lambda = 1.1$, and γ=1.0).

spectral selectivity observed along the ascending auditory pathway (Miller, Escabi et al. 2002, Mc Laughlin, Van de Sande et al. 2007, Rodriguez, Chen et al. 2010).

4.3.3 Hierarchical Organization of Auditory System versus Optimal HSNN

The scaling parameters of the optimal network indicate a substantial loss of temporal and minimal change in spectral resolution across network layers. This prompts us to ask how feature selectivity changes along the network layers and whether the sequential transformations in spectral and temporal selectivity are essential for optimal speech recognition in noise. To measure the sequential transformations in acoustic processing, we first measure spectro-temporal receptive fields (STRFs) of each neuron in the network (see Methods). Example STRFs are shown for two selected frequencies for the six network layers (Fig. 4.4**a**; 1.5 and 3 kHz). As a comparison, example STRFs from the auditory nerve (Kim, Young 1994), midbrain (inferior colliculus) (Rodriguez, Chen et al. 2010), thalamus and auditory cortex (Miller, Escabi et al. 2002) of cats are also shown in Fig. 4.4**e**. STRFs are substantially faster in early network layers lasting only a few milliseconds and mirroring STRFs from the auditory nerve, which have relatively short latencies and integration times (Kim, Young 1994, Rodriguez, Chen et al. 2010). STRFs have progressively longer integration times (paired t-test with Bonferoni correction, $p<0.01$; Fig. 4.4**b**) and latencies (paired t-test with Bonferoni correction, $p<0.01$; Fig. 4.4**c**) along the network layers, while bandwidths increase only slightly from the first to last layer (paired t-test with Bonferoni correction, $p<0.01$; Fig. 4.4**d**). These sequential transformations mirror changes in temporal and spectral selectivity seen between the auditory nerve (Kim, Young 1994, Rodriguez, Chen et al. 2010), midbrain(Rodriguez, Chen et al. 2010), thalamus and ultimately auditory cortex (Miller, Escabi et al. 2002, Depireux, Simon et al. 2001) (Fig. 4.4**e**-**h**). As for the auditory network model,

integration times (Fig. 4.4f) and latencies (Fig. 4.4g) increase systematically and smoothly between peripheral and central auditory levels (paired t-test with Bonferoni correction, $p<0.01$) while bandwidths show a small but significant increase between the auditory nerve and cortex (paired t-test with Bonferoni correction, $p<0.01$), analogous to results from the computational network. Furthermore, while STRFs in the early network layers are dominated by excitatory domains, analogous to those observed in the auditory nerve (Kim, Young 1994), STRFs in deep layers exhibit stronger and more varied inhibition / suppression, have increasingly more complex structure, thus mirroring hierarchical changes in selectivity seen in the central auditory system (Miller, Escabi et al. 2002, Depireux, Simon et al. 2001, Sen, Theunissen et al. 2001).

4.3.4 Optimal HSNN versus High Resolution HSNN

It is intriguing that the optimal network solution requires a large sequential loss of temporal and only a subtle loss of spectral resolution across network layers. First the overall trend and magnitude of the observed changes mirrors changes in selectivity seen in the ascending auditory system. Second, it is surprising that optimal performance is achieved by sequentially degrading temporal resolution across network layers, as this ought to limit the transfer of acoustic information across the network. One plausible hypothesis is that such a sequential decrease in resolution is necessary to extract invariant acoustic features in speech while rejecting noise and fine details in the acoustic signal that may contribute in a variety of hearing tasks (e.g., spatial hearing, pitch perception etc.),

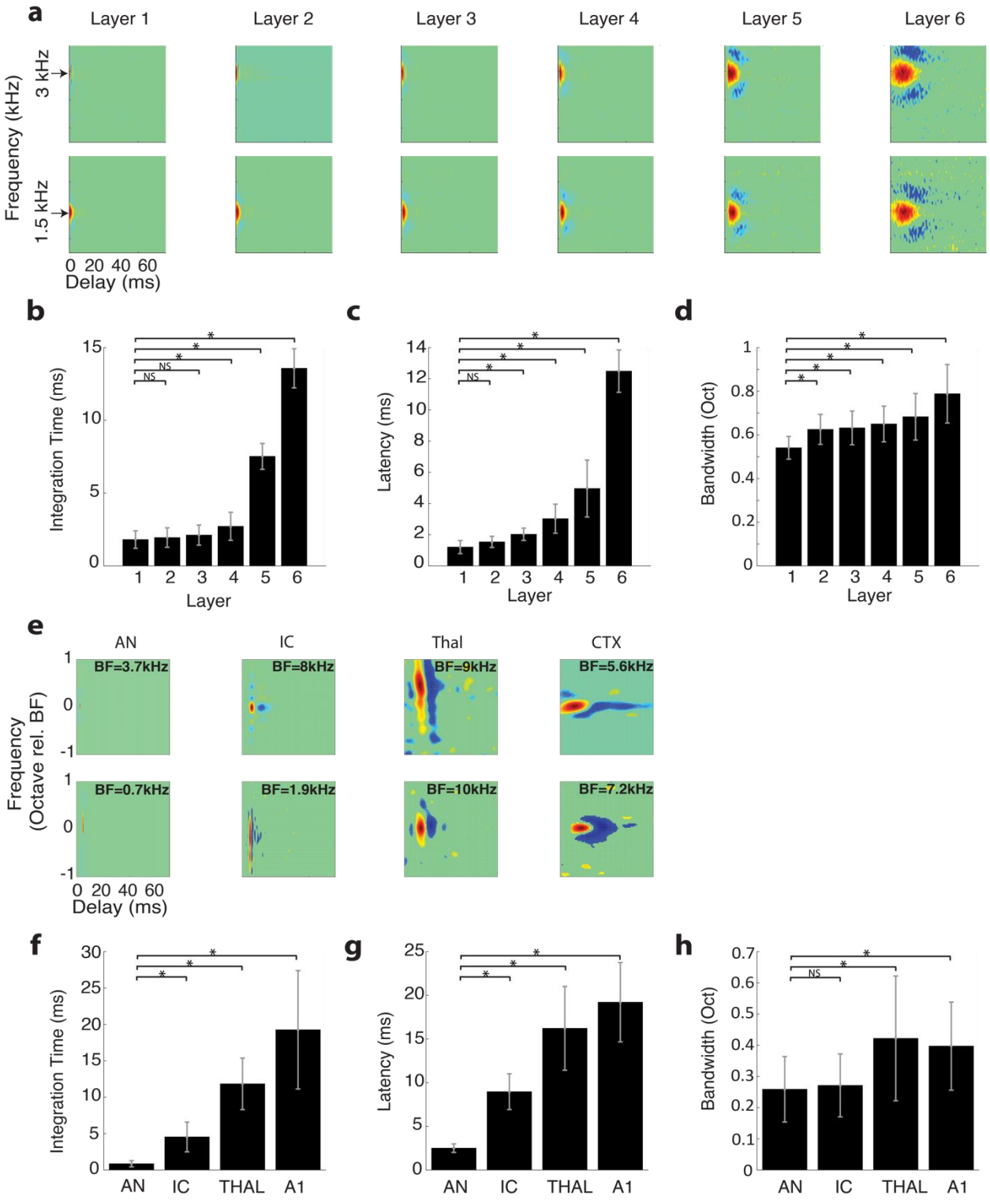


Figure 4.4. Receptive field transformations of the optimal hierarchical network mirror ascending transformations in feature extraction observed in the auditory pathway. (**a**) Example measured spectro-

temporal receptive field (STRF) for the optimal network shows systematically changes between consecutive network layers. All STRFs are normalized to the same color scale where red indicates an increase in activity (excitation), blue a decrease in activity (inhibition or suppression), and green tones a lack of activity. In the early network layers STRFs are relatively fast with short duration and latencies, and relatively narrowly tuned. STRFs become progressively slower, slightly broader, and have longer and more varied patterns of inhibition across the network layers, mirroring changes in spectral and temporal selectivity seen in the ascending auditory pathway. The measured (**b**) integration times, (**c**) latencies, and (**d**) bandwidths increase across the six network layers. (**e**) Examples STRFs from the auditory nerve (AN)(Kim, Young 1994), inferior colliculus (IC)(Rodriguez, Chen et al. 2010), thalamus (MGB) and primary auditory cortex (A1)(Miller, Escabi et al. 2002) become progressively longer and have progressively more complex spectro-temporal sensitivity along the ascending auditory pathway. Average integration times (**f**), latencies (**g**) and bandwidths (**h**) between AN and AC follow similar trends as the optimal network model (**b**-**d**). Asterisks (*) designate significant comparisons (t-test with Bonferoni correction, $p<0.01$) relative to layer 1 for the optimal network (**b**-**d**) or relative to the auditory nerve for the neural data (**f**-**h**) while error bars designate SD.

but ultimately don't contribute to speech recognition performance. This may be expected as human listeners require a limited set of temporal and spectral cues for speech recognition (Elliott, Theunissen 2009, Chi, Gao et al. 1999) and can achieve high recognition performance even when spectral and temporal resolution is degraded (Shannon, Zeng et al. 1995, Drullman, Festen et al. 1994). We thus investigate the above hypothesis by comparing the optimal network performance against a high-resolution network that lacks scaling ($\alpha = 1$, $\lambda = 1$, and $\gamma = 1$) and for which there should be a minimal loss of acoustic information across layers. Unlike the optimal network, STRFs from the high-resolution network are relative consistent and change minimally across layers (Fig. 4.5), which supports the idea that spectrotemporal information propagates across the high-resolution network with minimal processing.

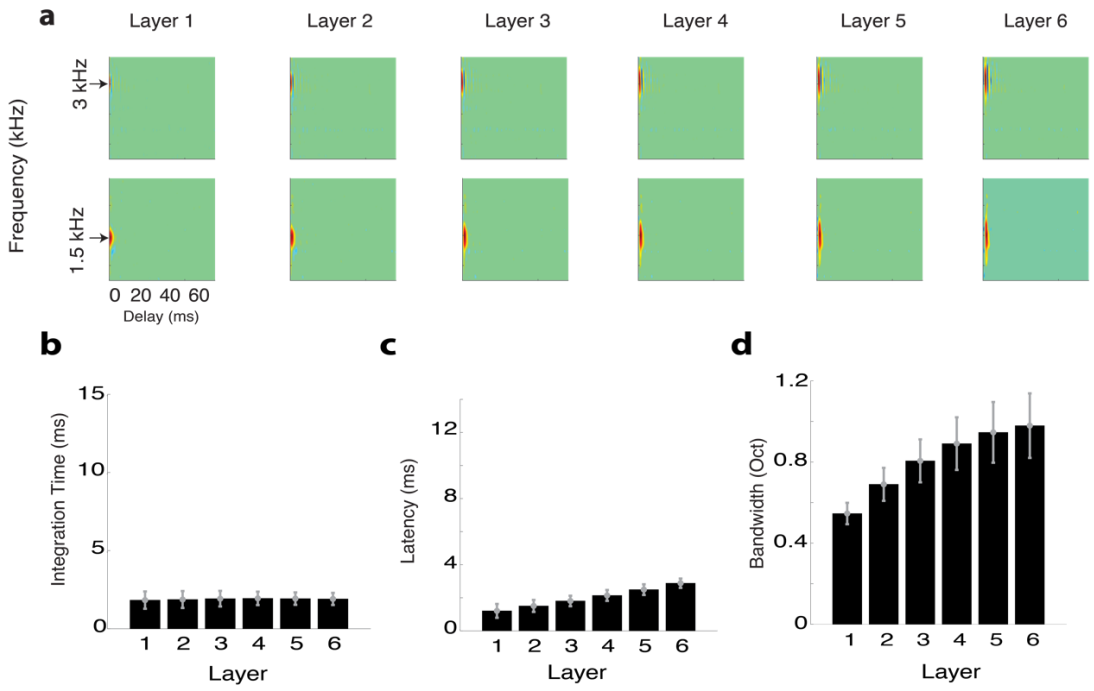


Figure 4.5. Receptive field transformations of the high-resolution network indicate that spectro-temporal information propagates with minimal processing across network layers. (**a**) Example spectro-temporal receptive field (STRF) measured for the optimal network maintain high-resolution and change minimally across network layers. Unlike the optimal network, the measured (**b**) integration times and (**c**) latencies change minimally and are relatively constant across the six network layers. (**d**) Bandwidths, by comparison, increase slightly across the six network layers and follow a similar trend as the optimal HSNN. The figure format follows the same convention as in Figure 4.4.

Example Bayesian posterior time-frequency histograms (average firing probability across all excerpts of each sound at each time-frequency bin) are measured at 5 dB SNR and are shown for the words “three”, “four”, “five” and “nine” for both the high-resolution (Fig. 4.6**a**) and optimal network (Fig. 4.6**b**) configurations along with selected spiking outputs from a single talker. Intriguingly, the Bayesian posterior for the high-resolution network are highly blurred in both the

temporal and spectral dimensions and have similar structure for the example words (Fig. 4.6**a**, right panels). This is also seen in the individual network outputs where the high-resolution network produces a dense and saturated firing pattern (Fig. 4.6**a)** that lacks the detailed spatio-temporal pattern seen in the optimal network (Fig. 4.6**b)**. Despite the fact that this network has high spectro-temporal integration resolution key spectral and temporal details that are evident in the cochlear model output, and which presumably are essential for speech recognition, are highly distorted in the overall network output. By comparison, the optimal network appears to preserve and even accentuate key acoustic elements such as temporal transitions for voice onset timing and spectral resonances (formants), while at the same time reject the background noise (Fig. 4.6**b**, right panels). Consequently, the word recognition accuracy of the optimal network is significantly higher than the high-resolution network for all of the tested SNRs (Fig. 4.6 **c**; $p<0.001$, t-test with Bonferroni correction). On average, there is a 27.6 % improvement in the word accuracy rates for the optimal scaling over the high-resolution network.

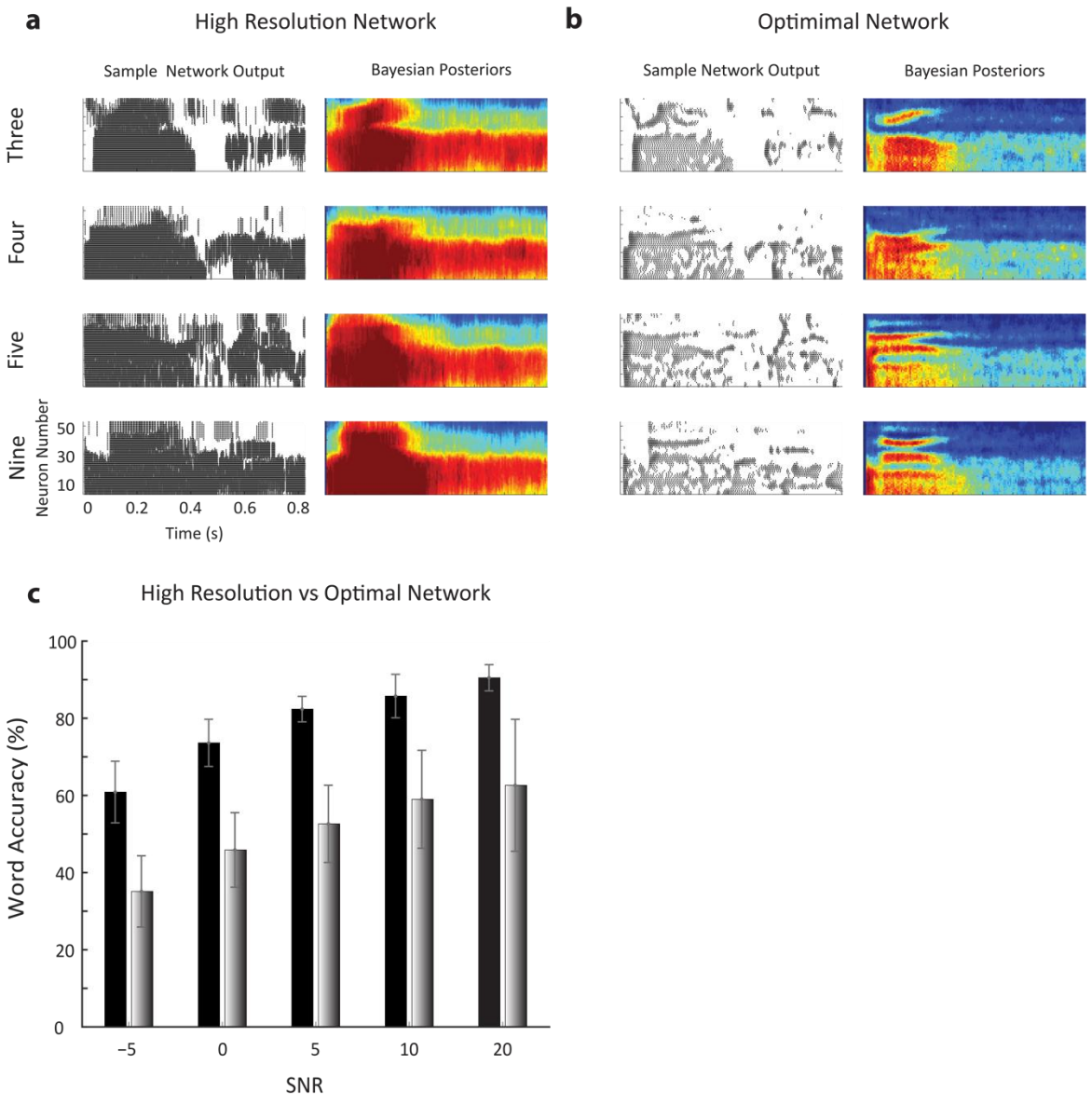


Figure 4.6. Optimal hierarchical network substantially outperforms a high-resolution network designed to preserve incoming acoustic information. Sample network spike train outputs and Bayesian posterior histograms for the words *three*, *four*, *five*, and *nine* are shown for the (**a**) high-resolution and (**b**) optimal hierarchical network at 5 dB SNR. The Bayesian posterior histograms correspond to the average probability of firing at each time-frequency bin for each digit. Firing patterns and Bayesian posterior of the high-resolution network are saturated and spatio-temporally blurred compare to the hierarchical network. (**b**) Details such as spectral resonances (e.g., formants) and temporal transitions resulting from voicing onset are accentuated in the hierarchical network output. (**c**) The hierarchical network outperforms the high-

resolution network in the word recognition task at all tested SNRs (black=hierarchical; gray=high-resolution).

4.3.5 Acoustic Transformation Between Consecutive Network Layers (Optimal vs. High Resolution)

The consequences of the scaling characteristics of each network can be identified by measuring how word recognition accuracy changes along the layers. By considering the output of each network layer and applying the Bayesian classifier to measure sequential changes in performance between layers we measure recognition accuracy for each layer. In the optimal network, word recognition accuracy systematically increases along layers with an average improvement of 15.5% between the first and last layer when tested at 5 dB SNR ($p<0.001$, t-test; Fig. 4.7**a**, blue; 13.7% average improvement across all SNRs). By comparison, for the high-resolution network, performance degrade sequentially across layers with an average decrease of 19.8% between the first and last layer ($p<0.001$, t-test; Fig. 4.7**a**, red; 18.1 % average reduction across all SNRs). This suggests that the optimal network is capable of sequentially extracting high-level acoustic features and enabling an invariant representation that enhances word recognition performance. By comparison, background noise is pervasive in the high-resolution network, which resulted in a reduction in performance across network layers.

Although the classifier performance gain benefits from the optimal scaling configuration, a similar trend is not observed for the transfer of acoustic information across network layers. First, firing rates decrease systematically across layers for the optimal scaling network consistent with a sparser output representation (Fig. 4.7**b**, blue), as proposed for deep layers of the auditory pathway (Schneider, Woolley 2013, Chen, Read et al. 2012, Hromádka, DeWeese et al. 2008). By comparison firing rates are relatively stable across layers for the high-resolution network

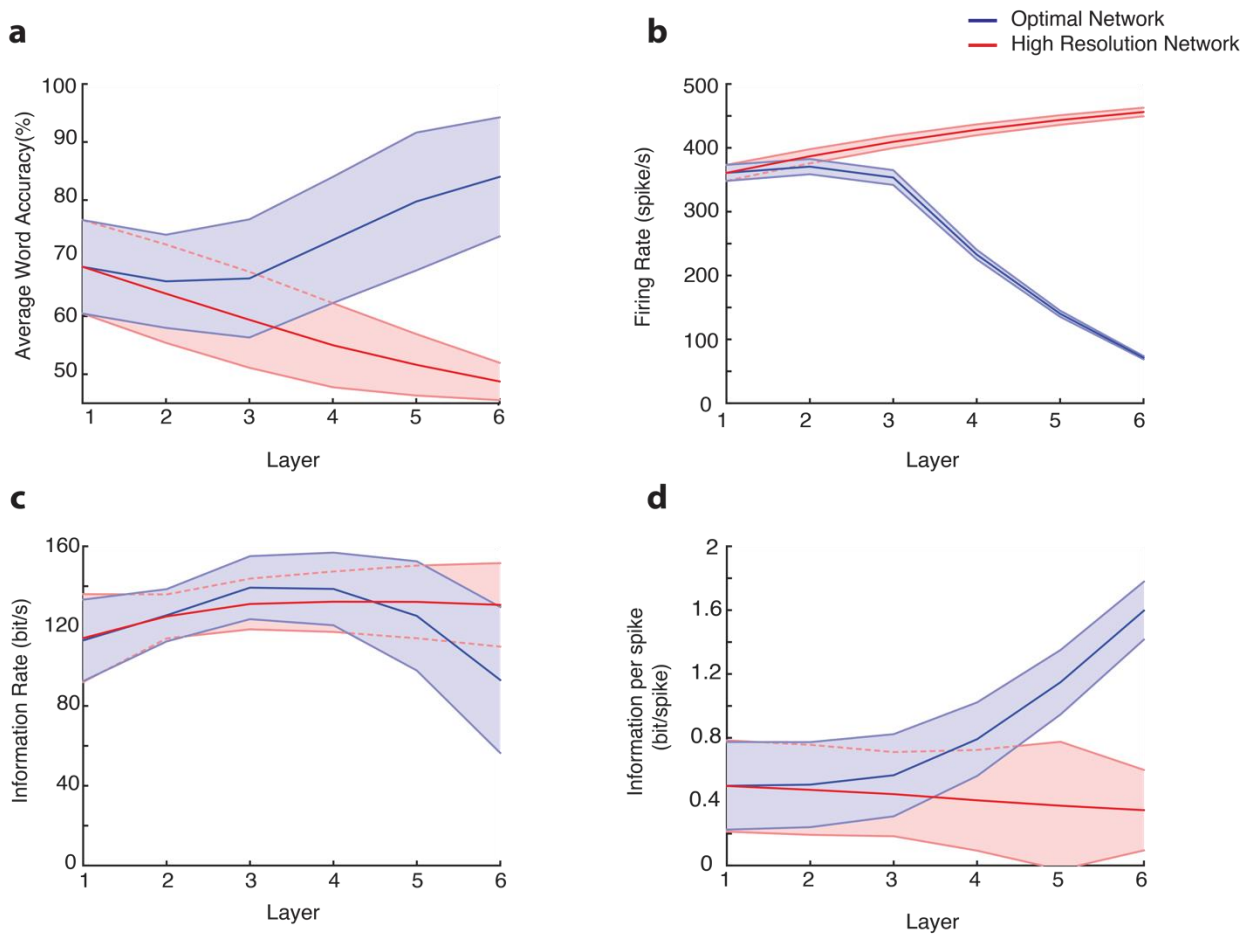


Figure 4.7. Sequential acoustic transformation between consecutive network layers enhances word recognition performance and robustness of the optimal hierarchical network. (**a**) The average word accuracy at 5 dB SNR systematically increases across network layers for the optimal network (**a**, blue) whereas the high-resolution network exhibits a systematic reduction in word recognition accuracy (**a**, red). For the high-resolution network average firing rates (**b**, red), information rates (**c**, red), and information per spike (**d**, red) are relatively constant across layers indicating minimal transformations of the incoming acoustic signal. In contrast, average firing rates (**b**, blue) and information rates (**c**, blue) both decrease between the first and last network layers of the optimal network, consistent with a sequential sparsification of the response and a reduction in the acoustic information encoded in the output spike trains. However, the information conveyed by single action potentials (**d**, blue) sequentially increase between the first and last layer so that individual action potentials can be quite informative. Continuous curves show the mean whereas error contours designate the SD.

(Fig. 4.7**b**, red). For both networks, we measure the average mutual information (see Methods) in the presence of noise (5 dB) to identify how incoming acoustic information is sequentially transformed from layer-to-layer. While the rate of information transmission is sequentially conserved for the high-resolution network across layers (Fig. 4.6**c**, red), the information transmission rate (i.e., bits / sec) decrease between the first and last layer for the optimal scaling network (Fig. 4.7**c**, blue). By comparison, information rates are relatively constant for the high-resolution network indicating that acoustic information is conserved across network layers despite the observed reduction in recognition accuracy (Fig. 4.7**a**, red). This paradoxical result suggests that the layer-to-layer increase in word recognition performance for the optimal network is accompanied by a loss of total acoustic information in the deep network layers. By comparison, recognition performance suffers for the high-resolution network even though it tends to preserve the acoustic information from layer-to-layer. We next measure the average information conveyed by individual action potentials to determine how acoustic features are represented by individual precisely timed spikes. Although information rates decrease for the optimal network across layers, conveyed information by single action potentials is higher and increase along the layers for the optimal network (Fig. 4.7**d**, blue). This contrasts the high-resolution network where information per spike remain relatively constant across layers (Fig. 4.7**d**, red). Despite the observed reduction in the total transmitted information for the optimal network between the first and last layer, this finding suggests action potentials become more informative from layer-to-layer. Taken together with the changes in spectro-temporal selectivity (Fig. 4.6), the findings are consistent with the general hypothesis that the optimal network produces a sparse code in which invariant acoustic features are represented with isolated spikes while selectively rejecting noise and acoustic features are not pertinent to the speech recognition task. By comparison, the high-resolution network has a

tendency to preserve incoming acoustic information, including the noise, and thus suffers in recognition performance.

4.3.6 Human Performance and Network Hierarchy and Nonlinearity Properties

We next asked whether the sequential layer-to-layer transformations of the optimal HSNN are required for robust coding of speech. Hypothetically, its plausible that similar performance could be achieved with a single layer network as long as each neuron accounts for the overall network receptive field transformations. To test this, we developed single-layer networks consisting of generalized linear model neurons[31] with either a linear receptive field and Poisson spike train generator (LP network) or a linear receptive field and nonlinear stage followed by Poisson spike train generator (LNP network) (Fig. 4.8**a**; see Methods). The performance of the LP network, which accounts for the linear transformations of the optimal HSNN, was on average 21.7% lower than the optimal HSNN indicating that nonlinearities are critical to achieve high word recognition accuracy (Fig. 4.8**b**). Its plausible that this performance disparity can be overcome by incorporating a nonlinearity that models the rectifying effects in the spike generation process of neurons (LNP network). Doing so improves the performance to within 2.1% of the optimal HSNN when there is little background noise (SNR=20 dB, 85.6 % for optimal HSNN versus 82.5 % for LNP network). However, the performance degraded when background noise was added when compared to the optimal HSNN, with an overall performance reduction of 13.8 % at -5 dB SNR (58.4 % for optimal HSNN versus 44.6 % for LNP network).

The robustness of each network was next examined by comparing the performance of each model against human performance trends. For each condition, we measured the relative accuracy change (RAC) between the model and human performance (Methods, Fig. 4.8**c)**. The RAC of the optimal HSNN was near zero with a small reduction in RAC of only 3.9% at -5 dB SNR. Thus,

the optimal HSNN follows a similar trend as humans across background noise levels. By comparison, both the LP and LNP performance diverged from human performance with increasing background noise with an overall RAC reduction of 22.2 % and 15.6% at -5 dB SNR, respectively. Thus, in contrast to the optimal HSNN trends which mirrors human data, the LP and LNP network performance diverged from the human trend with increasing background noise.

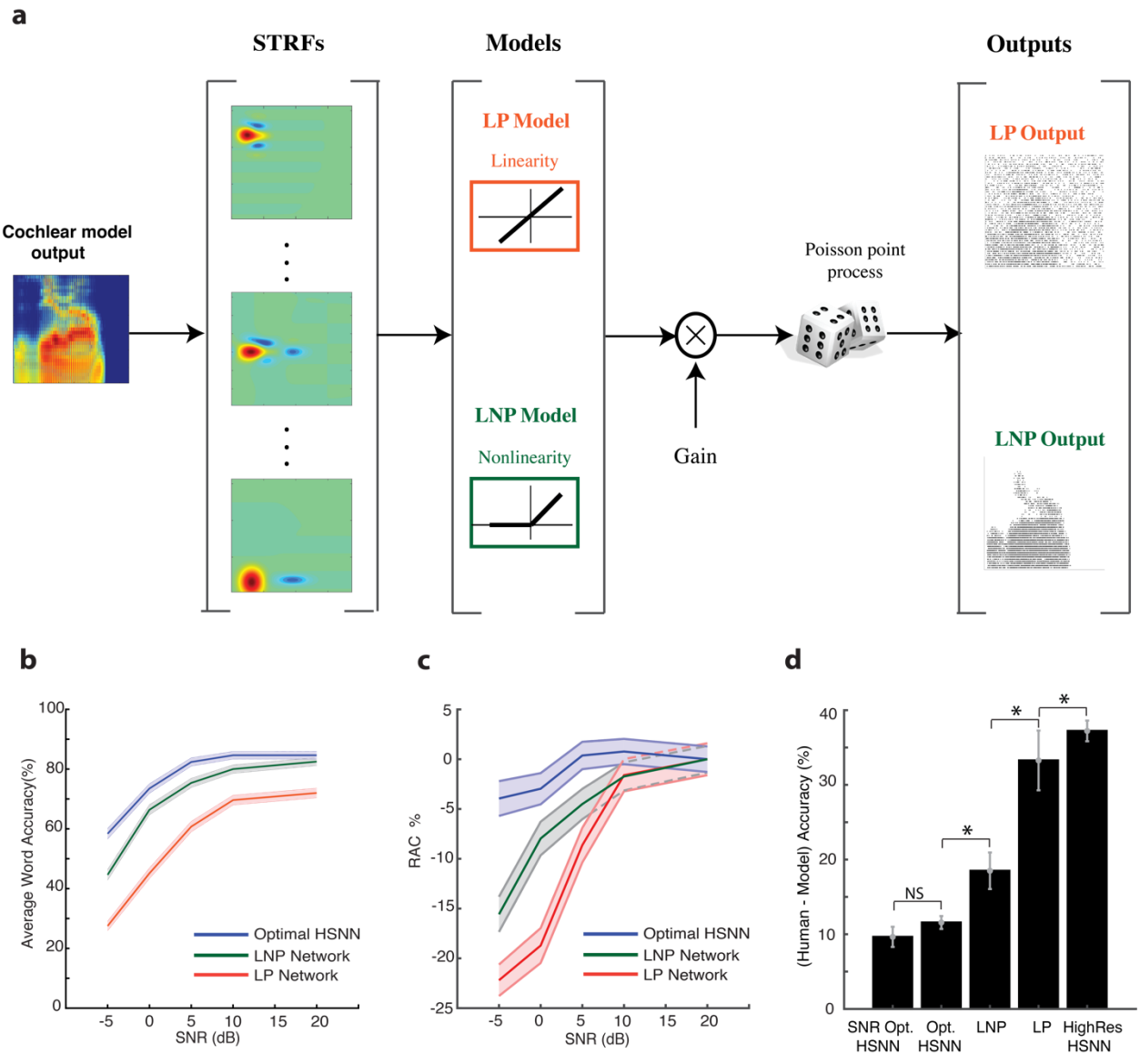


Figure 4.8. Optimal HSNN enhances robustness and outperforms single-layer generalized linear model networks with matched linear and nonlinear receptive field transformation. (a) Linear STRFs obtained at

the output of the HSNN are used as to model the linear receptive field transformation of each neuron (see Methods). The LP network consists of an array of linear STRFs followed by a Poisson spike generator. The LNP network additionally incorporates a rectifying output stage following each STRF. (b) The optimal HSNN outperformance the LP network with an average performance improvement of 21.7% across SNRs. Nonlinear output rectification in the LNP network improves the performance to within 2% of the HSNN at 20 dB SNR. However, the average LNP performance was 7% lower than the optimal HSNN and performance degraded systematically with increasing noise levels (13.75 % performance reduction at -5 dB SNR) demonstrating enhanced robustness of the optimal HSNN. (**c**) The relative change in accuracy ($RAC=(A_{model}-A_{human}) - (A^{20dB}_{model}-A^{20dB}_{human})$) was used to measure the divergence of each model across SNR when compared against human accuracy rates (Crandell, Smaldino 2000). An RAC of 0 across SNRs indicates that the model performance follows a similar noise robust trend when compared to humans. For the optimal HSNN, RACs were near zero across SNRs. RACs diverged substantially relative to human accuracy rates with increasing SNR for the LP and LNP networks. (**d**) Average accuracy difference between human and model data ($A_{human}-A_{model}$). Average performance of the SNR optimal (optimized for each SNR) and optimal HSNN (optimized across all SNRs) are within ~10 % of the human word accuracy rates. The LNP (18.5 %), LP (33.3%) and high-resolution HSNN (37.2%) performances are substantially lower relative to humans. Asterisks designate significant differences ($p<0.05$, t-test with Bonferroni correction) and error bars designate SEM.

The average performance of each network was also compared against human word recognition accuracy. The accuracy for the optimal and SNR optimal HSNNs are not significantly differences when compared against human accuracy rates with an average reduction of 9.7% and 11.5%, respectively ($p>0.05$, t-test). Furthermore, the optimal HSNN outperformed all other models tested. The LNP, LP, and high-resolution HSNN exhibited a rank order reduction in performance relative to human accuracy (18.45 %, 33.27%, 37.22% respectively; $p<0.05$, t-test with Bonferroni Correction).

Overall, the findings indicate that although the linear and nonlinear receptive field transformations both contribute to the overall network performance, the sequential layer-to-layer

transformations carried out by the optimal HSNN are critical for maintaining a noise robust representation that mirrors human performance trends.

4.3.7 Optimal Spiking Timing Resolution

Obviously, the degradation of spectro-temporal resolution and total information across network layers for the optimal network is reflected in the network outputs. One might expect tradeoffs between spike-timing resolution and recognition accuracy, as previously demonstrated when "reading out" neural activity in auditory cortex (Loftus, Bishop et al. 2004, Wehr, Zador 2003, Simoncelli, Paninski et al. 2004). We investigate this possibility by varying the temporal resolution of the network output spike trains while measuring the word classification performance at multiple SNRs (see Methods). Not surprisingly the classifier performance decrease with reducing SNR, however in all instances an optimal spike timing resolution could be identified within the vicinity of 4-14 ms for the optimal network (Fig. 4.9**a** and **b**). By comparison, the high-resolution network requires a higher temporal resolution of ~2 ms to achieve maximum performance (46.6% accuracy across all SNRs; Fig. 4.9**c**), which is ~ 31.8% on average lower than the optimal network (78.4 % accuracy across all SNRs, Fig. 4.9**a**). At high SNR, the classifier is relatively insensitive to the temporal resolution and the performance exhibits relatively broad tuning for the optimal network (Fig. 4.9**a** and **b**). The performance curve is substantially more tuned across temporal resolutions for low SNR which is indicating a critical resolution of 4-14 ms for conveying speech related information by the optimal network output. Presumably, employing too high of a resolution allows high frequency noise to permeate through the network, which decreases the performance. By comparison, limiting the temporal resolution beyond the optimal value removes key temporal information present in the spike trains further limiting performance. The network is far more sensitive to these effects for low SNR and the classifier needs to be

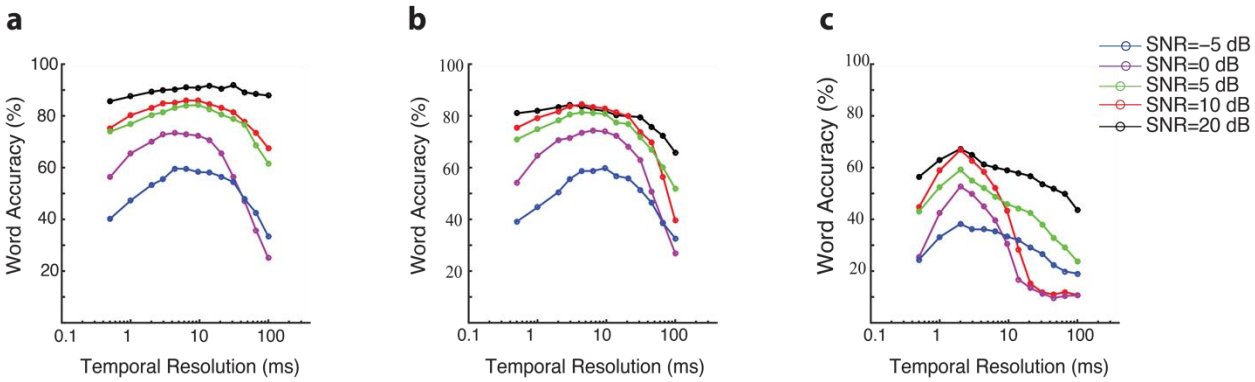


Figure 4.9. Optimal temporal resolution that maximize word recognition accuracy in noise. (**a**) Word accuracy rate as a function of spike train temporal resolution (bin widths 0.5-100 ms) and SNR (-5 to 20 dB) for the optimal (**a**) and high-resolution networks (**c**). Each curve is computed by selecting the optimal scaling parameters for each SNR and measuring the word accuracy rate from the network outputs at multiple temporal resolutions. (**b**) Same as (**a**), except that global optimal scaling parameters is used for all SNRs tested. The temporal resolution that maximizes the word accuracy rate of the global optimal model is 6.5 ms. (**c**) Word accuracy rate as a function of temporal resolution and SNR for the high-resolution network. The optimal temporal resolution for the high-resolution model is 2 ms.

precisely tuned to read out the spike trains and maximize performance when speech babble is present. Taken across all SNRs, the optimal temporal resolution that maximized discrimination performance is 6.5 ms for the optimal network (78.4% average accuracy across all SNRs), which is comparable to the spike timing resolution required for optimal speech and vocalizations recognition in auditory cortex (Loftus, Bishop et al. 2004, Simoncelli, Paninski et al. 2004).

4.4 Discussion

The results demonstrate that the hierarchical organization of the ascending auditory system forms a near optimal strategy for feature extraction that maximizes sound recognition performance and is relatively impervious to noise. Upon optimizing the network organization on a behaviorally relevant word recognition task, the HSNN achieves high recognition accuracy and follows a

similar noise robust trend that is within ~10% of human performance by sequentially refining the spectral and temporal selectivity from layer-to-layer. Similar noise robustness is not replicated with conventional receptive field based networks even when the receptive fields capture the linear integration of the optimal HSNN and a threshold nonlinearity was imposed. The sequential nonlinear transformations of the optimal HSNN preserve critical acoustic features for speech recognition while simultaneously discarding acoustic noise not relevant to the sound recognition task. These transformations mirror changes in selectivity along the ascending auditory pathway, including an extensive loss of temporal resolution (Joris, Schreiner et al. 2004), slight loss of spectral resolution (Miller, Escabi et al. 2002, Mc Laughlin, Van de Sande et al. 2007, Rodriguez, Chen et al. 2010), and increase in sparsity (Schneider, Woolley 2013, Levy, Reyes 2012). Thus, the orderly arrangement of receptive fields and sequential nonlinear transformations of the ascending auditory pathway are critical to achieve a noise robust code.

The theoretical findings suggest that the hierarchical scaling organization of the ascending auditory system forms a near optimal strategy for feature extraction that maximizes sound recognition performance which is relatively impervious to noise. The network achieves high recognition accuracy by sequentially refining the spectral and temporal selectivity from layer-to-layer and sequentially discards detailed acoustic information not relevant to the sound recognition task. The findings have major implications for theories of feature extraction and sound segregation by the auditory system as well as for sound and speech recognition applications in the presence of noise. Not only does the optimal network replicate several hierarchical auditory system trends, but also the findings indicate that this organization is essential for acoustic feature extraction and creating invariant neural representations that are impervious to babble noise, one of the most sophisticated noise. Moreover, the model helps us to investigate why some characteristics of

auditory system, such as degrading temporal resolution are necessary for robust speech recognition.

By maximizing word recognition performance in noise for multiple talkers, the model replicates several well-identified characteristics of the ascending auditory pathway, including an extensive loss of temporal resolution (Joris, Schreiner et al. 2004), slight loss of spectral resolution (Mc Laughlin, Van de Sande et al. 2007, Joris, Schreiner et al. 2004, Rodriguez, Chen et al. 2010, Miller, Escabi et al. 2002), increase in sparsity and increasing amount of spectro-temporal inhibition and/or suppression and receptive field complexity (Miller, Escabi et al. 2002, Rodriguez, Chen et al. 2010, Kim, Young 1994, Sen, Theunissen et al. 2001, Clopton, Backoff 1991). Modeling auditory system properties help us to investigate more in these properties and their effects not only on feature selection for noise robust recognition, but also on quantity and quality of information transferring from one stage of auditory system to the other stage. It is interesting that although sparsity of neurons' responses increases, and conveyed information decreases, which causes each response becomes more informative (Fig. 4.7d, blue) and as a result leads to better word recognition (Fig. 4.7a, blue). By using the fact that increasing of word accuracy rate is the result of better feature selection, based on the model's results, it is obvious that changing in temporal and spectral integration characteristics in auditory pathway is crucial for extraction more informative responses regarding to speech recognition. Although the findings mirror changes in spectral and temporal selectivity seen between auditory nerve, midbrain, thalamus, and cortex receptive fields, there are some differences between the network and neural data. In general, auditory receptive fields tend to be somewhat slower and narrower than the network despite the fact that the layer-to-layer changes and system wide trends are quite similar. Such disparity may be partly attributed to neural mechanisms not include in the network such as descending feedback

(Suga 2008), synaptic and dendritic nonlinearities (Reyes 2001) and adaptive mechanisms such as spike time dependent plasticity, synaptic depression, and gain normalization (Mesgarani, David et al. 2014, Rabinowitz, Willmore et al. 2011).

The proposed acoustic processing paradigm differs from conventional auditory processing networks, which employ large-scale neural networks for speech recognition (Hinton, Deng et al. 2012, Dahl, Yu et al. 2012) and the neural network structures are predominantly employed as output classifiers, not for feature extraction. While conventional models typically extract spectral features with static filters and then feed these features sequentially into recurrent networks to account for some temporal dependencies (LeCun, Bengio et al. 2015), the proposed model directly accounts for spectro-temporal dependencies in the acoustic signal. This model transforms spectro-temporal acoustic features of the cochlea into a spatio-temporal pattern of action potentials that are sequentially refined and processed through the auditory network using multiple integration time-scales, ultimately creating a relatively noise invariant neural representation. It is surprising that despite the relative simplicity of the network and the relatively few neurons elements (318 excitatory and 318 inhibitory), proposed network is capable of achieving such high performance in the presence of noise and multiple talkers. Furthermore, it is intriguing that training the network with only three parameters can achieve such high performance and robustness (Fig. 4.6c). Our approach for reducing the parameter dimensionality (1908 parameters) by treating network parameters as smoothly varying, analogous to auditory system organization, offers a plausible alternative to conventional learning rules that require high-dimensional parameter spaces and attempt to optimize the parameters of individual neuron elements (LeCun, Bengio et al. 2015). The high performance achieved through the hierarchical approach highlights how establishing a

biologically realistic organization in which time and frequency integration characteristics scale across network layers leads to a near optimal solution.

A challenge for future studies is to reveal the combined role of biologically realistic strategies for auditory signal processing, feature extraction, and classification that together can achieve superior performance in variable acoustic environments. For one, this feature extraction network could benefit from a biologically realistic neural classifier that mirrors decision-making stages in high-level cortices (Fritz, Elhilali et al. 2007). Future approaches also could benefit by extending the biological realism in the feature extraction models, such as incorporating descending feedback projections (Suga 2008), more realistic neuron models (Gerstner, Naud 2009), and adaptive mechanisms such as spike time dependent plasticity, synaptic depression, and gain normalization (Mesgarani, David et al. 2014, Rabinowitz, Willmore et al. 2011). Ultimately, a comprehensive theory of audition requires such large-scale integration in which multiple neural mechanisms and realistic architectures produce highly robust representations that mimic normal auditory system physiology and potentially match or surpass human performance in real-world sound recognition tasks.

4.5 Funding Sources

Research reported in this chapter was partly supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award Number R01DC015138. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Additional funding was also provided by a grant through the University of Connecticut Research Foundation.

References:

ATTIAS, H. and SCHREINERT, C., 1997a. Temporal Low-Order Statistics of Natural, *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference* 1997a, MIT Press, pp. 27.

ATTIAS, H. and SCHREINERT, C., 1997b. Temporal Low-Order Statistics of Natural, *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference* 1997b, MIT Press, pp. 27.

ATTIAS, H. and SCHREINER, C.E., 1998. Coding of naturalistic stimuli by auditory midbrain neurons. *Advances in neural information processing systems,* , pp. 103-109.

BARLOW, H., 1961. Published in MIT Press.

BRADBURY, J. and BUDNEY, G., *The Diversity of Animal Sounds (Macaulay Library of Natural Sounds, Cornell Laboratory of Ornithology, Ithaca, NY).*

BRANAGH, K. and DEARMAN, G., *Shakespeare W.BBC Radio Presents: Hamlet.*

BREGMAN, A.S., 1994. *Auditory scene analysis: The perceptual organization of sound.* MIT press.

CHEN, C., READ, H.L. and ESCABI, M.A., 2012. Precise feature based time scales and frequency decorrelation lead to a sparse auditory code. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **32**(25), pp. 8454-8468.

CHI, T., GAO, Y., GUYTON, M.C., RU, P. and SHAMMA, S., 1999. Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America,* **106**(5), pp. 2719-2732.

CLOPTON, B.M. and BACKOFF, P.M., 1991. Spectrotemporal receptive fields of neurons in cochlear nucleus of guinea pig. *Hearing research,* **52**(2), pp. 329-344.

COHEN, L., 1995a. *Time-frequency analysis.* Prentice hall.

COHEN, L., 1995b. *Time-frequency analysis.* Prentice hall.

CRANDELL, C.C. and SMALDINO, J.J., 2000. Classroom acoustics for children with normal hearing and with hearing impairment. *Language, speech, and hearing services in schools,* **31**(4), pp. 362-370.

DAHL, G.E., YU, D., DENG, L. and ACERO, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing,* **20**(1), pp. 30-42.

DELGUTTE, B. and KIANG, N.Y., 1984. Speech coding in the auditory nerve: I. Vowel-like sounds. *The Journal of the Acoustical Society of America,* **75**(3), pp. 866-878.

DEPIREUX, D.A., SIMON, J.Z., KLEIN, D.J. and SHAMMA, S.A., 2001. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of neurophysiology,* **85**(3), pp. 1220-1234.

DRULLMAN, R., FESTEN, J.M. and PLOMP, R., 1994. Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America,* **95**(2), pp. 1053-1064.

EDELMAN, G.M., 1987. *Neural Darwinism: The theory of neuronal group selection.* Basic Books.

ELLIOTT, T.M. and THEUNISSEN, F.E., 2009. The modulation transfer function for speech intelligibility. *PLoS comput biol,* **5**(3), pp. e1000302.

ENGINEER, C.T., PEREZ, C.A., CHEN, Y.H., CARRAWAY, R.S., REED, A.C., SHETAKE, J.A., JAKKAMSETTI, V., CHANG, K.Q. and KILGARD, M.P., 2008a. Cortical activity patterns predict speech discrimination ability. *Nature neuroscience,* **11**(5), pp. 603-608.

ENGINEER, C.T., PEREZ, C.A., CHEN, Y.H., CARRAWAY, R.S., REED, A.C., SHETAKE, J.A., JAKKAMSETTI, V., CHANG, K.Q. and KILGARD, M.P., 2008b. Cortical activity patterns predict speech discrimination ability. *Nature neuroscience,* **11**(5), pp. 603-608.

ESCABI, M.A., MILLER, L.M., READ, H.L. and SCHREINER, C.E., 2003. Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **23**(37), pp. 11489-11504.

ESCABI, M.A., NASSIRI, R., MILLER, L.M., SCHREINER, C.E. and READ, H.L., 2005. The contribution of spike threshold to acoustic feature selectivity, spike information content, and information throughput. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **25**(41), pp. 9524-9534.

ESCABI, M.A. and SCHREINER, C.E., 2002a. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **22**(10), pp. 4114-4131.

ESCABI, M.A. and SCHREINER, C.E., 2002b. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **22**(10), pp. 4114-4131.

FIELD, D.J., 1987. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A,* **4**(12), pp. 2379-2394.

FRITZ, J.B., ELHILALI, M., DAVID, S.V. and SHAMMA, S.A., 2007. Auditory attention—focusing the searchlight on sound. *Current opinion in neurobiology,* **17**(4), pp. 437-455.

GEFFEN, M.N., GERVAIN, J., WERKER, J.F. and MAGNASCO, M.O., 2011. Auditory perception of self-similarity in water sounds. *Frontiers in integrative neuroscience,* **5**, pp. 15.

GERSTNER, W. and NAUD, R., 2009. Neuroscience. How good are neuron models? *Science (New York, N.Y.),* **326**(5951), pp. 379-380.

GREEN, J.A., GUSTAFSON, G.E. and MCGHIE, A.C., 1998. Changes in infants' cries as a function of time in a cry bout. *Child development,* **69**(2), pp. 271-279.

GREENBERG, S., 1999. Speaking in shorthand–A syllable-centric perspective for understanding pronunciation variation. *Speech Communication,* **29**(2), pp. 159-176.

HEIL, P., 1997. Auditory cortical onset responses revisited. I. First-spike timing. *Journal of neurophysiology,* **77**(5), pp. 2616-2641.

HICE, C.L., 2000. Sounds of Neotropical Rainforest Mammals: An Audio Field Guide. *Journal of mammalogy,* **81**(1), pp. 284-286.

HINTON, G., DENG, L., YU, D., DAHL, G.E., MOHAMED, A., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P. and SAINATH, T.N., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine,* **29**(6), pp. 82-97.

HOUTGAST, T. and STEENEKEN, H.J., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America,* **77**(3), pp. 1069-1077.

HROMÁDKA, T., DEWEESE, M.R. and ZADOR, A.M., 2008. Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol,* **6**(1), pp. e16.

HUBEL, D.H. and WIESEL, T.N., 1959. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology,* **148**(3), pp. 574-591.

JORIS, P.X., SCHREINER, C.E. and REES, A., 2004. Neural processing of amplitude-modulated sounds. *Physiological Reviews,* **84**(2), pp. 541-577.

KENDALL, M.G., 1979. The advanced theory of statistics. *The advanced theory of statistics.,* (4th Ed),.

KIM, P.J. and YOUNG, E.D., 1994. Comparative analysis of spectro-temporal receptive fields, reverse correlation functions, and frequency tuning curves of auditory-nerve fibers. *The Journal of the Acoustical Society of America,* **95**(1), pp. 410-422.

LECUN, Y., BENGIO, Y. and HINTON, G., 2015. Deep learning. *Nature,* **521**(7553), pp. 436-444.

LEE, C.M., OSMAN, A.F., VOLGUSHEV, M., ESCABI, M.A. and READ, H.L., 2016. Neural spike-timing patterns vary with sound shape and periodicity in three auditory cortical fields. *Journal of neurophysiology,* **115**(4), pp. 1886-1904.

LESICA, N.A. and GROTHE, B., 2008. Efficient temporal processing of naturalistic sounds. *PLoS One,* **3**(2), pp. e1655.

LEVY, R.B. and REYES, A.D., 2012. Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **32**(16), pp. 5609-5619.

LEWICKI, M.S., 2002. Efficient coding of natural sounds. *Nature neuroscience,* **5**(4), pp. 356-363.

LIBERMAN, M.E.A., *NIST Speech Disc 7-1.1 (1disc).*

LIU, R.C., MILLER, K.D., MERZENICH, M.M. and SCHREINER, C.E., 2003. Acoustic variability and distinguishability among mouse ultrasound vocalizations. *The Journal of the Acoustical Society of America,* **114**(6), pp. 3412-3422.

LOFTUS, W.C., BISHOP, D.C., SAINT MARIE, R.L. and OLIVER, D.L., 2004. Organization of binaural excitatory and inhibitory inputs to the inferior colliculus from the superior olive. *Journal of Comparative Neurology,* **472**(3), pp. 330-344.

MALONE, B. and SCHREINER, C.E., 2010. Time-varying sounds: amplitude envelope modulations. *The auditory brain.Oxford University Press, Oxford, New York, ,* pp. 125-148.

MC LAUGHLIN, M., VAN DE SANDE, B., VAN DER HEIJDEN, M. and JORIS, P.X., 2007. Comparison of bandwidths in the inferior colliculus and the auditory nerve. I. Measurement using a spectrally manipulated stimulus. *Journal of neurophysiology,* **98**(5), pp. 2566-2579.

MCCALLUM, A. and NIGAM, K., 1998. A comparison of event models for naive bayes text classification, *AAAI-98 workshop on learning for text categorization* 1998, Madison, WI, pp. 41-48.

MCDERMOTT, J.H., SCHEMITSCH, M. and SIMONCELLI, E.P., 2013. Summary statistics in auditory perception. *Nature neuroscience,* **16**(4), pp. 493-498.

MCDERMOTT, J.H. and SIMONCELLI, E.P., 2011a. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron,* **71**(5), pp. 926-940.

MCDERMOTT, J.H. and SIMONCELLI, E.P., 2011b. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron,* **71**(5), pp. 926-940.

MERZENICH, M.M. and REID, M.D., 1974. Representation of the cochlea within the inferior colliculus of the cat. *Brain research,* **77**(3), pp. 397-415.

MESGARANI, N., DAVID, S.V., FRITZ, J.B. and SHAMMA, S.A., 2014. Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America,* **111**(18), pp. 6792-6797.

MILLER, L.M., ESCABI, M.A., READ, H.L. and SCHREINER, C.E., 2002. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of neurophysiology,* **87**(1), pp. 516-527.

OLIVER, D.L., 2000. Ascending efferent projections of the superior olivary complex. *Microscopy research and technique,* **51**(4), pp. 355-363.

OSWALD, A.M., SCHIFF, M.L. and REYES, A.D., 2006. Synaptic mechanisms underlying auditory processing. *Current opinion in neurobiology,* **16**(4), pp. 371-376.

QIU, A., SCHREINER, C.E. and ESCABI, M.A., 2003. Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. *Journal of neurophysiology,* **90**(1), pp. 456-476.

RABINOWITZ, N.C., WILLMORE, B.D., SCHNUPP, J.W. and KING, A.J., 2011. Contrast gain control in auditory cortex. *Neuron,* **70**(6), pp. 1178-1191.

READ, H.L., MILLER, L.M., SCHREINER, C.E. and WINER, J.A., 2008. Two thalamic pathways to primary auditory cortex. *Neuroscience,* **152**(1), pp. 151-159.

REYES, A., 2001. Influence of dendritic conductances on the input-output properties of neurons. *Annual Review of Neuroscience,* **24**(1), pp. 653-675.

RODRIGUEZ, F.A., CHEN, C., READ, H.L. and ESCABI, M.A., 2010. Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **30**(47), pp. 15969-15980.

RODRIGUEZ, F.A., READ, H.L. and ESCABI, M.A., 2010. Spectral and temporal modulation tradeoff in the inferior colliculus. *Journal of neurophysiology,* **103**(2), pp. 887-903.

RUDERMAN, D.L., 1997. Origins of scaling in natural images. *Vision research,* **37**(23), pp. 3385-3398.

RUDERMAN, D.L. and BIALEK, W., 1994. Statistics of natural images: Scaling in the woods. *Physical Review Letters,* **73**(6), pp. 814.

SACHS, M.B., VOIGT, H.F. and YOUNG, E.D., 1983. Auditory nerve representation of vowels in background noise. *Journal of neurophysiology,* **50**(1), pp. 27-45.

SCHNEIDER, D.M. and WOOLLEY, S.M., 2013. Sparse and background-invariant coding of vocalizations in auditory scenes. *Neuron,* **79**(1), pp. 141-152.

SCHWARTZ, O. and SIMONCELLI, E.P., 2001. Natural signal statistics and sensory gain control. *Nature neuroscience,* **4**(8), pp. 819-825.

SEN, K., THEUNISSEN, F.E. and DOUPE, A.J., 2001. Feature analysis of natural sounds in the songbird auditory forebrain. *Journal of neurophysiology,* **86**(3), pp. 1445-1458.

SHANNON, C.E., 1958. Channels with side information at the transmitter. *IBM journal of Research and Development,* **2**(4), pp. 289-293.

SHANNON, R.V., ZENG, F., KAMATH, V., WYGONSKI, J. and EKELID, M., 1995. Speech recognition with primarily temporal cues. *Science,* **270**(5234), pp. 303.

SIMONCELLI, E.P., 2003. Vision and the statistics of the visual environment. *Current opinion in neurobiology,* **13**(2), pp. 144-149.

SIMONCELLI, E.P., PANINSKI, L., PILLOW, J. and SCHWARTZ, O., 2004. Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences,* **3**, pp. 327-338.

SINGH, N.C. and THEUNISSEN, F.E., 2003. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America,* **114**(6), pp. 3394-3411.

SMITH, E.C. and LEWICKI, M.S., 2006. Efficient auditory coding. *Nature,* **439**(7079), pp. 978-982.

STRONG, S., VAN STEVENINCK, RR DE RUYTER, BIALEK, W. and KOBERLE, R., 1998. On the application of information theory to neural spike trains, *Pac Symp Biocomput* 1998, pp. 32.

SUGA, N., 2008. Role of corticofugal feedback in hearing. *Journal of Comparative Physiology A,* **194**(2), pp. 169-183.

TAN, A.Y., ZHANG, L.I., MERZENICH, M.M. and SCHREINER, C.E., 2004. Tone-evoked excitatory and inhibitory synaptic conductances of primary auditory cortex neurons. *Journal of neurophysiology,* **92**(1), pp. 630-643.

TER-MIKAELIAN, M., SEMPLE, M.N. and SANES, D.H., 2013. Effects of spectral and temporal disruption on cortical encoding of gerbil vocalizations. *Journal of neurophysiology,* **110**(5), pp. 1190-1204.

TRAPPENBERG, T., 2009. *Fundamentals of computational neuroscience.* OUP Oxford.

ULANOVSKY, N., LAS, L. and NELKEN, I., 2003. Processing of low-probability sounds by cortical neurons. *Nature neuroscience,* **6**(4), pp. 391-398.

VON TRAPP, G., BURAN, B.N., SEN, K., SEMPLE, M.N. and SANES, D.H., 2016. A Decline in Response Variability Improves Neural Signal Detection during Auditory Task Performance. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **36**(43), pp. 11097-11106.

VOSS, R.F. and CLARKE, J., 1978. ''1/f noise''in music: Music from 1/f noise. *The Journal of the Acoustical Society of America,* **63**(1), pp. 258-263.

WEHR, M. and ZADOR, A.M., 2003. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature,* **426**(6965), pp. 442-446.

WINER, J.A., SAINT MARIE, R.L., LARUE, D.T. and OLIVER, D.L., 1996. GABAergic feedforward projections from the inferior colliculus to the medial geniculate body. *Proceedings of the National Academy of Sciences of the United States of America,* **93**(15), pp. 8005-8010.

WÖHR, M., DAHLHOFF, M., WOLF, E., HOLSBOER, F., SCHWARTING, R.K. and WOTJAK, C.T., 2008. Effects of genetic background, gender, and early environmental factors on isolation-induced ultrasonic calling in mouse pups: an embryo-transfer study. *Behavior genetics,* **38**(6), pp. 579-595.

WÖHR, M. and SCHWARTING, R.K., 2008. Maternal care, isolation-induced infant ultrasonic calling, and their relations to adult anxiety-related behavior in the rat. *Behavioral neuroscience,* **122**(2), pp. 310.

WOOLLEY, S.M., FREMOUW, T.E., HSU, A. and THEUNISSEN, F.E., 2005. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature neuroscience,* **8**(10), pp. 1371-1379.

XIE, R., GITTELMAN, J.X. and POLLAK, G.D., 2007. Rethinking tuning: in vivo whole-cell recordings of the inferior colliculus in awake bats. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **27**(35), pp. 9469-9481.

ZHENG, Y. and ESCABI, M.A., 2013. Proportional spike-timing precision and firing reliability underlie efficient temporal processing of periodicity and envelope shape cues. *Journal of neurophysiology,* **110**(3), pp. 587-606.

ZHENG, Y. and ESCABI, M.A., 2008. Distinct roles for onset and sustained activity in the neuronal code for temporal periodicity and acoustic envelope shape. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* **28**(52), pp. 14230-14244.

ZYLBERBERG, J., PFAU, D. and DEWEESE, M.R., 2012. Dead leaves and the dirty ground: Low-level image statistics in transmissive and occlusive imaging environments. *Physical Review E,* **86**(6), pp. 066112.

[i]