Portland State University

PDXScholar

Dissertations and Theses

Dissertations and Theses

1-1-2010

Efficient Encoding Of Vocalizations In The Auditory Midbrain

Lars Andreas Holmstrom
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

Let us know how access to this document benefits you.

Recommended Citation

Holmstrom, Lars Andreas, "Efficient Encoding Of Vocalizations In The Auditory Midbrain" (2010).
Dissertations and Theses. Paper 27.
<https://doi.org/10.15760/etd.27>

This Dissertation is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

Efficient Encoding Of Vocalizations In The Auditory Midbrain

by

Lars Andreas Holmstrom

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Systems Science

Dissertation Committee:
James McNames, Chair
George G. Lendaris
Christine Portfors
Patrick Roberts
Wayne Wakeland
Gerardo Lafferriere

Portland State University
©2010

Abstract

An important question in sensory neuroscience is what coding strategies and mechanisms are used by the brain to detect and discriminate among behaviorally relevant stimuli. To address the noisy response properties of individual neurons, sensory systems often utilize broadly tuned neurons with overlapping receptive fields at the system's periphery, resulting in homogeneous responses among neighboring populations of neurons. It has been hypothesized that progressive response heterogeneity in ascending sensory pathways is evidence of an efficient encoding strategy that minimizes the redundancy of the peripheral neural code and maximizes information throughput for higher level processing. This hypothesis has been partly supported by the documentation of neural heterogeneity in various cortical structures.

This dissertation examines whether selective and sensitive responses to behaviorally relevant stimuli contribute to a heterogeneous and efficient encoding in the auditory midbrain. Prior to this study, no compelling experimental framework existed to address this question. Stimulus design methodologies for neuroethological experiments were largely based on token vocalizations or simple approximations of vocalization components. This dissertation describes a novel state-space signal modeling methodology which makes possible the independent manipulation of the frequency, amplitude, duration, and harmonic structure of vocalization stimuli. This methodology was used to analyze four mouse vocalizations and create a suite of perturbed variants of each of these vocalizations. Responses of neurons in the mouse inferior colliculus (IC) to the natural vocalizations and their perturbations were characterized using measures of both spike rate and spike timing. In order to compare these responses to those of peripheral auditory neurons, a data-driven model was developed and fit to each IC neuron based on the neuron's pure tone responses. These models were then used to approximate how peripheral auditory neurons would respond to our

suite of vocalization stimuli. Using information theoretic measures, this dissertation argues that selectivity and sensitivity by individual neurons results in heterogeneous population responses in the IC and contributes to the efficient encoding of behaviorally relevant vocalizations.

Acknowledgements

Without the assistance of faculty, friends, and family, I never would have taken on and achieved this substantial, life changing goal. While I sometimes wonder about the opportunity lost by spending so much time (and money!) engaged in this arguably selfish endeavor for the last number of years, I have remarkably few regrets about the process that I am now completing. Here I express my gratitude to the people that helped me the most on this journey.

Erin Colclazier, for almost single handedly keeping my spirits afloat throughout graduate school. I am truly indebted to her for her unfaltering support and determination to make me have fun when I needed it the most. These are the kinds of debts that I like to pay back.

My parents, for their support throughout all of my education. Not only did they instill in me a confidence that I could take on any educational challenge, but they supported me when I chose not to take these challenges on. I'm sure they worried about me when I dropped out of college in 1993, but their trust in my own decision making process was a better lesson for me to learn at the time.

My sister, for teaching me how to fight for the important things in life.

My entire committee, for challenging me to push my limits (which I discovered on many occasions). I have leaned on each and every one of them at different times through this process. I am very grateful to have had access to a group so diverse in their specialties, yet so consistent in their kindness, approachability, depth of knowledge, and eagerness to share this knowledge. I couldn't have done this without a single one of them.

Lonneke Eeuwes, for all of the long hours she spent collecting data for our exper-

iments and carefully reading through my manuscripts.

Sunghan Kim, for working with me through the technical details of our state space frequency tracking methods. My research was made possible by your work.

Roberto Santiago, for taking me under his wing as a fresh graduate student and imbuing me with a sense of adventure about research. He taught me that the challenges inherent in science are not a chore, but a privilege to be relished.

John Vasallo, for setting time aside for our weekly musical catharsis and for being flexible with respect to my constantly shifting schedule. He almost single handedly ensured that playing music will continue to be an important part of my life.

Todd Haynes, for his roving intellect and dedication to the creative process. I can only hope that I can tap into a similar wellspring of enthusiasm, kindness, and vision.

Kathleen McConnel and Casey Quinlan, for completing their Ph.D.'s before me. This greatly motivated me to get done before they started to gloat.

The New Bad Things, for providing me with much needed memories of reckless abandon in my youth. This was an important balance for all my time spent laboring in the salt mines during graduate school.

Peter Siracusa, for having one of the most inquisitive and probing minds I know, and for showing me the value of thinking and living outside the box.

Graham Ross, for being one of my earliest post-undergraduate mentors. Aside from teaching me much of what I know about the fundamentals of software development, his following words gave me an existentialist comfort throughout my graduate studies: "Getting a Ph.D. is an asymptotic approach towards a singularity. You know more and more about less and less until finally you know everything about nothing."

And, finally, to all of my friends that I did not single out here. Like it or not, you'll be seeing more of me now!

Contents

Abstract	i
Acknowledgements	iii
List of Tables	viii
List of Figures	ix
Definitions of Terms	x
Preface	xi
1 Introduction	1
2 Background and Significance	5
2.1 Neural Processing of Sound	5
2.1.1 Psychoacoustics	5
2.1.2 The Ascending Auditory Pathways	8
2.1.3 The Cochlea	8
2.1.4 The Auditory Nerve - Neural Processing at the Periphery . . .	10
2.1.5 Peripheral Auditory Nonlinearity	15
2.1.6 The Inferior Colliculus	17
2.1.7 Efficient Encoding Through Heterogeneity	19
2.2 Stimulus Design for Auditory Neuroethology	21
2.3 Frequency Analysis and Tracking	24
2.4 Spectro-Temporal Modeling of Neural Responses	26
2.5 Pure Tone STRF Models	31
2.6 Medical Significance	34
3 Methods	36
3.1 Animals	36
3.2 Surgical procedures	37
3.3 Stimulus Design	38
3.4 Acoustic stimulation	41
3.5 Data acquisition and analysis	43
3.6 Spike Metrics	44
3.7 Comparing responses in the IC to responses in lower auditory nuclei .	48
3.8 Discrimination efficiency among vocalization variants	54
3.9 Quantification of response heterogeneity across IC neurons	55

4	Results	57
4.1	Predicting responses to complex stimuli from responses to pure tones	59
4.2	Changing the spectral content of vocalizations altered neural responses	62
4.3	Removal of AM or FM from the vocalizations altered neural responses	66
4.4	Vocalization duration modulated neural responses	68
4.5	Population responses	70
4.6	Efficiency of vocalization discrimination in the IC	76
4.7	Heterogeneity of responses across the neural population	82
5	Conclusions	86
5.1	State-Space Stimulus Design	86
5.2	Heterogeneous response properties in the IC	90
5.3	Efficient encoding in the IC	93
5.4	Contributions	95
5.5	Future Directions	96
	Bibliography	98
	Appendices	116
A	State-Space Stimulus Design for Auditory Neuroethology	116
A.1	Overview	116
A.2	Methods	117
A.2.1	State-Space Signal Modeling	117
A.2.2	Harmonic Frequency State-Space Signal Model	119
A.2.3	Vocalization Analysis	122
A.2.4	Modifying Vocalizations	122
A.2.5	Synthesizing Vocalizations	123
B	Vocalization Analysis with the Extended Kalman Smoother	125
B.1	The Kalman Filter	125
B.2	The Extended Kalman Filter	126
B.3	The Extended Kalman Smoother	131
B.4	Performance Assessment of EKS Analysis of Vocalizations	132
B.4.1	Synthetic Signal Generation	133
B.4.2	Measurement Noise Models	135
B.4.3	Assessment of State-Space Tracking on Real Vocalizations	136
B.5	Results	139
B.5.1	Comparison of EKS and Spectrographic Performance	139
B.5.2	Harmonic Tracking of Hummingbird Vocalization	146
B.5.3	Morphing the Hummingbird Vocalization	148
B.5.4	Harmonic Tracking of Human Vocalization	152
B.5.5	Harmonic Tracking of Bat Vocalization	158

C	Vocalization Analysis with the Particle Filter	163
C.1	Overview	163
C.2	Notation	168
C.3	Recursions for Typical Particle Filtering Methods	169
C.3.1	Standard Resampling Particle Filter (PF)	169
C.3.2	Maximum A Posteriori Particle Filter (MAP-PF)	170
C.3.3	The Marginalized Particle Filter (MPF)	173
C.4	Optimal MAP Adaptive Marginalized Particle Filter	175
C.5	Fast MAP Adaptive Marginalized Particle Filter	178
C.6	Performance Assessment of Particle Filter Analysis of Vocalizations .	180
D	Neural Data Analysis and Visualization	192

List of Tables

4.1	Modifying acoustic parameters of vocalizations altered neural responses in a heterogeneous manner	73
-----	--	----

List of Figures

2.1	Schematic of the ascending neural pathways of the auditory system .	7
2.2	Schematic of the generation of a frequency tuning curve for an auditory neuron	13
2.3	Place and temporal encoding in a single auditory nerve	14
2.4	A STRF generated by ripple-noise	30
2.5	Pure tone STRF modeling in the Mustached Bat	33
3.1	Spectrograms and oscillograms of the natural ultrasonic mouse vocalizations used in the study	40
3.2	Analysis and synthesis of the 40 kHz harmonic vocalization	42
3.3	A comparison of STRF model optimization using the linear 3.8 and nonlinear 3.11 optimality criterion	53
4.1	Distribution of characteristic frequencies	58
4.2	PSTHs of six representative neurons in response to the natural vocalizations	60
4.3	Example of pure tone modeling of a neuron from this study	63
4.4	Altering the spectral content of the vocalizations explained the change in responsiveness in some neurons	65
4.5	AM and FM are important features for creating selective neural responses	67
4.6	Vocalization duration modulates the responses of neurons in the IC .	69
4.7	Characteristics of neural responses to the 30 kHz and 40 kHz vocalizations	72
4.8	Characteristics of neural responses to the female up-sweep vocalization	75
4.9	Specific examples of responses to variants of the female and male up-sweep vocalizations	77
4.10	Efficiency of neural discrimination among vocalizations	79
4.11	Heterogeneity of IC responses to a given vocalization	83

Definitions of Terms

CF: Characteristic Frequency. The frequency that elicits a response in an auditory neuron at the lowest sound pressure level.

Conspecific: Two or more individual organisms, populations, or taxa are conspecific if they belong to the same species.

EKF: Extended Kalman Filter. An iterative and causal method of estimating the states of a system given a series of observations

EKS: Extended Kalman Smoother. An iterative and non-causal method of estimating the states of a system given a series of observations

FRA: Frequency Response Area. For auditory neurons, this refers to the range of pure tone frequencies and intensities that elicit a significant response in the neuron.

IC: Inferior colliculus. The largest and most important nuclei of the auditory mid-brain.

ICC: The central nucleus of the inferior colliculus

Neural Selectivity: A neuron is considered selective if it responds preferentially to a particular stimulus compared to similar stimuli.

Neural Sensitivity: A neuron is considered sensitive if its responses are altered significantly when a particular stimulus is perturbed, but it does not show a preference to one of these stimuli.

Neuroethology: the evolutionary and comparative approach to the study of animal behavior and its underlying mechanistic control by the nervous system.

Nuclei: A physically and functionally distinct collection of neurons

PSTH: Peristimulus time histogram. A histogram used to visualize the rate and timing of neuronal spike discharges in relation to an external stimulus or event.

Receptive Field: The receptive field of a sensory neuron is a region of parameter space in which the presence of a stimulus will alter the firing of that neuron. Receptive fields have been identified for neurons of the auditory system, the somatosensory system, and the visual system.

Tonotopy: the spatial arrangement of where sound is perceived, transmitted, or received. It refers to the fact that tones close to each other in terms of frequency are represented in topologically neighboring neurons in the brain.

Preface

Systems Science Approach

One thing that I have had to answer throughout my graduate career is “what do you study.” Inevitably, this is quickly followed by “what is systems science.” While my answer has often been dependent upon what courses I was currently enrolled in and who my target audience was, my own ideas of what it is to be a systems scientist have gelled over time. More often than not, I now state “systems science is the study of methodologies that can be used to analyze complex and dynamic systems, whether they be human systems, computer systems, or neural systems.” This inevitably either ends the encounter or sparks an interesting conversation.

Throughout the process of working my way through this program, I often felt like I had trouble focusing on a common topic for very long. Taking a step back to look at my research over the past seven years, however, I recognize the reoccurring patterns and themes that have truly defined my experience as a student of systems science. My motivation has always been driven by gaining insight into how my own mind works, and the tools that will help me gain this insight. My early graduate research was focused on the study of artificial neural networks, and specifically on models of how we learn from our experiences with the environment [[Holmstrom et al., 2005](#), [2007a](#), [Toland et al., 2007](#)]. This period not only gave me insight into how I learn, but affected the way that I approach learning myself.

My interest in music led me to pursue a dissertation topic that could provide insight into rhythm and the perception of sound. When the opportunity arose to directly study how real brains process auditory information, I jumped at the opportunity. Looking back, I am surprised that I was willing to leap head first into an

entirely new domain that I had little background in. I realize, however, that this leap was made possible by embracing a systems science approach and leveraging the methodological tool chest that I was already starting to fill.

This program has offered me a unique opportunity to make a contribution to the rich and historical field of auditory neurophysiology. The fact that I have already been able to publish papers in some of the field's best journals speaks volumes about the broad applicability of the systems science methodologies and the interdisciplinary focus of the systems science approach. The collaborative effort required for this research drew from the disparate yet complimentary fields of statistical signal processing, mathematics, biology, computer science, artificial neuroscience and real neuroscience. This research used many key systems science methodologies, including modeling, simulation, and information theoretic analysis: 1) signal modeling and simulation are heavily leveraged in the design of the experiments and constitute one of the key contributions of this research to the field of auditory neurophysiology, 2) the modeling of neural responses to auditory input is a data driven system identification problem and is required both to characterize auditory neurons in our experiments and to test hypotheses about encoding strategies in the auditory system; and 3) information theoretic measures are employed to characterize the efficiency of the neural code in the auditory midbrain.

Moving forward, I feel greatly empowered by the tool chest that I have put together through my studies in the Systems Science Program. True to my nature, I will likely move onto a new area of focus after my completion. I have little doubt, however, that I will bring a unique perspective and skill set to any project I take on.

Organization of this Document

This dissertation includes a combination of methodological and scientific contributions to the field of auditory neurophysiology. The main focus, however, is a scientific study which can be largely decoupled from the details of several enabling methodologies developed or co-developed for this research. For the sake of readability, an organizational decision was made to move in depth descriptions and performance assessments of some of these methodologies to their own appendices and out of the main document body. This has the unusual side effect of creating an large appendix, but improves the approachability of the primary subject matter while providing sufficient detail for the curious or dedicated reader.

1 Introduction

Most animals face a complex sensory environment from which they have to extract behaviorally relevant cues for survival. A key question in sensory neuroscience is how the brain detects and discriminates among these cues. Sensory systems such as hearing, sight, taste, and touch often utilize broadly tuned neurons with overlapping receptive fields at the system’s periphery (e.g., fingertips, retina, etc.), resulting in a distributed and redundant encoding of stimuli among these neurons. This encoding strategy helps overcome the stochastic responses of individual neurons, but is subject to inter-neuron correlations and overlapping representations of similar and/or simultaneous stimuli, complicating discrimination among these external events [Sompolinsky et al., 2001, Shamir and Sompolinsky, 2004, Averbeck et al., 2006]. Somehow, these broad and redundant representations are re-represented as they ascend through neural pathways en route to the cortex, resulting in a sensory system which is honed to detect and discriminate among behaviorally relevant cues.

It has been hypothesized that ascending sensory pathways partially achieve this goal through increased neural selectivity and sensitivity to compound events and complex features found within external stimuli [Linsker, 1988, Rolls and Treves, 1990, Atick, 1992, Chelaru and Dragoi, 2008]. For example, where a peripheral visual neuron may respond whenever light passes over a region of the retina, a selective downstream neuron may only respond when motion of light across the retina in a particular direction is present. A sensitive neuron, on the other hand, would have different responses to different stimuli within its receptive field, but would not prefer one over the other. Sensitive and selective neurons may contribute to a heterogeneous population response, and theoretical simulations show that populations of neurons

with heterogeneous response properties result in an efficient encoding that minimizes neural correlations and maximizes information throughput [Gawne and Richmond, 1993, Bell and Sejnowski, 1995, Chelaru and Dragoi, 2008, Shamir and Sompolinsky, 2006]. Evidence of heterogeneous encoding has been documented in various cortical structures [Hubel and Wiesel, 1962, Ringach et al., 2002, Haupt et al., 2004], and understanding where heterogeneity emerges in sensory pathways is vital for understanding the neural mechanisms that drive efficient encoding of behaviorally relevant stimuli. In this study we focused on the auditory system, and examined whether vocalization stimuli is encoded more heterogeneously and efficiently in the auditory mid-brain than at the periphery of the auditory system.

The IC is the main nucleus in the auditory midbrain, receiving ascending input from the majority of brainstem nuclei [Adams, 1979] and descending projections from auditory cortex [Saldaña et al., 1996]. While neurons at the periphery of the auditory system respond to any stimuli containing frequencies within their frequency response area (FRA), efficient encoding of vocalization stimuli may be facilitated by neurons in the IC that respond to specific spectro-temporal features found in vocalizations such as frequency modulations [Woolley and Casseday, 2005, Andoni et al., 2007], amplitude modulations [McAlpine, 2004, Woolley and Casseday, 2005], duration [Brand et al., 2000, Pérez-González et al., 2006], and frequency combinations [Portfors and Wenstrup, 1999, Portfors, 2004].

One of the best examples of neural selectivity to species-specific vocalizations come from the study of songbirds. Auditory neurons in the song system of male songbirds are highly selective to the bird’s own song compared to other songs from the same species, other types of calls, or the bird’s own song played in reverse order [Doupe, 1997, Margoliash, 1986, Theunissen and Doupe, 1998]. While responses to sets of token vocalizations can be compelling examples of neural selectivity, they tell us

little about what components of the vocalization are responsible for a given response or how sensitive neurons are to the presence of these components. Furthermore, it is unknown whether these selective response properties contribute to a heterogeneous and efficient population encoding of vocalizations. These questions can be addressed in neurophysiological experiments by using auditory stimulus protocols generated from vocalizations and perturbed variants of these vocalizations. Experiments following this design philosophy are relatively new to the field of auditory neurophysiology, however, and the best practices in the field for creating these stimuli are inefficient and inflexible.

In this study, we characterized neural responses to four mouse vocalizations in the mouse IC. We developed a novel state-space stimulus design methodology that enabled us to manipulate the time-varying frequency, amplitude, duration, and harmonic structure of each vocalization [Holmstrom et al., 2009, 2010]. Based on measures of both spike rate and timing, we measured the neural responses to the natural vocalizations and their perturbed variants. Using information theoretic measures, we found that selective, sensitive, and heterogeneous response properties of IC neurons contribute to an efficient, pre-cortical encoding of vocalizations.

The research reported here represents a number of methodological and scientific contributions to the field of auditory neurophysiology. The first of these is the development and application of state-space methodologies for the generation of vocalization based stimuli. These methodologies are detailed in Appendices A-C. The second contribution is the implementation and refinement of a spectro-temporal receptive field model of IC neurons. This modeling approach enabled us to approximate how peripheral auditory neurons would respond to our stimulus set. The third contribution is an in-depth characterization of the response properties of neurons in the IC to vocalizations and perturbed variants of these vocalizations from both a spike rate and

spike timing perspective. Finally, the fourth contribution is an information theoretic analysis of the efficiency of the encoding of vocalizations in the IC.

2 Background and Significance

Hearing research is a rich and historical field with contributions from numerous disciplines including psychology, physiology, mathematics, physics, and biology. The sections of this chapter aim to provide a historical and conceptual context for the research reported in this dissertation.

2.1 Neural Processing of Sound

2.1.1 Psychoacoustics

The foundations of hearing research lie in the field of psychoacoustics, which is the study of the subjective human perception of sound. The pioneers of this field were Ohm, Seebeck, and Helmholtz, who formulated their classic theorems in the early to mid 19th century.

Ohm postulated his acoustic *phase law*, which states that the perceived quality of a complex sound depended only upon its power spectrum, and not the relative phases of the tonal components [Ohm, 1843]. While more recent psychoacoustic studies have shown that humans exhibit monaural phase sensitivity under special conditions [Hall and Schroeder, 1972], Ohm’s phase law is still one of the cornerstones of hearing perception. Implicit in this law, however, was another idea that the ear decomposes a complex sound into its frequency components which can be perceived separately by a trained ear (this psychoacoustic study has been systematically addressed in modern times [Plomp, 1964]). This is different than the visual system, in which mixtures of colors are perceived as a single new color, and the individual components cannot be independently resolved.

In opposition to Ohm’s phase law, Seebeck is credited with recognizing the phe-

nomenon of *periodicity pitch* [Seebeck, 1844], which has led to modern theories of the *missing fundamental* and *residue pitch* [Schouten et al., 1962]. This phenomenon refers to the psychoacoustic sensation of perceiving components within a complex sound even when the sound has little or no spectral power at the perceived pitch. Seebeck’s findings were not in accord with Ohm’s theories, because the Fourier decomposition of the the complex sound could not account for the perceived qualities of the sound. Rather, he argued that the perception of pitch was determined by periodicity in the temporal waveform. Seebeck’s work is the foundation of the *temporal encoding* (sometimes referred to as frequency or phase encoding) theories of the neural processing of sound information.

Helmholtz was a true renaissance man, who was interested in both the psychoacoustic and physiological aspects of hearing perception, in addition to myriad other fields. His investigations of the cochlea led him to believe that it was the instrument performing an approximate Fourier analysis on sound stimuli, although characterized by limited resolving power as indicated by psychoacoustic results. This formed the basis of his resonance theory of hearing, where the cochlea was imagined to function something akin to a set of piano strings, and only the strings that were tuned to a frequency present in the Fourier decomposition of the sound stimulus would be set into motion [Helmholtz, 1853]. This idea also formed the basis for the *place encoding* theory of the neural processing of sound information, in which different areas of the cochlea correspond to different perceived pitches.

The apparent irreconcilability between the temporal and place encoding theories generated years of debate until more empirical evidence was available during the middle of the 20th century.

Figure 2.1: Schematic of the ascending neural pathways of the auditory system. This figure was obtained from [Kandel et al., 2000].

2.1.2 The Ascending Auditory Pathways

Figure 2.1 shows a schematic of the ascending neural pathways of the auditory system from the central cochlear nucleus to the auditory cortex. Each nucleus contains thousands of neurons and has a distinct functional role in the processing of auditory information. Pictured on the bottom left, this figure shows the input from a single cochlea (the point of mechanical transduction from sound pressure waves to neural action potentials), and displays how ipsilateral (same hemisphere) and contralateral (cross-hemispheric) projections are made. The superior olivary nuclei is the first location where binaural interactions occur, and are involved in the localization of sounds in space. While the nuclei of the medulla and pons are sometimes bypassed by ascending projections, all neural pathways are routed through the inferior colliculus of the midbrain.

2.1.3 The Cochlea

The cochlea, found in the inner ear, is a spiraled, hollow, conical chamber of bone which is divided into three fluid-filled parts. Two are canals for the transmission of pressure and the third is the organ of Corti, which is situated on the basilar membrane. The organ of Corti is the point of transduction in the inner ear at which sound pressure waves are converted into neural action potentials (spikes) for further neural processing. Along the length of the spiraling organ of Corti is one row of inner hair cells and three rows of outer hair cells which protrude from its surface. In all, between 16,000 to 20,000 hair cells are found here, each of which is pulled back and forth when the basilar membrane is driven up and down in response to auditory stimuli. This, in turn, causes the hair cell to generate an action potential which propagates through the eighth cranial (auditory) nerve to the cochlear nucleus.

Pioneering work by Békésy during the middle of the 20th century allowed him to visually inspect, using a water immersion microscope and stroboscopic illumination, the frequency dependent vibrations of the human basilar membrane and to build physical and mathematical models of the activity [von Békésy, 1949b,a]. Extracting the cochleas from deceased humans and other mammals, Békésy found that auditory stimuli generated traveling waves on the basilar membrane, as opposed to Helmholtz’ theory of a bank of narrowly tuned resonators. Having been a telephone engineer early in his career, Békésy correctly recognized the medium of these traveling waves as a nonuniform transmission line, in which high frequencies travel only a short distance and are then rapidly attenuated, while low frequencies travel farther along before being stopped. This behavior generated a frequency selectivity in the vibrations (in accordance with place theories of neural encoding) whose envelope had a characteristic shape — slowly rising from the base of the basilar membrane (the high frequency end) until it reached a frequency dependent point of maximum resonance, at which point it quickly dampened.

Contemporary to Békésy, Gold performed rigorous psychoacoustic experiments in order to determine the frequency selectivity and discernibility of the human ear [Gold and Pumphrey, 1948]. These tests found, based on the *refractory* period of the nervous system, that the cochlea must perform some form of resonant frequency analysis and that the selectivity (the Q-factor) of the resonant elements should be proportional to frequency. Gold also took issue with the view of a passive and heavily damped basilar membrane, claiming that this theory could not explain the sharp tuning (high Q-factor) that the resonant elements must have [Gold, 1948]. Instead, he put forth his *regeneration hypothesis*, which claimed that his results could be reconciled with the physical properties of the basilar membrane if there was a source of positive mechanical feedback (i.e., an active resonator).

This conflict was partially resolved when techniques were developed which allowed for measurements of basilar membrane vibrations on a smaller scale than Békésy's techniques and within live specimens [Johnstone and Boyle, 1967, Rhode, 1971]. These experiments showed that the high intensity stimulus required by Békésy's techniques greatly broadened the range of the traveling wave's envelope. Furthermore, these experiments showed that the tuning of the basilar membrane broadened substantially within minutes after metabolism ceases.

Evidence of the *active* cochlea goes back to 1936, when it was found that a large oscillatory electric field applied from outside the inner ear results in a sensation of hearing at the same frequency as the applied field. In 1978, however, D. T. Kemp measured evoked otoacoustical emissions (i.e., sound) from within the human ear that, coupled with Gold's theoretical prediction of an active resonator, provided substantial evidence that the cochlea could generate its own mechanical energy and affect the response of the basilar membrane to auditory stimulus [Kemp, 1978]. In more recent studies, it has been determined that the outer hair cells, which are predominantly innervated by efferent (projecting towards the periphery) connections from the superior olivary complex, are mechanically responsible for improving cochlear sensitivity and frequency selectivity [Fettiplace and Hackney, 2006].

2.1.4 The Auditory Nerve - Neural Processing at the Periphery

The auditory nerve is formed by the axons of two types of spiral ganglion cells. Type I neurons have myelinated cell bodies and innervate inner hair cells. These are afferent neurons which carry signals generated in the organ of Corti to the cochlear nucleus. Type II neurons are unmyelinated and innervate outer hair cells. Type I neurons form the vast majority of this population (95% in cats). All of the reported results below refer to Type I neurons.

Responses of the auditory nerve to pure tone stimuli have been extensively studied using micro-electrode recording techniques [Wever and Bray, 1930, Wever, 1933, Galambos and Davis, 1943, Katsuki et al., 1958]. The response properties are well characterized and have predictable responses which include:

- *Spontaneous Activity.* Even in the absence of auditory stimuli, auditory nerve fibers fire spontaneously. It is thought that such neurons are perhaps “poised” at the brink of discharging in order to respond to auditory stimulus at very low intensities.
- *Latency.* Nerve response latencies to auditory stimuli range from less than 1 ms up to 2ms.
- *Habituation.* Similar to other sensory afferents (e.g., visual and pressure), auditory nerves respond to a continued stimulus of constant frequency and intensity by a burst of impulses correlated with the onset of the stimulus which gradually declines in rate to a steady state above the spontaneous activity throughout the duration of the stimulus.
- *Rate Encoding of Intensity.* As the intensity of auditory stimulus increases, so does the firing rate of the nerve, provided the stimulus is within the frequency tuning curve of the nerve. Thus, the frequency of the stimulus is encoded in the location or frequency band of the auditory nerve, while the intensity is encoded in the firing rate of the nerve. The relationship of the rate to the stimulus intensity (in decibels) is typically sigmoidal, indicating a saturation of the firing rate at high intensities.
- *Place Encoding.* For low auditory stimuli intensities, discharge rate increases above spontaneous only for a narrow range of frequencies. As the intensity

of the stimuli increases, so does the range of frequencies to which the fiber responds. By using stimuli of different frequencies and intensities, one can generate the frequency tuning curve for each nerve fiber. The frequency for which the response threshold is minimum is called the characteristic frequency (CF) of the nerve. When placed on a logarithmic intensity scale, these tuning curves tend to be V-shaped, with long tails as the frequency decreases below CF, and short tails as the frequency increases above. The shape of this tuning curve is approximately the shape of the traveling wave envelope on basilar membrane CF [Kiang and Moxon, 1974]. After recording the frequency tuning curves of 52 auditory neurons in 16 cochleas, Liberman generated a cochlear frequency map, showing that the length of the organ of Corti allotted to an octave of stimulus frequency increases steadily from low to high frequencies [Liberman, 1982]. This frequency specific response of auditory nerves tied to the physiology of the cochlea is consistent with the *place encoding* hypothesis of Helmholtz.

- *Temporal Encoding.* Auditory nerves generate action potentials in phase with the waveform of auditory stimulus. The phase of firing is also stereotyped, such that, for example, a particular auditory nerve may only fire during the compressive phase of the stimulus [Rose et al., 1967]. As the frequency of the stimulus increases above roughly 4kHz, the refractory period of the nerve makes it impossible to fire during each cycle. The temporal encoding persists, however, and the nerves will fire during the same stereotyped phase of the stimulus, skipping cycles when necessary [Johnson, 1980]. This phasic response of auditory nerves is consistent with the *temporal encoding* hypothesis of Seebeck.

With these studies, the century old debate between the *place encoding* and *temporal encoding* hypotheses was resolved by showing that the auditory nerve uses both

Figure 2.2: Schematic of the generation of a frequency tuning curve for an auditory neuron. The left pane shows the pure tone stimulus protocol, with increasing frequency from left to right and increasing intensity from bottom to top. The frequency tuning curve has the same axis, but displays the total number of spikes fired, on average, for each presentation of the stimulus. This figure was obtained from [Escabi and Read, 2003].

encoding strategies. Even though the basilar membrane is broadly tuned, the phase of tonal stimulus is encoded by auditory nerves with different CFs but overlapping tuning curves. Even though individual auditory nerves will not be able to uniquely encode the phase of the stimulus (even neurons tuned to low frequencies often don't fire during each cycle of the waveform), Weaver's *volley hypothesis* states that a population of nerves will still uniquely encode the phase and account for the discrepancy between broad auditory nerve tuning curves and narrow frequency discrimination in psychoacoustic studies [Wever, 1933]. Figure 2.2 displays the place-rate encoding of auditory neurons evident in the frequency tuning curve. The temporal-rate encoding in the responses of a single auditory nerve to different frequencies can be seen in Figure 2.3.

Figure 2.3: Place and temporal encoding in a single auditory nerve, as indicated by interspike interval histograms. The frequency varies from 877 Hz (top left), to 318 Hz (bottom right). The CF for this nerve is close to 600 Hz, resulting in the greatest firing rate in the middle left figure. The multiple peaks in the histogram indicate that the nerves fire in phase with the stimulus waveform, but often not during each cycle. This figure was obtained from [Rose et al., 1967].

2.1.5 Peripheral Auditory Nonlinearity

The concept of peripheral auditory nonlinearity has its history in two-tone psychoacoustic studies in which certain difference tones are perceivable for certain simultaneous two-tone presentations at frequencies f_1 and f_2 . Since they were first described by the Italian musician G. Tartini in 1714 (and used in a violin tuning technique), difference tones have been central to the question of pitch perception [Goldstein and Kiang, 1968].

The most influential view of the phenomena was introduced by Helmholtz' in the middle of the 19th century [Helmholtz, 1853]. He hypothesized that aural combination tones arise from mechanical nonlinearity within the middle ear. In principle, this nonlinearity can be expressed as a power series expansion where the response is linearly proportional to the sound pressure, plus a term proportional to the square of the sound pressure, cube of the pressure, etc. If two frequencies are present, the square term would produce intermodulation products equal to the sum and difference of the two frequencies ($f_1 + f_2$, $f_1 - f_2$, $f_2 - f_1$), the quadratic term would produce products equal to twice one frequency plus and minus the other frequency ($2f_2 - f_1$, $2f_1 - f_2$, etc.), and so on. This formulation suggested that nonlinear distortion increased with stimulus intensity and would be limited to loud auditory stimuli. Otherwise, however, the combination tones would be indiscernible from any other auditory stimuli.

More recent psychoacoustic studies have systematically explored the perceptibility of difference tones. In two-tone stimulus tests, Plomp found that all four of his observers could hear pitches corresponding to the frequencies $f_2 - f_1$, $2f_1 - f_2$, and $3f_1 - 2f_2$ [Plomp, 1965]. Due to the dependence of the perceivable threshold of these difference tones on the ratio f_2/f_1 , he suggested that the source of nonlinearity was in the mechanics of the cochlea, rather than the middle ear as proposed by Helmholtz.

Through careful experiments by Goldstein, perceivable difference tones were found to be cancelable by adding a third tone to the external stimulus with frequency equal to the difference tone but with opposite phase [Goldstein, 1967]. This experiment provided a mechanism to measure the amplitude of the distortion products, but it also indicated that the distortion tones behaved as if they had been externally generated and were resonating on the basilar membrane. This hypothesis has been confirmed by modern techniques for measuring the mechanical action of the basilar membrane in response to combination tones [Kim et al., 1980, Nuttall and Dolan, 1993].

Goldstein & Kiang extended this psychoacoustic experiment into the neurophysiological domain by recording from the auditory nerves of cats while applying their two tone stimuli and looking for responses to the $2f_2 - f_1$ distortion product [Goldstein and Kiang, 1968]. These authors found that a single auditory nerve fiber may respond to the presence of the cubic difference tone even when the two primary tones (f_1 and f_2) fail to generate a response in the nerve. As in the psychoacoustic experiment, this response could also be canceled by applying an external signal at the difference tone frequency with an appropriate amplitude and phase.

Recent studies have shown that the $2f_2 - f_1$ and, to a lesser degree, $f_2 - f_1$ distortion products contribute significantly to both the mechanical motion of the cochlea and auditory nerve firing patterns even at physiologically relevant stimulus intensities (40-95 dB SPL) [Cooper and Rhode, 1997].

In addition to facilitatory interactions between paired tones, there are also inhibitory effects. Using a two-tone protocol where a constant tone is played back at the CF of the neuron and a second probe tone is played over a range of frequencies including the CF, it has been shown that the constant tone response is inhibited when the probe tone has a similar frequency [Nomoto et al., 1964, Sachs and Kiang, 1968]. These are referred to as inhibitory side-bands and can be found in the two-tone re-

sponses of many auditory neurons throughout the various brain stem, midbrain, and cortex.

Cochlear models predicting many of these nonlinear effects have been realized both in physical media [von Békésy, 1949a] and mathematically [Lukashkin and Russell, 1998, Chadwick, 1998]. Regardless of the well documented existence of nonlinearities at the periphery of our auditory system, the responses of these nerves to arbitrary stimulus at low to moderate intensities can be accurately modeled by linear reverse correlation methodologies using broad-band noise or pure tones as stimulus [de Boer and de Jongh, 1978]. This is an indication of significant linearity in the encoding of auditory stimuli at this stage of processing at behaviorally relevant intensities.

2.1.6 The Inferior Colliculus

Physiology. While most sensory systems have only two relay stations between the periphery and the cerebral cortex, the auditory system has a minimum of three with several stages of convergence and divergence and at least seven levels of commissural (from one hemisphere to the other) crossings [Malmierca, 2002]. From the point of sound pressure wave transduction in the cochlea, the auditory nerve projects to the cochlear nucleus, which itself projects to the superior olivary complex and the lateral lemniscus. All of the projections from these nuclei eventually converge on the auditory midbrain, the inferior colliculus [Malmierca, 2004/2005].

The inferior colliculus (IC) is among the largest auditory nuclei in the vertebrate brain and plays the role of relay station for all ascending auditory information. Ascending projections into the IC originate in nearly all parts of the cochlear nucleus, from much of the olivary complex, and from each of the lateral lemniscus nuclei Adams [1979]. Descending projections are received from every auditory cortical area Saldaña et al. [1996]. In addition, the IC has extensive projections within its own nucleus (in-

trinsic), to the IC in the other hemisphere (commissural), and to nearly all brain stem nuclei. The IC contains three principal divisions which differ in neuronal structure, connection patterns, and likely functional roles. The central nucleus is exclusively auditory and is essential for normal hearing. The lateral nucleus receives non-auditory input and is multisensory. The roles of the dorsal cortex in hearing is unknown. The central nucleus of the inferior colliculus (ICC) is the most commonly studied of these divisions. The ICC is tonotopically organized, with frequency band laminae consisting of thousands of neurons with similar characteristic frequencies. These laminae are not uniform physiologically and may themselves contain functionally distinct regions. The axons of most ICC cell types have extensive local branches, resembling interneurons elsewhere in the brain [Winer and Schreiner, 2005, Schreiner and Langner, 1997].

Pure Tone Responses. Early microelectrode studies of the central nucleus of the inferior colliculus focused on responses to clicks and pure tones [Thurlow et al., 1951]. Frequency response tuning curves tend to be sharper than those found in auditory nerves [Katsuki et al., 1958] and the shapes of the curves are not stereotypic. Early studies found that the spike rates of many cells did not monotonically increase as the intensity of pure tone stimulus increased, indicating the effects of side-band inhibition [Rose et al., 1963]. Recent studies have identified 4 classes of tuning curves found in the ICC of awake mice using single and two-tone stimulus protocols to facilitate the measurement of inhibitory response areas. Class I neurons are characterized by a steep slope on the high-frequency side of the excitatory response area and broadening of the frequency response area towards the low-frequency side, which may or may not include a low-frequency tail (these tuning curves are similar to those found in the auditory nerve). Class II neurons have an excitatory response area that was restricted to a narrow bandpass with steep slopes, with areas of strong inhibition flanking both sides. Class III neurons have a symmetrical widening of the excita-

tory frequency response area (i.e. V-shaped) with increasing SPL. Class IV neurons have complex forms of excitatory and inhibitory receptive fields with more than one characteristic frequency (i.e., multiply tuned) [Egorova et al., 2001]. In addition, six classes of firing patterns have been identified [Malmierca, 2004/2005].

Responses To Complex Sound Stimuli. Neurons in the ICC have been shown to respond preferentially to complex spectro-temporal features in auditory stimuli. These features include, but are not limited to, frequency and amplitude modulations [Woolley and Casseday, 2005], stimuli duration [Pérez-González et al., 2006], specific delays between spectral components [Portfors and Wenstrup, 1999], facilitatory and inhibitory frequency interactions [Portfors and Wenstrup, 2002, Nataraj and Wenstrup, 2005], and even novelty [Pérez-González et al., 2005]. While there is evidence that the effects of cochlear distortion may influence responses in the ICC [McAlpine, 2004], many of the responses to complex sound stimuli point to the existence of nonlinear spectro-temporal interactions. These features are fundamental building blocks of mammalian vocalizations, indicating that a significant amount of spectro-temporal feature extraction required for speech perception occurs pre-cortically. The IC promises to be both a great challenge and a great resource in the quest for understanding the foundations of speech perception.

2.1.7 Efficient Encoding Through Heterogeneity

Peripheral sensory neurons are often coarsely tuned and stochastic, yet subtle discrimination among stimuli is critical for many species’ survival. One approach for overcoming the noisy responses of individual neurons is through population encoding in which sensory information is redundantly distributed across a large number of neurons. The benefits of population encoding at the periphery have been hypothesized in the visual [Lee et al., 1988], motor [Georgopoulos et al., 1986], and auditory sys-

tems [Wever, 1933]. However, correlations in neural responses limit the effectiveness of this approach [Sompolinsky et al., 2001, Shamir and Sompolinsky, 2004, Averbek et al., 2006] and distributed encoding schemes make discrimination among similar and/or simultaneously encountered stimuli difficult due to overlapping neural representations.

It has been hypothesized that sensory systems facilitate the discrimination task by implementing a factorial code in which information with considerable redundancy and correlation (as is true for most naturalistic stimuli) is recoded such that distinct stimulus events are represented with little or no statistical interdependence [Linsker, 1988, Nadal and Parga, 1994]. This process may be the result of an evolutionary optimization that maximizes the efficiency of the neural code, while subject to the physical constraints of neural networks and the statistical and behavioral qualities of the stimuli [Barlow, 1960, Linsker, 1988, Bialek et al., 1991, Atick, 1992]. This results in the migration from a redundant encoding at the periphery to an efficient, heterogeneous encoding in higher level processing that is driven by selectivity and sensitivity to increasingly complex features of external stimuli. This encoding strategy also dictates that higher level neural representations match the statistical and behavioral qualities of the stimuli [Barlow et al., 1989]. Stimuli that are not often encountered may be encoded less efficiently than stimuli that are repeatedly encountered. Stimuli that are never encountered may not be encoded at all.

Evidence of heterogeneous encoding has been documented in various cortical structures [Hubel and Wiesel, 1962, Ringach et al., 2002, Haupt et al., 2004] and in sensory modalities as diverse as taste [Rolls and Treves, 1990] and vision [Rolls and Tovee, 1995]. This dissertation investigates whether vocalization stimuli are encoded more heterogeneously and efficiently in the auditory mid-brain than at the periphery of the auditory system.

2.2 Stimulus Design for Auditory Neuroethology

In what is essentially a system identification problem, one of the goals of auditory neurophysiology is to fully characterize the responses of individual neurons to arbitrary sound stimuli. The hope of this reductionist approach is to gain insight into the functional role played by the complex networks of neurons and nuclei that constitute our auditory system. Although much effort has gone into the task of generating theories, algorithms, and metrics for quantifying neural responses, stimulus design is an equal partner in the process driving hypothesis testing, neural categorization, and model design.

The tonotopic organization of the auditory nerve and other auditory nuclei — particularly those near the periphery — has made pure tone stimuli of differing frequencies and intensities a common protocol for categorizing neurons based on their frequency response areas. Nuclei farther from the periphery, however, may have complex response characteristics that are not well characterized by their pure tone responses alone. To address this, the ubiquitous pure tone stimulus protocol has been expanded to include combinations of tones, frequency sweeps, amplitude and frequency modulations, and other synthetic sounds.

A downside to this approach is the exponential explosion of stimulus variants. Even if only two simultaneous tones are considered, a stimulus protocol that pairs a large number of pure tones of differing frequencies and amplitudes may take a prohibitively long period of time. The neuroethological approach to this “curse of dimensionality” is to constrain the range of stimulus variants to approximate or replicate sounds that are behaviorally relevant to the subject.

In particular, species-specific vocalizations have been explored for guiding stimulus design. For example, based on the harmonic structure of many species’ vocalizations,

harmonic combinations of pure tones have been used [Sinex et al., 2002, 2003, Ehret, 2002]. This greatly reduces the complexity of a multi-tonal stimulus protocol. Bats are commonly studied due to their highly specialized auditory systems, and stimulus protocols have been designed using combinations of tones, frequency sweeps, and delays inspired by their echolocation vocalizations. These studies have shown that some neurons in the inferior colliculus and auditory cortex are highly selective to these combinations of synthetic stimuli [Pollak et al., 1977, Suga and Jen, 1976, Suga et al., 1979, Portfors and Wenstrup, 1999, 2001, 2002].

Another approach to the curse of dimensionality in stimulus design has been addressed by the spectro-temporal receptive field (STRF) modeling methodologies [Eggermont, 1993, Eggermont et al., 1983, Young and Calhoun, 2005, Sen et al., 2001]. Instead of using structured stimuli built from pure tones, frequency sweeps, and other spectro-temporal structures, highly unstructured white noise is used in conjunction with reverse correlation and spike-triggered averaging techniques for fitting these models [de Boer and Kuyper, 1968, de Boer and de Jongh, 1978, Aertsen and Johannesma, 1981]. Similar to the problem posed by pure tone stimuli, however, white noise is of limited use because many auditory neurons display highly nonlinear response characteristics and do not respond robustly to broad-band stimuli [Theunissen, 2004]. To address this issue, the algorithms used to fit STRF models have been adapted to accommodate pure tones, narrow-band gaussian noise, ripple noise, and, more recently, natural stimuli such as recorded vocalizations [Klein et al., 2000, Escabi and Read, 2003, Holmstrom et al., 2007b, Theunissen et al., 2000].

Both of these stimulus design philosophies, one starting from the highly structured stimuli of pure tones and the other from the highly unstructured stimuli of white noise, have largely come to rely on species-specific vocalizations to guide their development.

The best examples of neural selectivity to species-specific vocalizations come from

the study of songbirds. Auditory neurons in the song system of male songbirds are highly selective to the bird’s own song compared to other conspecific songs, other types of calls, or the bird’s own song played in reverse order [Doupe, 1997, Margoliash, 1986, Theunissen and Doupe, 1998]. While responses to sets of token vocalizations can be compelling examples of neural selectivity, they tell us little about what components of the vocalization are responsible for a given response or how sensitive neurons are to perturbations in these vocalizations. Hypothesis testing is much more approachable when a single recorded vocalization can be parametrically modeled and modified spectro-temporally to produce variants that are quantifiably different.

One approach guided by this principle focuses primarily on manipulating vocalizations directly through techniques such as amplification or attenuation, shifting in frequency, and cropping [Suga, 1992, Wang et al., 1995, Portnoff, 1981]. A more flexible approach involves generation of synthetic stimuli based on a parameterized model of the vocalization. For example, spectro-temporal filter bank representations of vocalizations have been explored in bird song to decompose a vocalization into a sum of fixed frequency sinusoids with time dependent amplitudes and relative phase offsets. This decomposition enables the manipulation of the amplitude and relative phase offsets of each frequency channel independently during synthesis [Theunissen and Doupe, 1998].

Similarly, spectrographic decompositions based on the short-time Fourier transform (STFT) are often used for independently manipulating the frequency and time scale of speech and audio signals using a technique called phase vocoding [Flanagan and Golden, 1966, Portnoff, 1976]. Phase vocoding and related techniques rely on manipulations of the STFT magnitude. For many of these modifications, however, a valid inverse of the STFT does not exist [Hayes et al., 1980] — resulting in the proliferation of heuristic methods for maintaining phase coherence in the synthesis

phase [Crochiere, 1980].

In addition to issues associated with synthesizing signals from filter bank or STFT representations, these approaches can be unwieldy due to the large number of parameters required to model the vocalization. To address this issue, more compact signal models have been explored. Synthesized bird songs have been modeled as a single sinusoid varying continuously in amplitude and frequency [Margoliash, 1983]. Based on the common harmonic structure of many species' vocalizations, signal models consisting of harmonically related sinusoids of varying amplitude and frequency have also been explored for synthesizing vocalizations from species as diverse as the mustached bat [Ohlemiller et al., 1994] and the common marmoset [DiMattina and Wang, 2006]. In order to fit these signal models, however, the time-varying frequencies, amplitudes, and relative phases of the harmonic components must first be estimated from the original recording, which is itself a non-trivial task.

2.3 Frequency Analysis and Tracking

Fourier analysis techniques, developed during the early part of the 19th century and named after Joseph Fourier, revolutionized many fields of mathematics and physics by showing that an arbitrary function could be decomposed into an infinite sum of sinusoidal basis functions of differing amplitudes and phase. This idea led to Ohm's assertion in the same time period that the ear operates as a frequency analyzer decomposing complex sounds into its constituent parts [Ohm, 1843].

Early forays into the frequency analysis of complex sounds has its roots in the physical phenomenon of sympathetic resonance in which a body that naturally resonates at a frequency f can be excited by external vibrations that are harmonically related to f . External vibrations that are identical produce the strongest effect, although octaves, thirds, and fifths generate vibrations to a lesser degree. For performing fre-

quency analysis of complex sounds, Helmholtz developed a tunable resonating device consisting of a calibrated chamber of variable volume and a small opening at one end. When placed to one's ear while a complex sound is being played, most of the frequencies will be significantly damped except for the frequencies that resonate with the chamber. By slowly varying the volume of the chamber, one can pick out the relative amplitudes of the specific frequency components of the complex sound [Helmholtz, 1853].

More recently, the field of frequency analysis has been greatly influenced by digital recording techniques and the Fast Fourier Transform (FFT) — an efficient algorithm to compute the discrete Fourier transform (DFT) and its inverse. Spectral estimates obtained by an application of the FFT assume statistical stationarity in the signal being processed, however, which is a condition that is rarely upheld in behaviorally relevant sounds such as speech. To overcome this limitation when analyzing non-stationary signals (e.g., signals containing amplitude or frequency modulation), the FFT is often applied to overlapping, windowed segments of the full signal in a process called the Short-Time Fourier Transform (STFT). The result of this process is a time-frequency representation of the signal called a spectrogram.

There are, however, significant limitations to this estimation approach. For example there is a time-frequency resolution trade-off in the STFT which limits the accuracy of frequency estimates even in noise-free non-stationary signals [Manolakis et al., 2005]. Another limitation is the difficulty of estimating the amplitudes, and phases of noisy signals. These variables are required to fully characterize the signal.

State-space signal processing methods offer an opportunity to overcome these limitations while providing new possibilities for the re-synthesis and signal morphing requirements of auditory neurophysiological experiments. The application of state-space methods to track the amplitudes, phases, and frequencies of quasi-periodic

signals in noise was pioneered in [Parker and Anderson, 1990], which discusses two candidate state-space signal models and the application of the Extended Kalman Filter for tracking the signal parameter values. There have been many subsequent investigations into this basic theme [White, 1993, James et al., 1994, La Scala et al., 1995, 1996, La Scala and Bitmead, 1996, Bittanti and Savaresi, 2000]. Recently there have been several proposed methods based on sequential Monte-Carlo techniques — commonly referred to as particle filters [Fischler and Bobrovsky, 2006, Dubois and Davy, 2007]. These methods hold promise to overcome a major shortcoming of the EKF approaches which are based on local linearizations and often have trouble regaining lock once they loose track during the iterative process.

The research reported in this dissertation investigates the use of state-space methods for analyzing and synthesizing vocalizations, and highlights the benefits of these methods over the current best practice in the field based on spectrographic analysis. Appendix A provides an in depth description of state-space signal modeling and its applicability to experiments in the field of auditory neurophysiology. Appendix B discusses the theory and application of the Extended Kalman Smoother in this domain, while Appendix C discusses the theory and application of Particle Filters.

2.4 Spectro-Temporal Modeling of Neural Responses

A Volterra series is a model for nonlinear behavior whose output depends on the input to the system at all other times, unlike a Taylor series expansion, which is memoryless. A continuous time-invariant system with input $x(t)$ and output $y(t)$ can be expanded into a Volterra series as:

$$y(t) = k_0 + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} k_n(\tau_1, \tau_2, \dots, \tau_n) x(t - \tau_1) x(t - \tau_2) \dots x(t - \tau_n) d\tau_1 d\tau_2 \dots d\tau_n$$

where k_n is called the n^{th} order Volterra kernel. The first-order Volterra functional is the most familiar:

$$K_1[x(t)] = \int_{-\infty}^{\infty} k_1(\tau_1)x(t - \tau_1)d\tau_1 \quad (2.1)$$

and is the standard time-domain description of a linear time-invariant system with impulse response $k_1(\tau)$ [Volterra, 1930]. In order to simplify the calculation of the Volterra coefficients, Wiener created a useful series of functionals that are orthogonal with respect to white Gaussian noise (WGN) input [Wiener, 1958]. Using this set of functionals, the Wiener coefficients can be calculated using the cross-correlation method by applying GWN to the input of a nonlinear system and computing a series of first- and higher-order input-output cross-correlations [Lee and Schetzen, 1965, Gill et al., 2006].

These techniques were first used by de Boer to examine the linear transfer functions of sensory neurons using a method called reverse correlation [de Boer and Kuyper, 1968]. Reverse correlation can be used to simply and intuitively calculate the first-order Wiener kernel of a sensory neuron using GWN stimulus. The stimulus waveforms preceding each action potential fired by the neuron are averaged together to generate what is called the spike-triggered average. Due to the uncorrelated nature of the GWN input into the system, only the components of waveforms that are consistently present preceding an action potential remain in the estimated kernel in a process that increases in accuracy as the data record increases. This technique was successfully used by de Boer to model the linear transfer function of auditory nerves using low to moderate intensity GWN stimulus [de Boer and de Jongh, 1978]. After the computation of the kernel (called a revcor function by de Boer), this model was able to accurately predict the firing rate of the modeled neuron to novel stimulus.

There are, of course, numerous drawbacks to this technique. First of all, the data

records must be of sufficient length to accurately estimate kernel coefficients. Since there is an exponential increase of data required to estimate each additional order of the kernel, experiments are limited by how long a single neuron can be recorded from (which is itself dependent on what type of neuron is being recorded and the specific recording techniques). In addition to these practical considerations, there are theoretical constraints including the dependence of the estimated kernel on the parameters of the GWN stimuli and the lack of a convergence guarantee when the system has a point process output (e.g., spike trains) [Johnson, 1980]. Nonetheless, these techniques have provided insight into the neural encoding of sensory information and have been applied to other peripheral sensory systems such as vision [Rieke et al., 1999] and electrolocation [Gabbiani and Metzner, 1999] with significant success.

Although reverse correlation techniques have been successfully used at the periphery of the auditory system [de Boer and de Jongh, 1978], the first-order Wiener kernel is negligible if the response of the neuron is not precisely phase locked to the stimulus waveform [Eggermont, 1993]. This condition is largely held for auditory nerves tuned to frequencies below about 4-6 kHz, but becomes increasingly less so for neurons tuned to higher frequencies and neurons further from the periphery [Kim and Young, 1994]. Due to this limitation, focus has been brought to the second-order Volterra-Wiener functional:

$$K_2[x(t)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k_2(\tau_1, \tau_2) x(t - \tau_1) x(t - \tau_2) d\tau_1 d\tau_2 \quad (2.2)$$

K_2 can be interpreted as the linear transfer function between the output of the system and autocorrelation of the input, which is closely related (through a single Fourier transform) to a large class of time-frequency representations via the Wigner distribution [Eggermont, 1993]. Based on the tonotopic organization of auditory nuclei and

the substantial lack of phase locking of most neurons to the stimulus waveform, the second order kernel based on a spectro-temporal representation of auditory stimuli has been rigorously pursued [Aertsen and Johannesma, 1981, Eggermont, 1993, Eggermont et al., 1983]. This kernel, called the spectro-temporal receptive field (STRF), represents the best linear fit between a time-frequency decomposition of auditory stimuli and the firing rate of an auditory neuron.

STRF methodologies have been used to model neurons throughout all levels of the ascending auditory pathway from the auditory nerve [de Boer and de Jongh, 1978, Kim and Young, 1994, Young and Calhoun, 2005] to the auditory forebrain [Sen et al., 2001]. Many approaches have been taken for the spectro-temporal representation of auditory stimuli, including spectrographs, wavelet and gamma tone transformations, and representations with adaptive gain control to more closely model the cochlear response [Slaney and Lyon, 1993, Gill et al., 2006].

Another decision that affects the fitting of the STRF model is the choice of stimuli. Historically, broad-band gaussian noise has been used to simplify the analysis of reverse correlation and spike-triggered average techniques [Aertsen and Johannesma, 1981]. Broad-band noise has limited use as a stimulus protocol, however, because many auditory neurons do not respond robustly to this stimulus [Theunissen, 2004]. To address this issue, the algorithms used to fit STRF models have been adapted to accommodate pure tones, narrow-band gaussian noise, ripple noise, and, more recently, natural stimuli such as recorded vocalizations [Klein et al., 2000, Escabi and Read, 2003, Holmstrom et al., 2007b, Theunissen et al., 2000]. Figure 2.4 shows how a STRF is generated from ripple noise and the visual representation of the STRF, indicating regions of inhibition and excitation. The STRF is generated by finding the average spectro-temporal characteristics of the auditory stimulus preceding each spike (the three spikes are seen at the top of the figure).

Figure 2.4: A STRF generated by averaging the spectro-temporal characteristics of the ripple-noise auditory stimulus preceding each spike (the three spikes are seen at the top of the figure). This figure was obtained from [Escabi and Read, 2003].

A final important model design decision concerns the methods used to fit the model and to evaluate its performance. Spike-triggered averaging techniques break down when statistical correlations exist in the stimuli. This is particularly true in the cases of pure tone and vocalization stimuli because both are highly structured. Methods have been developed to address this issue and allow for optimal model fitting even when using correlated stimuli [Theunissen et al., 2000, Holmstrom et al., 2007b].

2.5 Pure Tone STRF Models

Within the dorsal nucleus of the lateral lemniscus (DNLL), which is only one synapse below the IC, responses to vocalization stimuli can largely be predicted by a STRF model trained on pure tone stimuli [Bauer et al., 2002, Pollak et al., 2003]. It has been shown, however, that cortical neurons are not well characterized by a neuron’s pure tone responses, and that STRF models trained on pure tone stimuli generate poor predictions to novel vocalization stimuli [Theunissen et al., 2000]. This reduction in model accuracy is largely attributed to increasing nonlinearity in the mapping between external stimuli and neural responses as you progress up the neural pathways of the auditory system. These nonlinearities cannot be captured by linear STRF models unless they are static and can be explicitly included or approximated in the representation of the stimuli and/or neural response [Theunissen et al., 2000]. For example, it is common to represent auditory stimuli in terms of signal power on a logarithmic scale to approximate the compressive properties of the basilar membrane [Gill et al., 2006, Theunissen et al., 2000]. Both the power calculation and the logarithmic scaling are non-linear operations on the sound pressure wave of the stimuli.

Significant selectivity and sensitivity to complex spectro-temporal features first appears in the IC. Responses in the IC can partially be explained by pure tone STRF models, indicating that IC neurons sometimes combine spectral information

from different frequency bands in an approximately linear fashion. While nonlinear integration of spectral information by IC neurons limits the accuracy of pure tone STRF model predictions, these models are still useful for testing hypotheses about neural encoding in the IC.

In a study conducted in the IC of the Mustached Bat, we recorded the responses of 50 neurons to both pure tones and complex vocalizations [Holmstrom et al., 2007b]. Using models trained from the pure tone responses alone, we found that predicted responses to complex vocalizations were poor in the NMSE sense, but captured the salient qualities of the recorded responses in 82% of the neurons for more than 80% of the vocalizations used in our stimulus protocol. By focusing on the linear predictions obtained from the STRF models, we were able to identify the specific frequency bands that contributed to the recorded responses. This study resulted in two significant findings. First, we found that IC neurons in the Mustached Bat often have complex frequency response areas (FRAs) with excitatory and inhibitory regions far from the neuron’s primary tuning region. This is in stark contrast to peripheral auditory neurons which are characterized by a single, v-shaped FRA. Second, we found that these secondary tuning regions contribute, and are sometimes solely responsible for, neural selectivity and sensitivity to complex social vocalizations (as opposed to the echolocation vocalizations emitted by these bats). This provided evidence of dual purpose neurons in the IC which have a sharp primary tuning around 60 kHz for echolocation encoding, and a broad tuning at lower frequencies for encoding social vocalizations. A visualization of this result for a representative neuron can be seen in Fig. 2.5. Note in Fig. 2.5.E1-6 that the portion of the model covering the primary tuning range of the neuron does not contribute to the responses to the vocalization stimuli (dashed lines). However, the full model, which also includes the secondary tuning region(s), captures the salient qualities of the responses (solid lines).

Figure 2.5: Comparison of predicted responses to social vocalizations of the primary (BF) and extended (full) models for one neuron. A. The tuning curve for the neuron. This neuron had sharp tuning in the 60 kHz range (the BF for the neuron) as well as broad tuning in the 10-30 kHz range. B. The predicted pure tone responses for a 61 kHz tone (top) and a 20 kHz tone (bottom), showing a good NMSE fit of the model to the pure tone data. C. The primary tuning model, which spans 50-68 kHz and indicates a strong response to 61 kHz stimulus. D. The extended model for the neuron, indicating additional strong excitation in the 10-30 kHz range. E1-6. Responses of the neuron to six different social vocalizations. Overlaid on the responses are the predictions from the primary (BF) and extended (full) models.

For the study highlighted in this dissertation, pure tone STRF modeling of IC neurons again plays an important role. Instead of focusing on the similarities between the recorded responses and those predicted by the pure tone STRF models, as we did in the Mustached Bat study, we focused on the differences. We attribute these differences to nonlinearities introduced at the level of the IC that are not present in peripheral auditory neurons, and use the model predictions as approximations of the linear responses expected in peripheral neurons.

2.6 Medical Significance

Profoundly deaf individuals who still have an intact auditory nerve have profited from the dramatic advances made over the past 30 years in the field of cochlear implants (CIs). The CI is a microelectrode array implanted in the cochlea that directly stimulates the auditory nerve. This technology has helped more than 40,000 individuals achieve varying levels of speech perception, including telephone communication and unassisted classroom participation [Rauschecker and Shannon, 2002]. CIs are not an option for those individuals whose deafness is caused by lesions beyond the cochlea, however. This includes patients with neurofibromatosis type 2 (NF2), a genetic disease that occurs in about 1 in 40,000 births [Evans et al., 2000]. NF2 produces Schwann cell tumors along afferent nerve tracts as they enter the brainstem and spinal cord. One of the defining symptoms of NF2 is the growth of bilateral tumors along the eighth cranial nerve (composed of the auditory and vestibular nerves). Removal of the tumors almost always necessitates a transection of the auditory nerve which results in total deafness. If the tumors are not removed, they produce compression of the brainstem that is ultimately fatal [Rauschecker and Shannon, 2002].

Auditory Midbrain Implants (AMIs) are becoming a feasible solution for patients suffering from sensorineural hearing loss where CIs are not an option [Lenarz et al.,

2006a,b, Samii et al., 2007]. Rather than stimulating the auditory nerve, AMIs stimulate the inferior colliculus central nucleus (ICC) of the midbrain. The primary motivation behind the choice of the ICC for auditory stimulation is threefold. First, the ICC serves as a critical point of convergence for auditory information ascending from the brain stem to the auditory cortex [Aitkin and Phillips, 1984, Ehret and Romand, 1996]. Second, the ICC has a highly organized tonotopic structure [Merzenich and Reid, 1974]. This should allow an implant to stimulate different frequency regions independently and provides a first-order approximation of the mapping between auditory stimulus at the periphery to neural excitation in the ICC. Third, for NF2 patients, AMI implantation can be performed after tumor removal with minimal added risk [Lenarz et al., 2006a,b].

While many of the surgical and mechanical aspects of this treatment are developed well enough for clinical trials [Lenarz et al., 2006a, Lim et al., 2007], the current practice of stimulating the ICC based on its tonotopic gradient alone (similar to the stimulus protocol used with CIs) largely ignores the functional role played by the ICC and the ascending auditory pathways of the brainstem and midbrain between the cochlea and the ICC. This is primarily due to the lack of knowledge of the role played by this system of nuclei in the processing of complex speech stimulus.

3 Methods

This chapter and the appendices highlight the methodologies used in this research. Our goal was to characterize the selectivity, sensitivity, and heterogeneity of responses in the mouse IC to vocalization stimuli, and to explore whether this characterization provides evidence of a progressive and efficient encoding strategy within the auditory system. Some of these methods, such as the surgical procedures and the recording techniques, were not pioneered for this research and have already been used in numerous other studies [Portfors, 2004, Portfors and Felix, 2005, Holmstrom et al., 2007b, Portfors et al., 2009]. Other methods presented here already existed in the literature, but are new and are just starting to be explored. These methods include similarity metrics of neural responses based on spike timing and the application of these metrics to estimate information theoretic measures of efficiency for individual neurons [Victor and Purpura, 1996, Huetz et al., 2009]. A significant number of the methodologies were developed specifically for this research, however, such as the application of state-space methods for stimulus design, sensitivity analysis of neural responses to perturbed vocalizations, pure tone modeling of peripheral responses, and information theoretic measures of heterogeneity [Holmstrom et al., 2010]. As a whole, these methods offer a comprehensive characterization of neural responses in the mouse IC to behaviorally relevant vocalization stimuli.

3.1 Animals

Twenty-four female CBA/CaJ mice, 7-27 weeks of age, were used in this experiment. The animals were housed with littermates of the same gender until the surgical procedure was performed. The mice were kept under a reversed 12:12 hours light-dark

regime so that the electrophysiological recordings occurred during their awake period. Food and water were provided ad libitum. All animal care and experimental procedures were in accordance with the guidelines of the National Institutes of Health and approved by the Washington State University Institutional Animal Care and Use committee.

3.2 Surgical procedures

The surgical procedures used in this study have previously been described in detail [Felix and Portfors, 2007]. Briefly, mice were anesthetized with isoflurane inhalation and a midline incision was made in the scalp. The skin was reflected laterally, and lidocaine gel was applied topically to exposed tissue. The positions of the left and right IC under the skull were located using stereotaxic coordinates and marked with India ink to guide the subsequent craniotomy. A tungsten ground electrode was inserted into the cerebral cortex, and a hollow metal rod was cemented onto the skull. The animal was allowed to recover from general anesthesia for at least one hour before the electrophysiology experiment began. Typically, experiments were conducted one day after surgery.

At the start of electrophysiological recordings, the mouse was given a mild sedative (acepromazine, 5 mg/kg injected intraperitoneally) and restrained in styrofoam molded to the animal's body. The headpin was bolted to the custom-designed stereotax in a sound-attenuating chamber. Using the stereotaxic coordinates, a craniotomy was made over one IC and the dura was removed. A glass micropipette electrode (impedance 8-20 M Ω) filled with 1 M NaCl was positioned over the IC. Electrode advancement through the IC was controlled by a hydraulic micropositioner (David Kopf Instruments, model 650D) located outside the sound-attenuating chamber. To protect the brain from dehydration, the hole in the skull was covered with petroleum

jelly during and between recording sessions. A supplemental dose of acepromazine (2.5 mg/kg i.p.) was given to animals that continued to struggle or showed other signs of discomfort. Animals that showed continued discomfort were removed for the day, however most mice remained calm throughout a recording session. Each session typically lasted 4-8 hours. Data was collected from the same animal on two or three consecutive days, but never more than twice in the same-side IC.

3.3 Stimulus Design

One of the contributions of the reported research to the field of auditory neurophysiology is the development of new methodologies for developing vocalization based stimulus protocols. These methodologies leverage statistical state-space signal models and state-space analysis methods for estimating the parameter values of these models. This approach to stimulus design provides three primary advantages over the current best practices in the field, which are based on spectrographic analysis and synthesis. The first of these is a means for efficiently and accurately de-noising and removing unwanted artifacts from vocalization recordings. The second advantage is the ability to provide better estimates of the frequency, phase, and amplitudes of the harmonic components of a vocalization, particularly in noisy recordings. The third advantage is the use of a flexible signal model which lends itself to numerous ways of generating modified versions of the vocalizations. This flexibility is critical for testing the sensitivity and selectivity of neural responses to behaviorally relevant stimuli. Details about these benefits are documented in [Appendix A](#).

For this study, four representative vocalizations ([Fig. 3.1](#)) commonly emitted by adult CBA/CaJ mice during social interactions were chosen from a collection of calls recorded in our lab [[Portfors, 2007](#)]. The 30 kHz harmonic, 40 kHz harmonic, and male upswing vocalizations were recorded during male-female pairings and are most

likely emitted by the male. The female upsweep vocalization was recorded from a female directly after her litter was removed from the nest. The 30 kHz harmonic vocalization (Fig. 3.1A) spans a frequency range of 28-74 kHz and is 118 ms in duration. The first (and only) harmonic contains significantly more power than the fundamental. The frequency contour of the vocalization starts low, builds to a peak at about 30 ms, and slowly falls until the vocalization terminates. The 40 kHz harmonic vocalization (Fig. 3.1B) spans a frequency range of 40-103 kHz and is approximately 85 ms in duration. This vocalization also has a single harmonic and has a low-high-low frequency contour. The majority of the spectral power, however, is found in the fundamental for the middle 66 ms of the vocalization. These two vocalizations were chosen because they were similar in duration and spectral structure, but had different frequency ranges. The female upsweep vocalization (Fig. 3.1C) spans a frequency range of 82-98 kHz and is 22 ms in duration. There is no harmonic structure and the frequency rises linearly. The male upsweep vocalization (Fig. 3.1D) spans a frequency range of 70-100 kHz and is 19 ms in duration. There is no harmonic structure, the frequency remains constant for the first 7 ms and then rises linearly, and most of the spectral power lies between 80-100 kHz. These two vocalizations were chosen because they are more simple than the harmonic vocalizations, are very similar in spectro-temporal content to one-another (with the exception of differences in amplitude modulation), and are emitted by opposite genders of CBA/CaJ mice.

These vocalizations were analyzed by extracting the time-varying frequency, amplitude, and phase information of their fundamental and harmonic (if present) components with custom-written Matlab code implementing a harmonic state-space signal model and the Extended Kalman Smoother [Holmstrom et al., 2009]. This enabled the synthesis of the original vocalizations in the absence of background noise (Fig. 3.2A,B). These synthesized yet unmodified vocalizations are referred to as the

Figure 3.1: Spectrograms (top) and oscillograms (bottom) of the natural ultrasonic mouse vocalizations used in the study. **A**, **B** and **D** were recorded during male-female pairings and were most likely emitted by the male, whereas **C** was emitted by a female in isolation after her pups were removed. All vocalizations were synthesized from the original recordings in order to remove excessive background noise.

natural vocalizations throughout this study. Furthermore, by manipulating the extracted parameter values, a set of variants were generated for each vocalization. Vocalizations were upshifted or downshifted in frequency by 5, 10, 15, and 20%. Duration was doubled and halved (Fig. 3.2C), amplitude modulation (AM) was removed (Fig. 3.2D), and frequency modulation (FM) was removed (Fig. 3.2E). For two vocalizations, harmonic structure was disrupted by removal of either the fundamental or the harmonic frequency components. In addition, the fundamental component was upshifted by 5-25% in 5% increments relative to the harmonic frequency while holding the harmonic component constant, or the harmonic frequency was downshifted in identical steps while the fundamental frequency was kept constant (Fig. 3.2F). Details of these analysis and synthesis methods can be found in Appendices A and B.

3.4 Acoustic stimulation

Stimulus generation was controlled by custom-written software on a PC located outside of the sound-attenuating chamber. The stimuli were fed into a high-speed D/A converter (Microstar Laboratories, 400,000 samples/s), a programmable attenuator (Tucker-Davis Technologies, model PA5), a power amplifier (Parasound, model HCA-1000A), and a leaf tweeter speaker (Infinity). Acoustic stimuli were presented in free-field; the speaker was positioned 10 cm from the ear contralateral to the IC under investigation. The speaker output was calibrated using a quarter inch condenser microphone (Bruel & Kjaer, model 4135) placed at the position that would normally be occupied by the ear of the animal. Sound pressure levels were obtained for frequencies between 6-100 kHz at 0 dB and 10 dB attenuation. Sound pressure levels showed a gradual decline of 3.2 dB per 10 kHz. In our data analysis, a correction was applied based on this decline. A fast Fourier transform of the acoustic signals was done to check for harmonic distortion in the output signals. Distortion products were buried

Figure 3.2: Analysis and synthesis of the 40 kHz harmonic vocalization. (A) Background noise and small recording artifacts are present in the original recording. (B) Synthesis of the signal faithfully regenerated the harmonic content of the vocalization in the absence of the noise present in the original recording. (C) Duration doubled without altering the frequency. (D) AM removed without altering the FM. (E) FM removed without altering the AM. (F) An inharmonic vocalization generated by multiplying the fundamental frequency by 150%.

in the noise floor (> 50 dB below signal level), and therefore considered negligible.

3.5 Data acquisition and analysis

Extracellular action potentials were recorded from well-isolated single neurons. All neurons were assumed to be located in the central nucleus of the IC based on the stereotaxic coordinates and penetration depth of the recording sites. Before digitization (Microstar Laboratories, 10,000 samples/s), spikes were amplified (Dagan, model 2400), and band-pass filtered (Butterworth 600 Hz-6 kHz, Krohn-Hite, model 3364). Data collection software was custom-written and provided real time displays of spike trains, post-stimulus time histograms (PSTHs), raster plots, and statistics (see Appendix D). Raw waveforms were stored for offline analysis and examined after recordings to ensure that only well-isolated single units were included in the data analysis. Pure tones varying between 6 and 60 kHz (100 ms duration, 1 ms rise-fall time) at intensities of 10-70 dB SPL were used as search stimuli. Once a single unit was isolated, characteristic frequency (CF) and minimum threshold were determined audio-visually. CF was defined as the frequency that evoked spikes to 50% of the stimulus presentations at the lowest intensity, and minimum threshold was defined as the lowest intensity that evoked spikes to 50% of the stimulus presentations at CF. Tone burst stimuli (50-150 ms duration, 1 ms rise-fall time, 3-4 Hz repetition rate, 200-300 ms recording window) from 6 to 100 kHz were presented in 1-2 kHz steps at 10, 30, and 50 dB above threshold to construct FRAs. Each frequency-intensity pair was presented 20 times.

All vocalization variants (2 ms rise-fall time, 20 repetitions at a rate of 3-4 Hz, 200-300 ms recording window) were then presented to the neurons at 60-80 dB SPL in 10 dB increments. This range was chosen because the vocalizations are emitted at these intensities in laboratory settings. In order to classify a neuron as being responsive

to a particular vocalization, the neuron had to fire in response to at least 50% of the stimulus presentations and these spikes could not be attributable to spontaneous activity (determined audio-visually).

3.6 Spike Metrics

Three metrics were chosen to objectively evaluate neural responses to the natural and perturbed vocalizations; one for measuring temporal consistencies among a neuron’s responses to repeated presentations of an identical stimulus, one for measuring temporal similarities between a neuron’s responses to two different stimuli, and the other for measuring firing rate similarities between a neuron’s responses to two different stimuli.

The first, based on the temporal encoding hypothesis, assumes that the timing of spikes conveys important information. This hypothesis is grounded in observations of phase-locked responses throughout the auditory system, originating in the auditory nerve [Galambos and Davis, 1943] and extending to auditory cortex [Steinschneider et al., 1999]. Information theoretic analyses of real and simulated sensory neurons have quantified how variability in spike timing reduces the information content of the neural code [Bialek et al., 1991, Rieke et al., 1993, Rieke et al., 1995, Rokem et al., 2006, Shamir et al., 2007].

The *correlation-based similarity measure* (R_X), as proposed in [Schreiber et al., 2003] and used in [Wang et al., 2007, Huetz et al., 2009], measures how well two binary spike trains of equal duration are temporally correlated. The temporal similarity between two binary spike trains, $\vec{b}_1$ and $\vec{b}_2$, is found by first convolving each spike train with a Gaussian kernel with mean 0 and standard deviation σ . The correlation

between the two filtered spike trains, $\vec{x}_1$ and $\vec{x}_2$ can then be calculated as follows:

$$R_x = \frac{\vec{x}_1 \cdot \vec{x}_2}{\|\vec{x}_1\| \|\vec{x}_2\|}$$

This results in a temporal similarity measure that is bound between 0 and 1 and is dependent on σ . A value of 0 indicates that the responses are uncorrelated, while a value of 1 indicates perfect correlation (the timing of the spikes in the two responses are identical). Two spike trains with zero spikes each are assumed to be uncorrelated and assigned a value of 0. As σ increases from zero, the metric becomes less and less sensitive to differences in spike timing. In this study, σ was set equal to 3 ms. This value was chosen partially based on the results of Wang, et al.(2007). In our own analysis, smaller values of σ consistently resulted in negligible values of R_x even for similar spike trains. Higher values of σ , on the other hand, were not sensitive enough to temporal spike jitter and reduced the contrast between responses that were phase locked and those that were not.

Because a single vocalization was presented a total of N times to a neuron, generating N spike trains, the response consistency for each neuron was calculated as:

$$R_X = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N \frac{\vec{x}_i \cdot \vec{x}_j}{\|\vec{x}_i\| \|\vec{x}_j\|} \quad (3.1)$$

which is equal to the mean of all possible distinct and non-redundant correlations of the N spike trains and excludes all self-comparisons ($i \neq j$). This metric is again bound between 0 and 1 and indicates how similar the N responses were. A perfectly consistent response indicates that each presentation of the vocalization evoked spikes that all occurred at exactly the same time.

When two distinct vocalizations were each presented N times to a single neuron,

the cross-correlation metric for each neuron was calculated as follows:

$$R_{XY} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \frac{\vec{x}_i \cdot \vec{y}_j}{\|\vec{x}_i\| \|\vec{y}_j\|} \quad (3.2)$$

which is equal to the mean of all possible cross-correlations between the two sets of N spike trains and is again bound between 0 and 1. This metric indicates the temporal similarity of an individual neuron's responses to two different vocalizations.

For classification purposes, it was important to specify a condition for which a neuron's response to one vocalization was considered temporally different than the response to a different vocalization. Because a high spontaneous rate to either variant may result in a high response similarity, a simple threshold on R_{XY} does not satisfy this requirement. Based on a combination of response consistency and response similarity, our temporal similarity measure was defined as

$$S_{XY} = \frac{R_{XY}}{R_X + R_Y} \quad (3.3)$$

This measure addresses bias implicit in R_{XY} when neurons have high spontaneous firing rates. Our objective classification rule for determining whether a response to vocalization X was temporally altered relative to vocalization Y was given by

$$S_{XY} \leq 0.25 \quad (3.4)$$

and resulted in very similar classification results as those obtained audio-visually.

Other metrics for comparing the temporal qualities of binary spike trains were also investigated, such as the Victor-Purpura spike timing metric [Victor and Purpura, 1996] and the Van Rossum spike distance [van Rossum, 2001]. We chose the correlation based spike metrics because they are normalized. This quality is important for

the current study because a primary goal was to calculate and compare these metrics across populations of neurons with differing response properties.

The third metric, based on the rate encoding hypothesis, assumes that the total number of spikes generated by a neuron conveys important information about the acoustic stimuli. This hypothesis is grounded in observations of rate encoding originating in the auditory nerve [Galambos and Davis, 1943] and is the foundation of the ubiquitous frequency response area; a measure of neural selectivity for pure tone stimuli used in all auditory nuclei. Information theoretic analysis of real and simulated neurons have also measured the information capacity of neurons using a rate code [Stein et al., 1972, Bialek et al., 1991].

The *selectivity index* [Wang and Kadia, 2001, Philibert et al., 2005, Castellanos et al., 2007] was chosen to quantify whether a neuron fired more often in response to a natural vocalization as opposed to one of its variants. This metric completely disregards the temporal qualities of the neural response $\vec{b}$ and was calculated using the estimated spike rate r of $\vec{b}$ over the complete recording window. The selectivity index d was given by:

$$d = \frac{(r_1 - r_2)}{(r_1 + r_2)} \quad (3.5)$$

where r_1 is the spike rate of the response to the natural vocalization, and r_2 is the spike rate of the response to the modified vocalization. This metric is equal to the normalized spike rate difference and is bound between -1 and 1. A value of 1 indicates that the neuron responded only to the natural vocalization, a value of -1 indicates that the neuron responded only to the modified vocalization, and a value of 0 indicates no difference in the spike rate between the two responses. If both r_1 and r_2 are zero, they are assumed not to differ and are assigned a value of 0.

Because two distinct vocalizations were each presented N times to a neuron, the

selectivity index for the neuron’s responses was calculated as follows:

$$D = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \frac{(r_{1,i} - r_{2,j})}{(r_{1,i} + r_{2,j})} \quad (3.6)$$

which is equal to the mean spike rate distance between the two sets of N spike trains and is bound between -1 and 1. A value above 0 indicates an elevated spike rate in response to the natural vocalization.

Using the above spike metrics, three measures were used to quantify and compare the population response characteristics to each vocalization variant: the median response consistency to the natural vocalization, $\overline{R_X^N}$, the median response consistency to each modified variant, $\overline{R_X^M}$, and the mean selectivity index for each modified variant, $\overline{D}$. The median was used to characterize R_X^N and R_X^M due to the presence of outliers and the non-normality of these distributions. The bootstrap method was used to place 95% confidence intervals on the estimate of the medians. Non-parametric, two-sample Kolmogorov-Smirnov tests were performed to determine whether distributions of R_X values were statistically different. To determine whether the distributions of D were statistically different from the responses to the natural vocalization two-tailed, two-sample t-tests were used and the standard deviation of the distribution was used to generate the confidence intervals.

3.7 Comparing responses in the IC to responses in lower auditory nuclei

One of the goals of the study was to measure whether responses in the IC to complex vocalization stimuli are more selective and heterogeneous than responses in lower auditory nuclei. Because the IC was the primary focus of this study and it required a year of data collection to record from 60 IC neurons that responded to vocalization stimuli, additional recording from other nuclei would have either required significantly

more recording time or a drastic reduction in our sample size of IC neurons. This made it intractable to perform a head to head comparison of responses from different nuclei. To address this, we used a model based approach grounded in results showing that responses to complex vocalizations by neurons in most lower nuclei can accurately be predicted by the pure tone responses of these neurons. At the periphery, this has been shown to be true for auditory nerve fibers [Sachs et al., 1983]. Within the dorsal nucleus of the lateral lemniscus (DNLL), which is only one synapse below the IC, responses to vocalization stimuli can also be explained by a linear summation of responses to pure tones [Bauer et al., 2002, Pollak et al., 2003]. This is strong evidence that selectivity to complex features of vocalizations does not exist within these nuclei. Furthermore, these studies reported that different cells with similar CFs responded to vocalizations with very similar latencies and firing patterns, resulting in a homogeneous and redundant encoding among these neurons. Based on these results, we used the pure tone responses of neurons in the IC to predict how neurons with identical FRAs would respond in lower auditory nuclei and compared these modeled responses to the actual recorded responses.

For each neuron, a model was optimized to approximate the relationship between the pure tone input and the resulting firing rate of the neuron, as approximated by the PSTH. The model was a discrete (in both frequency and time) linear finite impulse response (FIR) filter h_i such that:

$$\hat{r}(t) = c + \sum_{i=1}^{n_f} \sum_{j=1}^{n_t} h_i(j) s_i(t-j) \quad (3.7)$$

where $\hat{r}(t)$ was the predicted time-varying firing rate, c was the spontaneous firing rate of the neuron, i was the frequency band index of the stimulus, j was the time lag index, $s_i(t)$ was the discrete time-frequency representation of the time-varying

stimulus, n_f was the number of frequency bands, and n_t was the number of time lag indices. The parametrization of h_i can be interpreted as a Spectro-Temporal Receptive Field (STRF) [Aertsen and Johannesma, 1981, Eggermont et al., 1983, Theunissen et al., 2001, Gill et al., 2006], although our implementation of this model includes key differences from common STRF methods (see Section 2.4 for background on STRF methodologies). After fitting the model, arbitrary stimuli (e.g., complex vocalizations) can be converted into a spectrographic representation and convolved with the filter, generating a predicted response.

By choosing the linear FIR model, we made a number of simplifying assumptions about the stimulus/response relationship of the neurons in the inferior colliculus of the awake mouse. One of these assumptions is that each neuron receives inputs from n_f distinct frequency channels, each of which can be modeled independently and may be either excitatory or inhibitory. This assumption is partially justified by the fact that incoming sounds are approximately decomposed into individual frequency components at the point of transduction in the cochlea and that different frequency channels have been shown to have excitatory and/or inhibitory effects on the neural response. Another assumption is that the modeled responses of each frequency channel are linearly summed to produce the overall firing rate of the neuron in response to the stimulus. These limitations are understood and accepted and the model is assumed to be the best linear transfer function between the spectro-temporal characteristics of the pure tone training data and the neural responses.

Even with these simplifying assumptions in place, a number of important design decisions were made in order to implement this model. These included the time-frequency representation of the input, the temporal resolution of the output, and the algorithm used to estimate values of the model parameters h_i and c given an optimality criterion for the model predictions. Many approaches have been taken

for the spectro-temporal representation of auditory stimuli, including spectrographs, wavelet and gamma tone transformations, and representations with adaptive gain control to more closely model the cochlear response [Slaney and Lyon, 1993, Gill et al., 2006]. The model described here used a spectrographic representation of the input.

A two step process was used to calculate the optimal model parameter values for each neuron. In the first step, the parameter values for each model were chosen such that the normalized mean squared error (NMSE) between the predicted firing rate, $\hat{r}(t)$, and the actual firing rate, $r(t)$, was minimized:

$$\text{NMSE}_{\text{linear}} = \frac{1}{\sigma_r^2} \sum_t (r(t) - \hat{r}(t))^2 \quad (3.8)$$

where σ_r^2 was the variance of the actual firing rate of the neuron.

For each time step t of each stimulus presentation, the spectrogram of the stimulus signal was calculated for the current time step back to the maximum time lag, resulting in an $n_f \times n_t$ time-frequency (spectrographic) representation of the stimulus. Consecutive $n_f \times n_t$ blocks extracted in this fashion will overlap in $n_t - 1$ columns. This spectrogram was reshaped to form a single $n_f \cdot n_t$ length row of the input matrix P . For each time step t , the neuron's normalized PSTH at time t became an entry in the output column vector O . Each row of P composed entirely of zeros was removed, along with the corresponding entry in O , because they played no part in the calculation of the parameter values and their removal reduced the complexity and memory requirements of the optimization procedure.

Optimal values of h_i and c were uniquely determined for the optimality criterion defined in (3.8) by solving the Normal Equations [Manolakis et al., 2005]. From P and O we calculated the time averaged correlation matrix R of P and the cross correlation

vector d of P and O . The unbiased estimate was used for each of these calculations. The relationship between R , d , and the optimal LMMSE model parameters, a , could be written in the form of the Normal Equations:

$$Ra = d \quad (3.9)$$

The unique optimal parameter vector, a , which encompasses both h_i and c from (3.7) was then solved for:

$$a = R^{-1}d \quad (3.10)$$

For the second step of model optimization, a nonlinearity was introduced into the criterion function which forces negative predicted rates to be equal to zero:

$$\text{NMSE}_{\text{nonlinear}} = \frac{1}{\sigma_r^2} \sum_t (r(t) - H(\hat{r}(t)))^2 \quad (3.11)$$

where $H()$ is the Heaviside function. Using the model parameter values found in the first step as a starting point, a gradient descent method was then employed to further refine the fit. This nonlinearity (in the optimality criterion, not the model) was introduced to reflect the interpretation of negative predicted firing rates as a state of variable inhibition in the neuron. Since the measured PSTH's of the neurons could never themselves be negative, negative model output is seriously penalized by the LMMSE fit. This results in a model that under-predicts the responses of the neurons. Due to this, we chose an optimality criterion that would better match the always positive PSTH's of the neurons at the expense of, perhaps, over-emphasizing the state of inhibition in the cell.

For each training epoch (one pass through the input/output data set) the gradient of the error surface (defined by (3.11)) with respect to each model parameter in a ,

Figure 3.3: A comparison of model optimization using the linear 3.8 and nonlinear 3.11 optimality criterion. As can be seen, the model optimized using the nonlinear criterion provided a much better fit to the recorded PSTH of the neuron.

was determined. a was then updated iteratively in the direction of this gradient to reduce the error with respect to the new optimality criterion:

$$a_{i+1} = a_i + l \nabla_a \quad (3.12)$$

where l is the learning rate of the algorithm and ∇_a is the gradient of the error surface with respect to the model parameters. A simulated annealing approach was used which gradually lowered the learning rate l as the training proceeded. Figure 3.3 shows a comparison of the LMMSE fit and the nonlinear fit for the modeled response to a 60 kHz pure tone in an auditory neuron of a mustached bat from a previous study. As can be seen, the nonlinear error criterion and gradient descent method provides a much better fit to the PSTH of the actual response. In two different modeling exercises (one using bat vocalizations and the other using mouse vocalizations), the non-linear optimization step significantly improved the accuracy of the models with respect to both the pure tone training data and the generalization of these models to novel vocalization stimuli.

3.8 Discrimination efficiency among vocalization variants

To address how efficiently individual neurons discriminated among the modified variants of a vocalization we asked the question “provided a single spike train recorded from a given neuron, can we correctly classify which vocalization variant generated it?”. To quantify the discrimination efficiency of the encoding strategy employed by the IC, we used a measure of information transfer between stimuli and neural responses based on a classification task dependent on a chosen measure of similarity [Victor and Purpura, 1996, Huetz et al., 2009]. Given all of the responses of a neuron to each variant of a vocalization, a single response was removed from this set. Classification was then performed by measuring which vocalization variant’s responses were, on average, most similar to the held-out spike train based on the similarity measure S_{XY} defined in (3.3). By iterating over each spike train from each of the N_{vocs} vocalization variants, a confusion matrix M was generated. Each element of this matrix, $M_{i,j}$, indicated how many spike trains generated from vocalization i were classified as being generated by vocalization j . In the case of a tie, an equal value was added to each tying entry in M , summing to one. Using this confusion matrix, the neuron’s information transfer H was calculated:

$$H = \frac{1}{M_{total}} \sum_{i,j} [M_{i,j}(\log M_{i,j}) - \log \sum_a M_{a,j} - \log \sum_b M_{i,b} + \log M_{total}] \quad (3.13)$$

where M_{total} was the total number of spike trains being classified and H was bound between 0 and $\log_2 N_{vocs}$. If the neuron’s responses to each vocalization variant were temporally consistent yet distinct (low similarity values S_{XY}), spike trains were relatively easy to classify, M approached a diagonal matrix, and H was maximized. If spike trains were difficult to classify due to weak or highly correlated responses, the

values of $M_{i,j}$ approached a uniform distribution and H was minimized. For each neuron, a different value of H was calculated to indicate how efficiently it discriminated among the variants of a specified vocalization.

To assess how well the population of neurons could discriminate among the set of vocalizations, a similar experiment was performed where, for a given stimulus, a composite response was generated consisting of one response from each neuron. Since 20 recordings were made from each neuron to each vocalization, we created 20 composite population responses to each vocalization. The information transfer of these composite responses, H_{pop} , could then be calculated to measure the efficiency of the population encoding as a whole at the task of discriminating among the vocalization variants.

To correct for the bias H_{bias} inherent in the calculation of H using a finite amount of recorded data, H_{bias} was estimated using a shuffling technique [Victor and Purpura, 1997, Huetz et al., 2009]. The spike trains for a given neuron were randomly assigned to the different vocalization variants and the information transfer analysis was performed as described above. This shuffling process was repeated 100 times to calculate the mean ($\overline{H_{bias}}$) and standard deviation of H_{bias} . Only information transfer values greater than $\overline{H_{bias}} + 2SD$ were considered significant.

3.9 Quantification of response heterogeneity across IC neurons

A similar information theoretic analysis was performed to address the heterogeneity of responses to vocalizations across the neural population. This analysis focused on the question “provided a single spike train recorded in response to a given vocalization, can we correctly classify which neuron it was recorded from?”. If the responses of all neurons to a given vocalization were weak or redundant, classification was difficult and the information transfer H was low. As the heterogeneity of the responses

increased, so did H . The methodology for this analysis was similar to the vocalization discrimination efficiency analysis performed above. In this case, however, one value of H was calculated for each vocalization variant presented in the study (instead of one for each neuron in the study). H was bound between 0 and $\log_2 N_{neurons}$, where $N_{neurons}$ was the number of neurons responding to the vocalization class (e.g., 30 kHz harmonic) being analyzed.

4 Results

In this study, we were interested in understanding how biologically relevant, complex sounds are encoded in the IC. Mice are a model system for studying this because they possess a typical mammalian auditory system and they emit a wide repertoire of vocalizations under various social contexts [Panksepp et al., 2007, Wang et al., 2008]. Despite the ubiquity and behavioral relevance of mouse ultrasonic (>20 kHz) vocalizations, we found an under-representation of neurons with CFs that matched the spectral content of the vocalization stimuli (25 kHz to 102 kHz). Moreover, of the 111 neurons included in this report, 54% (n=60) responded to one or more of the natural and/or modified variants even though the majority (n=45) had a CF below 25 kHz (Fig. 4.1). Fourteen neurons that responded to the vocalizations did not respond to pure tones above 25 kHz.

The natural 30 kHz harmonic (n=28 responding neurons) and 40 kHz harmonic (n=9) vocalizations elicited responses in neurons across tonotopic layers, whereas the CFs of the neurons that responded to the natural female upswing (n=8) were clustered around 15 kHz in spite of the high spectral content (82-98 kHz) of this vocalization. The median response consistency of neurons responding to the natural female upswing ($\overline{R}_X^N = 0.56 \pm 0.21$) was higher than that of neurons responding to the natural 30 kHz harmonic and 40 kHz harmonic vocalizations ($\overline{R}_X^N = 0.27 \pm 0.04$ and 0.28 ± 0.11) for a similar median spike rate ($\bar{r} = 15.0 \pm 5.1$, 18.7 ± 5.5 , and 14.0 ± 3.6 spikes/s, respectively). This indicates that the discharge patterns of neurons responding to the natural female upswing vocalization showed less temporal variability than those responding to either natural harmonic vocalization, although the short duration of this vocalization contributes to this effect. The natural male upswing vocalization

Figure 4.1: Distribution of characteristic frequencies. Only 10% ($n=12$) of the neurons, had characteristic frequencies in the range of the ultrasonic vocalizations presented in this study, yet 54% ($n=60$) of the neurons responded to one or more vocalizations (natural or modified).

failed to elicit substantial responses in our sample of IC neurons, even though both the spectral content and duration are similar to that of the female upswing (Fig. 3.1C and D).

The different natural vocalizations evoked a heterogeneous pattern of responses in our neural population with respect to discharge rate and temporal firing pattern. In addition, the natural 30 kHz harmonic, 40 kHz harmonic, and female upswing vocalization elicited responses in different subsets of neurons. There was moderate overlap between the subsets; of the 31 neurons that responded to the natural vocalizations, 12 responded to more than one. Neurons with similar CFs and/or FRAs did not always respond to the same natural vocalizations (Fig. 4.2, FRAs not shown). This is an indication of population heterogeneity, and is in contrast to auditory nuclei like the DNLL [Bauer et al., 2002, Pollak et al., 2003]. Besides diversity in the degree of selectivity for the natural vocalizations across our population, we found response heterogeneity within individual neurons that responded to multiple vocalizations. Some neurons encoded the various vocalizations with different discharge rates, some with different discharge patterns, and some displayed differences in both. An example of a neuron that responded to multiple vocalizations generating diverse discharge rates and patterns is shown in the bottom row of Figure 4.2.

4.1 Predicting responses to complex stimuli from responses to pure tones

Spectro-temporal receptive field (STRF) models were created for all neurons in the experiment using each neuron’s pure tone responses. The normalized mean squared error (NMSE) on the training (pure tone) data was 0.43 with a standard deviation of 0.16. NMSE is normalized by the variance of the target data (the pure tone responses) such that any value below 1 is an improvement on a predicted response equal to each neuron’s mean firing rate. In most cases, the resulting model provided a good fit

Figure 4.2: PSTHs of six representative neurons in response to the natural vocalizations. The spectrograms of the vocalizations are displayed at the top. The characteristic frequency of each neuron is indicated on the left of each row of responses. Each of the vocalizations elicited a response in a subset of the neurons, and the responses were heterogeneous with respect to discharge rates and patterns. Even neurons with similar CFs responded differently to the vocalizations.

between the pure tone stimulus and responses across the full range of frequencies and intensities.

The models were then used to predict the responses to novel vocalization stimuli. These predictions approximated the responses of neurons in lower auditory nuclei with the same FRAs. The NMSE on these predictions was 0.92 with a standard deviation of 0.3. Even though the NMSE was higher on average than that of the training data (as would be expected from generalizing the model to novel complex stimuli), the predictions were often qualitatively similar to the actual responses to the harmonic vocalizations. The models could capture the delay and magnitude of onset responses, adaptation, inhibitory rebounds, and sideband inhibition. This is due to the overlap between the spectral content of these vocalizations and the FRAs of many of the neurons in the study. In some cases, however, the predictions deviated significantly from the recorded responses, even when the model fit the pure tone responses accurately. Furthermore, the models incorrectly predicted no response to the female upsweep vocalization variants, except for a few cases where the frequency downshifted variant did overlap with the FRA of neurons. This provides evidence that neural responses in the IC can be predicted by pure tone responses in some cases, but must be dependent on other mechanisms at other times.

Figure 4.3 provides an example of the recorded and modeled responses of a neuron to the natural variants of the 30 kHz harmonic, 40 kHz harmonic, and female upsweep vocalizations. The FRA of the neuron ranged between 7-35 kHz (Fig. 4.3A). Using the pure tone responses of this neuron, a model was generated that was the best linear transfer function between the spectro-temporal characteristics of the pure tone training data and the neural responses (Fig. 4.3B). The vocalization stimuli were converted into a spectrographic representation and convolved with the model to generate the predicted responses. The prediction to the 30 kHz harmonic vocalization closely

matched the recorded response (Fig. 4.3C). This close fit suggests that the encoding of this vocalization could be approximated by a linear summation of independent responses to different frequency components of the vocalization. The lack of overlap between the FRA of the neuron and the spectral content of the 40 kHz harmonic vocalization resulted in negligible recorded and predicted responses (Fig. 4.3D). The model cannot explain, however, the pronounced response to the female upswEEP vocalization (Fig. 4.3E). This suggests that a different encoding strategy was employed that is not based on linear or non-linear summations of the frequency components of the vocalization.

4.2 Changing the spectral content of vocalizations altered neural responses

It was common that neurons that were unresponsive to the natural vocalizations showed a response to at least one of their modified variants. This was the case for 24, 22, 5, and 7 neurons for the 30 kHz harmonic, 40 kHz harmonic, female upswEEP, and male upswEEP vocalizations, respectively. For both harmonic vocalizations, the increase in the number of responsive neurons could sometimes be explained by changes that occurred in the spectral content of these vocalizations after perturbing their acoustic parameters.

The effect of altering a vocalization’s spectral content on neural responsiveness is illustrated in Figure 4.4. In these examples, there was little (Fig. 4.4A) or no overlap (Fig. 4.4B, C) between the FRAs of the neurons and the spectral content of the natural 30 kHz harmonic vocalization. This is reflected in the weak (Fig. 4.4D) or absent (Fig. 4.4F, H) response to this vocalization. Removal of the AM makes the intensity of each harmonic equal to the mean intensity of the harmonic throughout the entire duration of the vocalization. This resulted in increased energy in the

Figure 4.3: Example of pure tone modeling of a neuron from this study. The output of the model approximates the expected response to arbitrary stimuli assuming that the pure tone responses can be used to characterize the response properties of the neuron. (A) The FRA of the neuron, showing a tuning range between 7-35 kHz. (B) A visualization of the parameterization of the model, which can be interpreted as a spectro-temporal receptive field (STRF). The vocalization stimuli were converted into a spectrographic representation and convolved with the model to generate the predicted response. (C) The actual and predicted responses to the 30 kHz harmonic vocalization. The model predicted a response that was an accurate representation of the recorded response. (D) The actual and predicted responses to the 40 kHz harmonic vocalization. In this case, there was no overlap between the frequency of the vocalization and the FRA of the neuron and the model accurately predicted the absence of a response. (E) The actual and predicted responses to the female upswEEP vocalization. Because the frequency content of the vocalization does not overlap with the FRA of the neuron, the model again predicted the absence of a response. In this case, the neuron unexpectedly had a strong response to the vocalization that cannot be explained by its pure tone responses.

lower frequency tail of the 30 kHz harmonic vocalization, shifting it into the FRA of the neuron shown in Figure 4.4A. As a result, this neuron drastically increased its discharge rate ($D = -0.36 \pm 0.14$) and thereby also altered its temporal response pattern ($S_{XY} = 0.22 \pm 0.08$, Fig. 4.4E). This response was accurately modeled by the pure tone model. Similarly, downshifting the 30 kHz harmonic vocalization in frequency increased the overlap between its spectral content and the FRA of the neuron in Figure 4.4B, causing a response (Fig. 4.4G). The transition from unresponsive to the natural vocalization ($R_X^N = 0.00 \pm 0.01$) to highly responsive to a variant with increased energy in lower frequencies is again reflected in the higher response consistency ($R_X^M = 0.29 \pm 0.08$), low response similarity ($S_{XY} = 0.00 \pm 0.02$), and negative selectivity index ($D = -0.62 \pm 0.16$). Once more, this response was accurately predicted by the pure tone model. In contrast to the vocalization variants shown in Figure 4.4E,G, the spectral content of the inharmonic variant in Figure 4.4I has shifted away from this neuron’s FRA (Fig. 4.4C), yet generates a consistent response ($R_X^M = 0.38 \pm 0.08$) whereas the natural vocalization does not ($R_X^N = 0.06 \pm 0.02$, $D = -0.36 \pm 0.10$). Furthermore, the pure tone model predicts the absence of a response to this variant (Fig. 4.4I). This apparent discrepancy may be due to the nonlinear amplification of sounds with multiple frequency components in the cochlea. The distortion product with the highest intensity generated through this mechanism (f2-f1) falls within the FRA of this neuron, perhaps causing its response. Neural sensitivity to cochlear distortion products has previously been documented throughout the auditory system [Goldstein and Kiang, 1968, McAlpine, 2004, Abel and Kössl, 2009, Portfors et al., 2009] and has been proposed as a mechanism for providing sensory cues in communication [Warren et al., 2009, Portfors et al., 2009].

Figure 4.4: Altering the spectral content of the vocalizations explained the change in responsiveness in some neurons. (A-C) The FRAs of 3 different neurons had little or no overlap with the natural 30 kHz harmonic vocalization. (D, F, and H) Low response consistencies to the natural vocalization are indicated by the low values for R_X^N . (E, G, and I) Each of the neurons responded reliably to a modified version of this vocalization, as indicated by the high values of R_X^M and the negative values of the selectivity index D . The responses in E and G resulted from the introduction of power in the FRA of the corresponding neuron when the vocalization was modified, as indicated by the close fit of the pure tone model predictions to the actual neural response. The FRA of the neuron in I does not contribute to the robust response, as attested the by lack of response predicted by the model. Nonlinear cochlear amplification of the different spectral components could result in the production of a difference tone, explaining the response.

4.3 Removal of AM or FM from the vocalizations altered neural responses

Of the neurons that responded to a natural vocalization or a no AM variant, 80% changed their discharge rate, discharge pattern, or both when the AM was removed. The discharge rate decreased in 31% and increased in 40% of the responses (Table 4.1). Likewise, 81% of neurons that responded to a natural vocalization or a no FM variant changed their discharge rate, discharge pattern, or both when the FM was removed. The discharge rate decreased in 34% and increased in 32% of the responses. This indicates that both AM and FM are important acoustic features in creating selectivity to natural vocalizations within individual IC neurons. Figure 4.5 displays two such cases.

The first neuron’s FRA (Fig. 4.5A) overlapped with the spectral content of the 40 kHz harmonic vocalization, resulting in a consistent response ($R_X^N = 0.40 \pm 0.13$) that was predicted by the pure tone model (Fig. 4.5C). After removal of the AM, the recorded response disappeared and the modeled response was significantly reduced (Fig. 4.5D, $R_X^M = 0.01 \pm 0.03$). Response selectivity for the natural vocalization translated into a positive selectivity index ($D = 0.37 \pm 0.10$) and low response similarity ($S_{XY} = 0.02 \pm 0.02$). The no AM variant of this vocalization had more energy in the 30-52 kHz band at onset, and triggered sideband inhibition that was captured by the model. The second neuron showed selectivity to the natural variant of the 30 kHz harmonic vocalization even though its FRA did not overlap with the frequency content of the vocalization (Fig. 4.5B). The pure tone model did not predict the strong response to the natural variant (Fig. 4.5E). Removal of the FM diminished the response (R_X decreased from 0.26 ± 0.08 to 0.07 ± 0.02), generating a positive selectivity index ($D = 0.23 \pm 0.10$) and low response similarity ($S_{XY} = 0.18 \pm 0.06$). The prediction

Figure 4.5: AM and FM are important features for creating selective neural responses. (A and B) The FRAs of 2 neurons. (C) This neuron consistently responded to the natural 40 kHz harmonic vocalization ($R_X^N = 0.40$), which was predicted by the pure tone model. (D) The lack of response to the no AM variant is indicated by the reduced R_X^M value, low response similarity ($S_{XY} = 0.02$) and large, positive selectivity index ($D = 0.37$). The predicted response was also attenuated, due to increased power introduced into the inhibitory sideband captured by the model. (E) This neuron showed a robust response to the natural 30 kHz vocalization, which was not captured by the model because the spectral power of the vocalization lies outside the FRA of the neuron. (F) The response was significantly reduced after removal of the FM from this vocalization.

of a weak response to the no FM variant remains relatively unchanged (Fig. 4.5F) because the spectral content of the variant is above the FRA of this neuron. These results suggest that the FM in the vocalization is responsible for evoking the neural response even though the frequency content of the vocalization is outside the neuron’s FRA. This neuron may be responding to cochlear distortions generated by the FM sweep as we have previously shown to occur in IC neurons [Portfors et al., 2009].

4.4 Vocalization duration modulated neural responses

As with removal of AM or FM, modifying the duration of the vocalizations changed the responses in the majority of the recorded neurons. For the most part, neurons decreased their discharge rate in response to the compressed vocalizations and increased their discharge rate in response to the stretched vocalizations (Table 4.1). To illustrate the response heterogeneity across our neural population, several ways that neurons modulated their response to duration-altered variants of the 40 kHz harmonic vocalization are shown in Figure 4.6. The neuron in Figure 4.6A was selective to the time stretched variant. The response consistency R_X of this neuron for the compressed, natural, and stretched vocalization were 0.02 ± 0.01 , 0.08 ± 0.02 , and 0.54 ± 0.10 , respectively. This indicates that the response pattern to the stretched version had the least temporal variability upon repeated presentations of the stimulus. Furthermore, the selectivity index ($D = -0.63 \pm 0.18$) signifies an increased discharge rate in response to the stretched vocalization. The strong response to the stretched variant was not predicted by the pure tone model, indicating that a mechanism other than facilitory frequency interaction may be contributing to the robust response.

The high response consistencies ($R_X = 0.55 \pm 0.14 - 0.60 \pm 0.12$) together with the selectivity indices close to zero ($D = -0.015 \pm 0.07 - 0.03 \pm 0.01$) of the neurons shown in Figure 4.6B, C suggest that these neurons do not prefer one duration-altered

Figure 4.6: Vocalization duration modulates the responses of neurons in the IC. (A, B, and C) Responses to the time compressed (half as long, left column), natural (middle column), and time stretched (twice as long, right column) 40 kHz harmonic vocalization are shown for 3 neurons. The neuron in A was selective for the stretched variant ($R_X^M = 0.54$ and $D = -0.63$), even though the pure tone model predicts no response to this variant. The neurons in B and C displayed consistent responses regardless of the duration of the stimulus (R_X ranges from 0.55-0.60), resulting in low values of D . However, both these neurons shifted the timing of the response in accordance with the duration of the vocalization ($S_{XY} \leq 0.25$). Note that the neuron in C displayed a sustained inhibitory response and that the inhibitory rebound is most apparent in response to the natural vocalization.

variant over another. When the response similarity is taken into account however, it becomes clear that the temporal response patterns changed from one variant to the next in both neurons ($S_{XY} \leq 0.25$). This encoding scheme would be particularly useful for discriminating between vocalization variants with different durations.

4.5 Population responses

By characterizing responses to natural vocalizations and perturbed variants of these vocalizations, we have shown that individual neurons are sensitive to particular acoustic parameters of the vocalizations. In some cases, this sensitivity can be explained by the pure tone responses of the neurons, while in other cases it cannot. Here, we address whether there were trends in the sensitivity across the population of responding neurons which may indicate population selectivity for one variant over the other. This population analysis was performed for each vocalization independently by focusing on the subset of neurons that responded to the vocalization or one of its variants. Of the 60 neurons that responded to our suite of vocalization stimuli (including all natural vocalizations and their variants), 52 responded to the 30 kHz harmonic, 31 to the 40 kHz harmonic, 13 to the female upsweep, and 7 to the male upsweep vocalizations.

For the two harmonic vocalizations, we found a heterogeneous mixture of responses to the natural and modified vocalizations (Fig. 4.7). The evoked response consistencies for each vocalization variant were distributed broadly with many neurons having low response consistencies and fewer having high values (Fig. 4.7A,C). The evoked selectivity indices for each vocalization variant were also distributed broadly (Fig. 4.7B,D), and this metric was approximately distributed normally. Individual neurons often varied significantly in their response consistencies and selectivity indices across the set of variants, as exemplified by the black line in Figure 4.7 following a

single neuron’s values.

The median response consistency was highest to the no AM variant and the distribution of these response consistencies (R_X^M) was statistically larger than the distribution of response consistencies (R_X^N) to the natural vocalizations (two-sample Kolmogorov-Smirnov test, $\alpha = 0.05$). All other variants evoked distributions of response consistencies that were not statistically different than the responses to the natural vocalization. Because the overlap between the spectral content of the no AM variants and the neurons FRAs is larger than for any other variant, almost half of the neural population increased their discharge rate in response to these vocalizations (Table 4.1). This led to a statistically significant increase in discharge rate ($\bar{D} = -0.09 \pm 0.10$, two-tailed, two-sample t-test, $\alpha = 0.05$) for the 30 kHz harmonic vocalization variant without AM. Similarly, downshifting the 40 kHz harmonic vocalization increased the overlap between spectral content and FRAs, which resulted in a statistically significant increase in discharge rate ($\bar{D} = -0.11 \pm 0.09$, two-tailed, two-sample t-test, $\alpha = 0.05$).

Disrupting the harmonic structure resulted in an overall decrease in discharge rate. This decrease was significant (two-sample Kolmogorov-Smirnov test, $\alpha = 0.05$) for the variant with the fundamental removed for the 30 kHz harmonic vocalization ($\bar{D} = 0.09 \pm 0.05$) and the variant with the harmonic removed for the 40 kHz harmonic vocalization ($\bar{D} = 0.10 \pm 0.08$). For the no fundamental variant of the 30 kHz vocalization, this is probably largely due to a decrease in spectral energy within the neurons’ FRAs. However, removal of the harmonic from the 40 kHz vocalization does not affect overlap between FRAs and spectral content, indicating that the harmonic structure is important for generating a response.

These results indicate no selectivity for the natural variant of each vocalization. On the contrary, they indicate that nearly all perturbations of the harmonic vocal-

Figure 4.7: (A and C) Population response consistencies (R_X) for neurons responding to at least one variant of the 30 kHz (A) or 40 kHz (C) harmonic vocalization, sorted from most consistent (top) to least consistent (bottom). (B and D) Population selectivity indices (D), sorted from least selective to the natural vocalization (top) to most selective (bottom). Box plots indicate the range of the second and third quartile of the distribution along with the median. The black line connects the metric values for a single neuron across the vocalization variants, indicating how variable a single neuron may be to perturbations in the vocalization characteristics. There is evidence of significant heterogeneity in the responses to the vocalizations from both a spike timing (A and C) and a spike rate (B and D) perspective. Except for the no AM variant, mean response consistencies did not differ significantly from the responses to the natural vocalizations (similar temporal jitter), nor was there a significant difference in mean spike rate between the responses to the natural vocalizations and most modified vocalizations. The (*) symbol indicates distributions that were significantly different than the natural variant's distribution (two-sample Kolmogorov-Smirnov test, $\alpha = 0.05$)

Table 4.1: Modifying acoustic parameters of vocalizations altered neural responses in a heterogeneous manner. Number of neurons that altered their spike rate or timing to each modification of each natural vocalization. Only neurons responding to one or more variant of each vocalization are included. $\uparrow$ indicates that the number of spikes increased by more than 20%. $\downarrow$ indicates that the number of spikes decreased by more than 20%. Δ indicates that the temporal pattern changed significantly, as determined by the classification rule defined in (3.4). $\sim$ indicates that no significant changes occurred in either spike rate or temporal pattern.

	$\uparrow$	$\uparrow/\Delta$	$\downarrow$	$\downarrow/\Delta$	Δ	$\sim$
30 kHz Harmonic						
AM Removed	9	11	9	2	2	13
FM Removed	8	6	13	2	2	15
Time Compressed	9	3	13	4	3	18
Time Stretched	10	10	9	1	3	12
Harmonic Removed	4	5	5	6	2	15
Fundamental Removed	5	3	9	6	1	14
40 kHz Harmonic						
AM Removed	6	7	7	1	1	6
FM Removed	10	2	7	1	1	7
Time Compressed	5	2	7	3	2	11
Time Stretched	4	5	10	1	2	6
Harmonic Removed	4	0	5	4	0	9
Fundamental Removed	5	0	4	2	2	9
Female Upsweep						
AM Removed	2	0	0	8	0	3
FM Removed	2	0	2	6	0	3
Time Compressed	3	0	4	3	0	3
Time Stretched	4	2	1	0	1	5

izations resulted in similarly robust responses from both a spike timing and spike rate perspective. Each variant, however, was encoded very differently, with different subsets of neurons responding to each variant or significantly different responses by individual neurons responding to multiple variants.

In contrast to responses to the harmonic vocalizations, responses to the female upsweep displayed strong selectivity to the natural vocalization (Fig. 4.8). With the exception of the time stretched variant, the median response consistency and mean discharge rates were highest to the natural variant ($\overline{R}_X^N > \overline{R}_X^M$ and $\overline{D} > 0$). This suggests that the female upsweep is encoded by neurons that are very sensitive to the spectro-temporal properties of this vocalization and that similar stimuli may be filtered out precortically.

This result is particularly striking when considering how similar the male and female upsweep vocalizations are with respect to frequency and duration (Fig. 3.1C,D). The only noteworthy differences between these two vocalizations are AM and FM, suggesting that the subpopulation of neurons that respond to the female upsweep are highly AM or FM selective. However, the time stretched variant of the male upsweep is very similar to the natural female upsweep (with respect to FM sweep rate, range and bandwidth), yet does not elicit responses in neurons that do respond to the female upsweep vocalization. This finding suggests that selectivity to AM or AM rate is the underlying reason for the differences in responses to the two vocalizations. Again, the spectral content of the upsweep vocalization is outside the FRA of the neurons. One possible explanation for these responses is that AM can create cochlear distortion products that evoke responses in IC neurons even when the carrier-frequency of the signal is outside the neural FRA [McAlpine, 2004]. Consequently, when AM is removed, the cochlear distortions are not generated and no response is elicited.

The differences in selectivity to the female upsweep vocalization (Fig. 4.9A) com-

Figure 4.8: (A) Population response consistencies (R_X) for neurons responding to at least one variant of the female upsweep vocalization, sorted from most consistent (top) to least consistent (bottom). (B) Population selectivity indices (D), sorted from least selective to the natural vocalization (top) to most selective (bottom). Box plots indicate the range of the second and third quartile of the distribution along with the median. The black line connects the metric values for a single neuron across the vocalization variants. There is significant evidence of population selectivity to the natural vocalizations from both a spike timing (A) and a spike rate (B) perspective. With the exception of the stretched variant, response consistency values differed significantly from the responses to the natural vocalizations (more temporal jitter), and there were significant differences in mean spike rate between the responses to the natural vocalizations and the modified vocalizations. (*) Distributions that were significantly different than the natural variant's distribution (two-sample Kolmogorov-Smirnov test, $\alpha = 0.05$)

pared to the male upstroke vocalization (Fig. 4.9B) are displayed for seven representative neurons. As can be seen in the first column of Figure 4.9A, the responses to the natural female upstroke vocalization showed very little variability in the presence and timing of spikes across stimulus presentations. This results in a high $\overline{R}_X^N$. Responses to the time stretched variant (Fig. 4.9A, last column) were similar in temporal pattern but were often slightly higher in discharge rate. This accounts for the negative but statistically insignificant value of $\overline{D}$ for this variant (Fig. 4.8). Changes in AM, FM, and frequency shifting evoked little or no response from the neurons, resulting in low values of $\overline{R}_X^M$ and high values of $\overline{D}$. Figure 4.9B shows the responses of the same neurons to the male upstroke vocalization. At best, the neurons responded weakly to the modified vocalizations, resulting in the low values of $\overline{R}_X^N$, $\overline{R}_X^M$, and $\overline{D}$ (data not shown).

4.6 Efficiency of vocalization discrimination in the IC

A primary motivation behind this study was to objectively test whether the IC encodes information more efficiently than lower auditory nuclei by increasing the heterogeneity of the neural responses to complex vocalization stimuli. This hypothesis is based on evidence that the IC is the first site in the ascending auditory system that shows selectivity to complex features found in vocalizations [Klug et al., 2002] and that increasing the heterogeneity and variability of neural responses results in a theoretical increase in information throughput [Linsker, 1988, Gawne and Richmond, 1993, Bell and Sejnowski, 1995, Shamir and Sompolinsky, 2006, Chelaru and Dragoi, 2008].

For each neuron in the study, the information transfer H was calculated to quantify how efficiently the neuron could discriminate among the variants of each natural vocalization (e.g., 30 kHz harmonic). If a particular neuron had weak or correlated

Figure 4.9: Specific examples of responses to variants of the female (A) and male (B) upsweep vocalizations showing that the neural population was selective to the natural (A, first column) and time stretched (A, last column) variants of the female upsweep vocalization. The spectrograms of the vocalizations are displayed across the top in the following order: natural, +5%, -5%, -20%, no AM, no FM, time compressed, time stretched. The characteristic frequency of each neuron is displayed on the right. Neurons with a wide range of CFs responded to the female upsweep vocalization, even though its spectral content is far outside the FRA of each neuron and the pure tone models predicted no response. Despite the similarity in duration and spectral content to the female upsweep vocalization, the natural male upsweep vocalization did not elicit substantial responses in the neuron population, and very few neurons responded to any of its variants.

responses to each variant of a vocalization, then it was difficult to classify what vocalization was responsible for a particular response and H was low for the neuron. If the responses were highly heterogeneous, then classification was possible and H was increased.

It has been shown that the pure tone response characteristics of neurons from lower auditory nuclei are sufficient to estimate their responses to complex vocalization stimuli [Bauer et al., 2002, Pollak et al., 2003]. In order to approximate how the neurons in our study of the IC would respond if they had been in a lower auditory nucleus, we used a pure tone spectro-temporal modeling methodology (see Methods) to predict their responses to our vocalization stimuli. The output of the model was an estimated PSTH with a bin width of 2 ms (chosen to approximate the neuron’s refractory period). We were then able to generate simulated spike trains for each neuron and vocalization by interpreting these modeled PSTHs as the probability of finding a spike within each 2 ms window. Using this method, we generated a set of simulated spike trains of equal number to the recorded data (20 spike trains per stimulus) and performed an identical assessment of the information transfer, H , of each neuron for each vocalization class. This enabled us to directly compare the discrimination efficiency of neurons in the IC with the discrimination efficiency of neurons from lower auditory nuclei. Figure 4.10 summarizes the results of this analysis.

For the 30 kHz harmonic vocalization (Fig. 4.10A,B,C), the 52 neurons that responded to at least one variant of this vocalization were included in the analysis. H , and $\overline{H_{bias}} \pm 2SD$ were calculated for both the recorded data (Fig. 4.10A) and the modeled data (Fig. 4.10B). The maximum information transfer possible was $\log_2 N_{vocs} = 3.7$ bits, which is attainable if each spike train could unambiguously be matched with the associated vocalization using the similarity measure S_{XY} . For the

Figure 4.10: IC neurons employ a more efficient encoding strategy for discrimination among vocalizations than an encoding strategy based on modeled responses by lower auditory nuclei. (A) Information transfer (H) between stimuli and responses for each neuron that responded to at least one of the 30 kHz harmonic vocalization variants. The empty circle and error bars indicate $\overline{H_{bias}} \pm 2SD$. The x-axis spans the bounds of H ($[0 \log_2 N_{vocs}]$ bits). The dashed line indicates H_{pop} , the information transfer of the neural population as a whole. (B) H for each neuron measured from the modeled responses to the 30 kHz harmonic vocalization variants. (C) Histogram comparing the recorded and modeled distributions of H for the neurons with significant information transfer ($H > \overline{H_{bias}} + 2SD$). H for the recorded data is significantly greater than H for the modeled data (two-sample Kolmogorov-Smirnov test, $p < 0.005$). (D) H between stimuli and response for each neuron which responded to at least one of the female upsweep vocalization variants. (E) H for each neuron measured from the modeled responses to the female upsweep vocalization variants. (F) Histogram comparing the recorded and modeled distributions of H for the neurons with significant information transfer ($H > \overline{H_{bias}} + 2SD$). H for the recorded data is significantly greater than H for the modeled data (two-sample Kolmogorov-Smirnov test, $p < 0.05$).

recorded data, 41 of the 52 neurons had significant values of H ($H > \overline{H_{bias}} + 2SD$). The maximum information transfer H_{max} was 2.01 bits and the mean information transfer $\overline{H}$ was 0.83 bits. For the modeled data, 34 of the 52 neurons had significant values of H , $H_{max} = 1.31$ bits, and $\overline{H} = 0.57$ bits. The distributions of H for the recorded and modeled data (Fig. 4.10C) were significantly different (two-sample Kolmogorov-Smirnov test, $p < 0.005$). When taken a whole, as if they fired in synchrony to each vocalization stimuli, the information transfer of the neural population, H_{pop} , was 3.62 bits for the recorded data and 2.91 bits for the modeled data. This result shows that the set of 52 neurons could almost perfectly classify each vocalization (2.6% error rate). This result also shows that the increased heterogeneity of the recorded responses led to a more efficient population encoding than the less heterogeneous modeled responses.

For the female upswing vocalization (Fig. 4.10D,E,F), the 13 neurons that responded to at least one variant were included in the analysis. H , and $\overline{H_{bias}} \pm 2SD$ were calculated for both the recorded data (Fig. 4.10E) and the modeled data (Fig. 4.10F). The maximum information transfer possible was $\log_2 N_{vocs} = 3$ bits. There were fewer vocalizations in this analysis because frequency upshifted variants were not included and the female upswing is not harmonically structured. For the recorded data, 11 of the 13 neurons had significant values of H ($H > \overline{H_{bias}} + 2SD$). The maximum information transfer H_{max} was 0.92 bits and the mean information transfer $\overline{H}$ was 0.51 bits. For the modeled data, 7 of the 13 neurons had significant values of H , $H_{max} = 0.28$ bits, and $\overline{H} = 0.22$ bits. The distributions of H for the recorded and modeled data (Fig. 4.10F) were significantly different (two-sample Kolmogorov-Smirnov test, $p < 0.05$). For this vocalization, H_{pop} was 1.40 bits for the recorded data and 0.37 bits for the modeled data. This again shows that the increased heterogeneity of the recorded responses led to a more efficient population encoding than

the less heterogeneous modeled responses. Because these responses were not as heterogeneous as the responses to the 30 kHz harmonic vocalization, however, it was not possible to uniquely determine the source vocalization using our recorded neural population and classification methodology.

Both of these cases indicate that individual neurons encoded the 30 kHz harmonic and female upswing vocalizations such that they provided useful information for discriminating among the variants of these vocalizations. Even though the modeled responses to the 30 kHz harmonic vocalization variants were often qualitatively similar to the recorded responses, significantly more information was available within the recorded responses for the task of discriminating among the vocalization variants. Furthermore, the analysis based on the recorded responses of the neural population resulted in more neurons with significant values of H than did the analysis based on the responses generated by each neuron's pure tone model ($n=41$ vs. $n=34$ for the 30 kHz harmonic vocalization and $n=11$ vs. $n=7$ for the female upswing vocalization), meaning that a higher percentage of the neural population contributed to the encoding of this vocalization. This suggests that the actual responses were further modulated by the system to increase their heterogeneity and information throughput. For the female upswing vocalization, the differences between the recorded and modeled H populations were even more significant because the models almost always predicted the absence of any response. Classification between the natural, time compressed, and time expanded variants was difficult due to the correlation of these responses (Fig. 4.9A), however. In this case, most of the information transfer can be attributed to the ease of classifying between the set of variants that elicited a response and the set of those that did not.

4.7 Heterogeneity of responses across the neural population

While the previous efficiency measure was focused on the heterogeneity of responses from a single neuron to a suite of vocalization stimuli, another efficiency measure is based on the heterogeneity of responses across the neural population to individual vocalizations. An increase in this type of heterogeneity has been shown theoretically to increase information throughput of neural populations by reducing within-trial noise correlations [Shamir and Sompolinsky, 2006, Chelaru and Dragoi, 2008] and inter-trial signal correlations [Gawne and Richmond, 1993, Bell and Sejnowski, 1995, Chechik et al., 2006].

For each vocalization variant in the study, the information transfer H was calculated to quantify how heterogeneous the responses were across the whole neural population. If each neuron in the study had weak or correlated responses to a particular vocalization variant, then it would be difficult to classify which neuron a given response could be attributed to and H would be low for the vocalization variant. If the responses were highly heterogeneous, then classification was possible and H was high. Figure 4.11 summarizes the results of this analysis.

For the 30 kHz harmonic vocalization (Fig. 4.11A,B,C), the 52 neurons that responded to at least one variant were included in the analysis. H , and $\overline{H_{bias}} \pm 2SD$ were calculated for both the recorded data (Fig. 4.11A) and the modeled data (Fig. 4.11B). The maximum information transfer possible was $\log_2 N_{neurons} = 5.2$ bits, which occurs if each spike train could unambiguously be matched with the associated neuron using the similarity measure S_{XY} defined in (3.4). For both the recorded and modeled data, all 13 vocalization variants had significant values of H ($H > \overline{H_{bias}} + 2SD$). The recorded data had a maximum information transfer H_{max} of 1.98 bits and a mean information transfer $\overline{H}$ of 1.46 bits. For the modeled data, $H_{max} = 1.53$ bits

Figure 4.11: IC responses are more heterogeneous to a given vocalization than is predicted by modeled responses from lower auditory nuclei. **(A)** Information transfer (H) for each variant of the 30 kHz harmonic vocalization, measured across all neurons that responded to at least one of the variants. The empty circle and error bars indicate $\overline{H_{bias}} \pm 2SD$. The x-axis spans the bounds of H ($[0 \log_2 N_{neurons}]$ bits). **(B)** H for each variant of the 30 kHz harmonic vocalization using the responses modeled from the pure tone response properties of each neuron. **(C)** Histogram comparing the recorded and modeled distributions of H for the vocalization variants with significant information transfer ($H > \overline{H_{bias}} + 2SD$). H for the recorded data is significantly greater than H for the modeled data (two-sample Kolmogorov-Smirnov test, $p < 0.001$). **(D)** Information transfer (H) for each variant of the female upsweep vocalization, measured across all neurons that responded to at least one of the variants. The three variants with the elevated values of H are: natural (bottom), time stretched, and time compressed (top). **(E)** H for each variant of the female upsweep vocalization using the responses modeled from the pure tone response properties of each neuron. **(F)** Histogram comparing the recorded and modeled distributions of H for the vocalization variants with significant information transfer ($H > \overline{H_{bias}} + 2SD$). H for the recorded data and H for the modeled data are not significantly different. This is primarily due to the small sample size ($n = 8$).

and $\overline{H} = 1.17$ bits. The distributions of H for the recorded and modeled data were significantly different (two-sample Kolmogorov-Smirnov test, $p < 0.001$).

For the female upsweep vocalization (Fig. 4.11D,E,F), the 13 neurons that responded to at least one variant were included in the analysis. H , and $\overline{H}_{bias} \pm 2SD$ were calculated for both the recorded data (Fig. 4.11D) and the modeled data (Fig. 4.11E). The maximum information transfer possible was $\log_2 N_{neurons} = 3.5$ bits. For both the recorded and modeled data, all 8 vocalization variants had significant values of H ($H > \overline{H}_{bias} + 2SD$). The recorded data had a maximum information transfer H_{max} of 1.09 bits and a mean information transfer $\overline{H}$ of 0.54 bits. For the modeled data, $H_{max} = 0.39$ bits and $\overline{H} = 0.33$ bits. The distributions of H for the recorded and modeled data were not significantly different in this case (two-sample Kolmogorov-Smirnov test), even though the data indicates that H was greater for the recorded responses. The lack of statistical significance is primarily due to the small number of samples ($n = 8$).

Even though the modeled responses to the 30 kHz vocalization captured many of the salient features of the recorded responses, they were not as heterogeneous, resulting in lower values of H for every vocalization and a less efficient encoding. For the female upsweep vocalization, H was very low for the 5 vocalization variants that elicited insignificant responses in the population for both the recorded and modeled responses. Even though the recorded responses to the natural, time compressed, and time expanded variants were highly correlated across the neurons (Fig. 4.9A), it was still often possible to classify which neuron generated a particular response based on small differences in the temporal pattern of the response, resulting in higher values of H for these vocalizations.

Both of these cases again indicate that neurons in the IC have responses that are more heterogeneous to a given vocalization than would be expected in lower auditory

nuclei, resulting in a more efficient encoding.

5 Conclusions

5.1 State-Space Stimulus Design

Stimulus design in auditory neuroethology is dependent upon mathematical methods for analyzing and synthesizing behaviorally relevant acoustic signals. In particular, modified versions of conspecific vocalizations are increasingly being leveraged to test hypotheses about high level neural mechanisms responsible for the perception of animal and human vocalizations. Even though signal models have been used previously in auditory neuroethology for synthesizing vocalizations [DiMattina and Wang, 2006], one must first analyze recorded vocalizations to accurately fit these models.

The methods detailed in Appendices B-C represent a step forward in the field of auditory stimulus design by providing a unified state-space framework that can be used to both analyze and synthesize vocalizations. The performance measures detailed in B.4 and C.6 show that a state-space modeling framework used in conjunction with the Extended Kalman Smoother (EKS) or Fast Maximum a Priori Marginalized Particle Filter (Fast MAM-PF) is capable of accurately estimating the frequency, phase, and amplitudes of vocalizations – even in the presence of excessive measurement noise. Spectrographic frequency tracking can sometimes provide reasonable estimates of the frequency and amplitude, but an additional optimization step must be taken to fit the phase in order to accurately reproduce the original vocalization [Laroche et al., 1993]. Adopting a unified state-space approach for analysis and synthesis makes this additional and complex step unnecessary.

For the rigorously tested case of tracking the fundamental frequency of synthetic hummingbird vocalizations, the EKS approach was more accurate than the common Short-Time Fourier Transform (STFT) method used previously in this field (see

Appendix B.4). This result holds true for both noise-free and significantly noisy (0 dB SNR) signals using either white Gaussian or $1/f$ (pink) measurement noise models (see Figs. B.3 and B.4).

However, the accuracy of the estimates provided by the EKS are, in part, conditional upon the local linearization of the state measurement equation (B.7) at each time step. If the state estimates ever stray too far from the true values and this linearization is calculated about the wrong point in state-space, a loss of track may occur which can be difficult to recover from. This can be the result of either poorly specified tracking parameters or a vocalization that is particularly difficult to track due to excessive noise, recording artifacts, periods of silence, or abrupt changes in frequency. In these cases, it is possible for the algorithm to lock onto a super- or sub-harmonic of the signal and never regain track, resulting in poor state estimates for all of the subsequent harmonic model parameters.

Because many vocalizations contain abrupt frequency jumps and periods of silence, we also explored the use of particle filter methods for the task of analyzing these signals. Over the course of our exploration, we developed a new method that specifically addresses the computational and memory limitations of common approaches to sequential Monte-Carlo methods. Our method leverages a particle filter to estimate the non-linear portion of the state-space signal model (the frequency) and the Kalman Filter to estimate the linear portion (the phase and amplitudes) in what is called a marginalized particle filter. Furthermore, our novel method uses the Viterbi algorithm to provide the maximum a posteriori (MAP) estimate of the state at each time sample. This hybrid method, which we call the maximum a posteriori adaptive marginalized particle filter (MAM-PF), is discussed in detail in Appendix C.

The performance of the Fast MAM-PF methodology was assessed by generating an ensemble of simulated mouse vocalizations which included abrupt frequency jumps

(see Appendix C.6). While this experiment showed that the PF method was significantly better than the EKS at regaining track on the fundamental frequency of these signals after abrupt frequency jumps, the PF did not track the fundamental frequency as accurately as the EKS when the EKS was locked on.

One explanation for this shortcoming is the relatively small number of particles used in each simulation. PF methods generate asymptotically optimal state estimates as the number of particles goes to infinity, but they are computationally expensive. The Fast MAM-PF algorithm used in this experiment scales exponentially with the number of particles. In order to run enough simulations to achieve the performance assessment results reported in Appendix C.6, the number of particles for each simulation had to be limited to 1000.

A consequence of using the harmonic signal model described in A.2.2 is that the tuning process can sometimes be challenging. This model describes a signal with random, normally distributed deviations in frequency, phase, and amplitude. While often *possible*, this model is not always a *probable* description of many vocalizations, which typically have a more complex and non-stationary statistical structure. For vocalizations with significant statistical variability, there may not be a static set of model parameters that results in accurate tracking throughout the whole signal. For example, increasing the variance of the frequency noise may improve the ability of the algorithm to track a large jump in frequency, but it may also make the algorithm track sub-optimally on sections of the vocalization that have little variability in frequency.

Another concern about this state-space model is that it assumes that the analyzed signal is composed of harmonically related components and is not an appropriate model for vocalizations that include anharmonic components or broad band noise. The residual analysis of the EKS analysis/synthesis of the human vocalization described in B.5.4, however, showed that the broad band component of the vocaliza-

tion can be accurately analyzed and synthesized. This is not too surprising, because Fourier’s harmonic analysis states that any periodic function can be represented as a superposition of harmonics. Since, in this case, 40 harmonics were specified in the model, it was possible to model even the broad band noise sections of the vocalization. In this case, the human vocalization recording was understood to have noisy components that were part of the vocalization itself rather than a measurement artifact. To account for this, the measurement noise variance was set to be low in the tracking parameters, and the EKS algorithm integrated more of the random noise directly into the state estimates. Accurate analysis and synthesis of human vocalizations may enable this technique to be used in psychoacoustic studies as well.

Aside from variability within each signal, the ensemble of synthetic mouse vocalizations included significant variability between vocalizations. It was often possible to improve the tracking performance on individual vocalizations through careful selection of the tracking parameters. This type of optimization is time consuming, however, and was not possible for the 1500 simulations that were completed for the performance assessment. Regardless, both the EKS and Fast MAM-PF methods outperformed the current best practices in the field and represent a significant contribution to the auditory community’s toolchest.

In addition to the accurate tracking of the frequency, phase, and amplitudes of vocalizations, the state-space harmonic model used in this study offers extensive flexibility for generating modified versions of vocalizations by altering behaviorally relevant parameters of analyzed vocalizations. Since vocalizations are modeled as sums of harmonically related signals with independent amplitudes and phase offsets, it is trivial to modify any of these characteristics prior to the synthesis phase, resulting in a quantifiably altered vocalization. This approach is considerably less complex and more accurate than comparable methods for frequency and time base shifting

such as phase vocoding [Flanagan and Golden, 1966] – a popular technique which is limited by the time-frequency tradeoff of the STFT calculation in the analysis phase and requires complicated and/or heuristic algorithms to maintain phase coherence in the synthesized signal. In addition, while the harmonic structure of the vocalizations is leveraged by the state-space tracking methods during the analysis phase for accurate parameter tracking, anharmonic modifications of these vocalizations can be generated during the synthesis phase by relaxing the constraint that the signal components are harmonically related. The frequency, length, amplitude, and phase of each harmonic component can be modified independently resulting in a flexible yet compact vocalization signal model.

Event though there are known limitations, the methods outlined in this dissertation represent a significant advance forward in the ability to analyze and synthesize vocalizations from a variety of animals obtained under a variety of environmental conditions. Being able to synthesize, de-noise, and easily modify the acoustic features of natural vocalizations will enhance the ability to use vocalization stimuli in behavioral and neurophysiological experiments.

5.2 Heterogeneous response properties in the IC

In this study, we used the aforementioned stimulus design methodology to analyze a set of four mouse vocalizations and to generate a set of modified variants of these vocalizations. Measuring the neural responses to these variants revealed that IC neurons were both selective and sensitive to small perturbations in the spectro-temporal characteristics of these stimuli. Furthermore, we found that these perturbations often affected each neuron’s response in a different way – even for neurons with similar characteristic frequencies (CFs).

The 30 and 40 kHz harmonic vocalizations used in our study are 4-6 times longer

than the upswing vocalizations and have significant AM and FM variability. These vocalizations generated responses in large subsets of neurons: 47% (n=52) and 28% (n=31), respectively. The populations that responded to each variant had highly heterogeneous mixtures of spike rates and temporal response consistencies. While the distributions of spike rates and temporal consistencies were similar for the variants of each vocalization, the responses of individual neurons often varied significantly with each perturbation. These results show that the IC is highly sensitive to perturbations of acoustic features in these stimuli, resulting in a distinct neural representation of each vocalization.

Response heterogeneity has been reported in numerous cortical structures [Chelaru and Dragoi, 2008]. In primary visual cortex there is a broad diversity of neural selectivity for complex stimuli across all cortical layers [Hubel and Wiesel, 1962, Ringach et al., 2002]. Similarly, the somatosensory cortex has substantial variability of receptive field size, overlap, and position across layers and within layers [Haupt et al., 2004]. Our results suggest that heterogeneity in receptive fields and response types is already present in precortical auditory pathways.

In contrast to the harmonic vocalizations, the female and male upswing vocalizations used in our study are short in duration, have monotonic FM, no harmonic structure, and simple AM envelopes. We found significant differences between the encoding of these and the harmonic vocalizations. Only 13 and 7 neurons responded to variants of the female and male upswing vocalizations, respectively. We found strong evidence of neural selectivity for the natural and duration modified variants of the female vocalization. While the median number of spikes present in each neuron’s response to the natural variant of this vocalization was not significantly different than the harmonic vocalizations ($\bar{r} = 15.0 \pm 5.1$ spikes/s vs. 18.7 ± 5.5 and 14.0 ± 3.6 spikes/s for the 30 kHz and 40 kHz harmonic vocalizations), the timing precision was

significantly greater ($\overline{R_X^N} = 0.56 \pm 0.21$ vs. 0.27 ± 0.04 and 0.28 ± 0.11). All other variants of the vocalization elicited negligible responses. This suggests an encoding of this vocalization that filters out all similar, and possibly irrelevant, stimuli.

These results are in agreement with those reported by [Liu and Schreiner, 2007] in mouse auditory cortex, who found that the timing of spikes, and not just the average spike count, is an important aspect of the neural encoding of vocalizations. Furthermore, they observed selectivity for the natural versions of their ultrasonic vocalization stimuli, which were simple in structure and similar to the female upswEEP vocalization we used.

Selectivity for the natural variants of conspecific vocalizations has been reported in the auditory cortex [Gehr et al., 2000, Wang and Kadia, 2001, Grace et al., 2003, Wang et al., 2005, Amin et al., 2007] and the IC [Castellanos et al., 2007] of mammals and birds. Contradictory results have been reported, however, in both the auditory thalamus [Philibert et al., 2005, Huetz et al., 2009] and the IC [Suta et al., 2003] using anesthetized animals. This has led to speculation that selectivity may be species, state, and nuclei specific [Philibert et al., 2005]. Our finding of neural population selectivity only for the simple vocalizations indicates that these previous results are not necessarily contradictory.

Despite being very similar in frequency and duration, the natural variant of the male upswEEP vocalization did not generate substantial responses in the neural population. It is possible that our sample set of neurons did not include any neurons that responded to this vocalization. Alternately, it is possible that this vocalization is filtered out by this stage of processing, thereby simplifying the discrimination task for higher level processing.

Different neural mechanisms likely underlie the selective responses to vocalizations seen in the IC. In the simplest case, some neurons respond to all vocalizations that

have energy within the neuron’s frequency tuning curve. In other cases, neural inhibition and non-linear integration of spectro-temporal features may explain selectivity to different acoustic features found in vocalizations such as AM, FM, duration and spectral content [Andoni et al., 2007, Holmstrom et al., 2007b, Brand et al., 2000, Portfors and Wenstrup, 2002, Woolley and Casseday, 2005, Portfors and Felix, 2005, Pérez-González et al., 2006, Andoni et al., 2007]. Finally, IC neurons may utilize cochlear distortions that are generated by combinations of ultrasonic frequencies in vocalizations to generate responses even when the spectral content of the vocalization is outside the FRA of the neuron [Portfors et al., 2009].

Selectivity is not sufficient to generate heterogeneity, however. The principal cells in the Dorsal Cochlear Nucleus (DCN), for example, process auditory information non-linearly and have been shown to respond selectively to certain complex features in vocalizations. These cells all exhibit the same non-linearities, however, and cells with similar CFs respond in the same way [Pollak et al., 2003]. By independently manipulating the acoustic features of our mouse vocalizations, we showed that each neuron was sensitive to these perturbations and that these responses were modulated in a different ways – even for cells with similar CFs. Our findings support the argument that this type of heterogeneity originates in the IC, and results in heterogeneous representations of each vocalization.

5.3 Efficient encoding in the IC

Our finding that neurons in the IC have heterogeneous responses to vocalizations argues for the presence of efficient encoding in the mammalian midbrain.

Our information theoretic analysis compared the efficiency of the encoding of vocalization stimuli in the IC to an encoding that was predicted by the pure tone responses of each neuron. We used the pure tone response model as our reference

because previous studies have shown that responses in most nuclei below the IC (e.g, the auditory nerve and the dorsal nucleus of the lateral lemniscus) can be predicted by linear summations of neural responses to pure tones alone [Klug et al., 2002, Bauer et al., 2002, Pollak et al., 2003]. Based on the techniques of these previous studies, our model approximates the responses of neurons in peripheral nuclei that have identical pure tone responses to the IC neurons used in this study.

Our results showed that the mouse IC encodes vocalizations more efficiently than a peripheral encoding. Information transfer was significantly improved due to the increase in the heterogeneity of the responses across the neural population. Other studies have theoretically shown that increasing the heterogeneity of the receptive fields of individual neurons increases the information contained in the population response due to the decrease in neural correlation [Linsker, 1988, Gawne and Richmond, 1993, Bell and Sejnowski, 1995, Shamir and Sompolinsky, 2006, Chechik et al., 2006, Chelaru and Dragoi, 2008].

The encoding of the female vocalization, however, was efficient due to selectivity rather than heterogeneity. This vocalization, perhaps because of its simple spectro-temporal structure and lack of natural variability, was encoded by a small number of highly precise neurons. Most perturbations to this vocalization resulted in little or no response. Because the pure tone model for each neuron predicted no response from this vocalization, our information theoretic analysis showed that the encoding employed by the IC was more efficient than our reference encoding. The male vocalization, on the other hand, did not generate responses in any of our 111 neurons despite its similarity to the female vocalization, perhaps indicating that it was filtered out by this level of auditory processing because it was not often encountered or was behaviorally irrelevant to our subject mice.

Our findings suggest that the IC plays a critical role in the efficient encoding of au-

ditary information by facilitating the discrimination of behaviorally relevant sounds. Considering that mice emit a much more extensive repertoire of vocalizations than were used as stimuli in this study, we expect that the IC is even more heterogeneous in its responses to complex sounds in the natural environment.

5.4 Contributions

The research reported in this dissertation represents a number of methodological and scientific contributions to the field of auditory neurophysiology.

The use of state-space methodologies for analyzing and synthesizing vocalizations. Despite a significant focus on using vocalization based stimuli, the best practices in the field have been limited by the use of the STFT for vocalization analysis. Additionally, spectrographic representations only lend themselves to simple frequency and time scaling, which limits the possibilities for testing neural sensitivities to other perturbations of these signals. The state-space methodologies detailed in this dissertation overcome many of these limitations and offer new possibilities for neurophysiological experiments [Holmstrom et al., 2009].

Improvements to STRF modeling techniques. While STRFs have been discussed in the literature since 1981 [Aertsen and Johannesma, 1981], there is still no canonical STRF implementation. Instead, the STRF methodology encompasses a family of models which are defined by their spectro-temporal representation and the algorithm(s) used for determining the parameter values. Historically, STRF models were fit through reverse correlation techniques using broad band noise for input. This technique breaks down, however, when the stimulus is no longer uncorrelated, as is true for ripple noise, pure tones, and vocalization stimuli. A framework for handling correlated stimuli in the fitting of STRF models was first proposed in [Theunissen et al., 2000]. The method we developed is functionally equivalent but simpler to

describe and implement [Holmstrom et al., 2007b]. Furthermore, we introduced a new step in the STRF fitting process that is based on the addition of a nonlinearity in the optimality criteria and a gradient descent method to carry out the fitting process. We found that this addition significantly improved the prediction accuracy of our models to novel vocalization stimuli (see Fig. 3.3).

The characterization of neural responses from both a spike rate and spike timing perspective. Historically, studies of neural selectivity and sensitivity have used only the total number of spikes in the comparison of neural responses to different stimuli. The correlation based metrics used in this study embrace a different perspective in which the precise timing of spikes is used in this comparison. The research reported here represents the first time this type of similarity measure has been used in a study of the IC [Holmstrom et al., 2010].

An information theoretic measure of the efficiency of the encoding of vocalizations in the IC. By perturbing a set of vocalizations using our novel state-space methodologies, we were able to characterize the selectivity, sensitivity, and heterogeneity of responses in the IC to behaviorally relevant stimuli. This was the first detailed experiment of its kind, and our results provide strong evidence that the IC plays a large role in the efficient encoding of vocalizations through both selectivity and sensitivity [Holmstrom et al., 2010].

5.5 Future Directions

State-space stimulus design. While the state-space stimulus design methodologies detailed in this dissertation represent a significant leap forward for stimulus design in auditory neurophysiology, there is still considerable work to be done in order to create an efficient tool capable of analyzing numerous types of vocalizations with minimal user interaction. Currently, both the EKS and PF approaches require significant

tuning to achieve their goals. Furthermore, the computational requirements of the PF methods make it prohibitive to analyze lengthy records or to iterate over a large number of variables in order to tune the algorithm correctly. While the EKS approach is the more practical in its current state, the PF approach holds the most promise because it is theoretically capable of handling the abrupt jumps in frequency and moments of silence that are common in vocalizations. To handle these types of vocalizations with the EKS approach requires splitting vocalizations up into smaller sections which can each be analyzed independently. This is a time consuming process.

There are two approaches to finding a better PF solution. The first is to come up with a new state-space harmonic model which implicitly includes probable jumps in frequency and moments of silence. Currently, these events are possible but not probable events of the state space model, requiring trade-offs in the tuning process and limiting the accuracy of the state estimates. The second approach is to further refine the PF algorithms with applications to vocalization-like signals in mind. We made a lot of progress in the development of the MAM-PF algorithm which drastically improved our ability to track vocalizations. Further work in this direction will undoubtedly be productive.

Multi-unit recording. The neurophysiological results reported in this dissertation were all based on a lab setup where only a single high impedance electrode was used. This setup results in high fidelity recordings of individual neurons, but makes it difficult to record from neurons that are known to be close to one another. Furthermore, this setup only facilitates the recording of inter-trial response correlations between different neurons, and not within-trial correlations. However, heterogeneity has also been investigated as means for neural networks to overcome the effects of within-trial correlations [Shamir and Sompolinsky, 2006, Chelaru and Dragoi, 2008]. Used in conjunction with multi-electrode or low impedance recordings, a similar ex-

perimental design as the one reported in this dissertation could measure both the inter-trial and within-trial correlations on neighboring neurons which are likely in the same iso-frequency laminae of the ICC. This would provide more insight into the true heterogeneity of responses in the ICC.

More knowledge of the vocalizations used as stimuli. It has been hypothesized that higher level neural representations match the statistical and behavioral qualities of external stimuli [Barlow et al., 1989]. Based on this hypothesis, one possible explanation for the selectivity we discovered for the natural variant of the female upsweep vocalization was that there was little natural variability in this vocalization, making it unnecessary to encode the other perturbations within the ICC. It could also be hypothesized that we did not find any neurons that responded to the male upsweep vocalization because it was not behaviorally important for the female mice that we recorded from. A better understanding of the behavioral contexts and natural variability of these vocalizations would allow us to assess how this affects the distribution of neural responses in the ICC.

Bibliography

- Cornelius Abel and Manfred Kössl. Sensitive response to low frequency cochlear distortion products in the auditory midbrain. *J Neurophysiol*, 101:1560–1574, Nov 2009. doi: 10.1152/jn.90805.2008. URL <http://dx.doi.org/10.1152/jn.90805.2008>.
- J. C. Adams. Ascending projections to the inferior colliculus. *J Comp Neurol*, 183(3):519–538, Feb 1979. doi: 10.1002/cne.901830305. URL <http://dx.doi.org/10.1002/cne.901830305>.
- A. M. Aertsen and P. I. Johannesma. The spectro-temporal receptive field: A functional characteristic of auditory neurons. *Biol Cybern*, 42(2):133–143, 1981.
- L. M. Aitkin and S. C. Phillips. Is the inferior colliculus an obligatory relay in the cat auditory system? *Neurosci Lett*, 44(3):259–264, Feb 1984.
- Noopur Amin, Allison Doupe, and Frdric E Theunissen. Development of selectivity for natural sounds in the songbird auditory forebrain. *J Neurophysiol*, 97(5):3517–3531, May 2007. doi: 10.1152/jn.01066.2006. URL <http://dx.doi.org/10.1152/jn.01066.2006>.
- Sari Andoni, Na Li, and George D Pollak. Spectrotemporal receptive fields in the inferior colliculus revealing selectivity for spectral motion in conspecific vocalizations. *J Neurosci*, 27(18):4882–4893, May 2007. doi: 10.1523/JNEUROSCI.4342-06.2007. URL <http://dx.doi.org/10.1523/JNEUROSCI.4342-06.2007>.
- J.J. Atick. Could information theory provide an ecological theory of sensory processing? *Network-Comp Neural*, 3:213–251, 1992.
- Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nat Rev Neurosci*, 7(5):358–366, May 2006. doi: 10.1038/nrn1888. URL <http://dx.doi.org/10.1038/nrn1888>.
- H. B. Barlow. *Current Problems in Animal Behaviour*, chapter The coding of sensory messages, pages 331–360. Cambridge University Press, 1960.
- H. B. Barlow, T. P. Kaushal, and G. J. Mitchison. Finding minimum entropy codes. *Neural Comput.*, 1(3):412–423, 1989. ISSN 0899-7667.
- Eric E Bauer, Achim Klug, and George D Pollak. Spectral determination of responses to species-specific calls in the dorsal nucleus of the lateral lemniscus. *J Neurophysiol*, 88(4):1955–1967, Oct 2002.
- A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput*, 7(6):1129–1159, Nov 1995.

- W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland. Reading a neural code. *Science*, 252(5014):1854–1857, Jun 1991.
- Sergio Bittanti and Sergio M. Savaresi. On the parameterization and design of an extended Kalman filter frequency tracker. *IEEE Transactions on Automatic Control*, 45(9):1718–1724, Sep 2000.
- A. Brand, R. Urban, and B. Grothe. Duration tuning in the mouse auditory midbrain. *J Neurophysiol*, 84(4):1790–1799, Oct 2000.
- A. E. Bryson and M. Frazier. Smoothing for linear and nonlinear dynamic systems. Technical Report TDR 63-119, Aero. Sys. Div. Wright-Patterson Air Force Base, Ohio, 1963.
- Olivier Cappé, Simon Godsill, and Eric Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. In *Proceedings of the IEEE*, volume 95, pages 899–924. IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC, May 2007.
- George Casella and Christian P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, Mar. 1996.
- T. A. Pincherli Castellanos, J. Aitoubah, S. Molotchnikoff, F. Lepore, and J-P. Guillemot. Responses of inferior collicular cells to species-specific vocalizations in normal and enucleated rats. *Exp Brain Res*, 183(3):341–350, Nov 2007. doi: 10.1007/s00221-007-1049-2. URL <http://dx.doi.org/10.1007/s00221-007-1049-2>.
- Ali Taylan Cemgil and Bert Kappen. Monte carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research (JAIR)*, 18:45–81, 2003.
- R. S. Chadwick. Compression, gain, and nonlinear distortion in an active cochlear model with subpartitions. *Proc Natl Acad Sci U S A*, 95(25):14594–14599, Dec 1998.
- Gal Chechik, Michael J Anderson, Omer Bar-Yosef, Eric D Young, Naftali Tishby, and Israel Nelken. Reduction of information redundancy in the ascending auditory pathway. *Neuron*, 51(3):359–368, Aug 2006. doi: 10.1016/j.neuron.2006.06.030. URL <http://dx.doi.org/10.1016/j.neuron.2006.06.030>.
- Mircea I Chelaru and Valentin Dragoi. Efficient coding in heterogeneous neuronal populations. *Proc Natl Acad Sci U S A*, 105(42):16344–16349, Oct 2008. doi: 10.1073/pnas.0807744105. URL <http://dx.doi.org/10.1073/pnas.0807744105>.
- N. P. Cooper and W. S. Rhode. Mechanical responses to two-tone distortion products in the apical and basal turns of the mammalian cochlea. *J Neurophysiol*, 78(1):261–270, Jul 1997.

- R. Crochiere. A weighted overlap-add method of short-time fourier analysis/synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):99–102, Feb 1980.
- E. de Boer and H. R. de Jongh. On cochlear encoding: potentialities and limitations of the reverse-correlation technique. *J Acoust Soc Am*, 63(1):115–135, Jan 1978.
- R. de Boer and P. Kuyper. Triggered correlation. *IEEE Trans Biomed Eng*, 15(3):169–179, Jul 1968.
- N. de Freitas. Rao-blackwellised particle filtering for fault diagnosis. In *Aerospace Conference Proceedings, 2002. IEEE*, volume 4, pages 4–1767– 4–1772, 2002.
- Christopher DiMattina and Xiaoqin Wang. Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. *J Neurophysiol*, 95(2):1244–1262, Feb 2006. doi: 10.1152/jn.00818.2005.
- Petar M. Djurić, Jayesh H. Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F. Bugallo, and Joaquín Míguez. Particle filtering. *IEEE Signal Processing Magazine*, 20(5):19–38, Sep 2003.
- Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, Jul 2000.
- A. J. Doupe. Song- and order-selective neurons in the songbird anterior forebrain and their emergence during vocal development. *J Neurosci*, 17(3):1147–1167, Feb 1997.
- Corentin Dubois and Manuel Davy. Harmonic tracking using sequential monte carlo. In *IEEE Workshop on Statistical Signal Processing Proceedings*, pages 1292–1297, Bordeaux, France, Jul 2005. Institute of Electrical and Electronics Engineers Computer Society.
- Corentin Dubois and Manuel Davy. Joint detection and tracking of time-varying harmonic components: A flexible bayesian approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1283–1295, May 2007.
- J. J. Eggermont. Wiener and Volterra analyses applied to the auditory system. *Hear Res*, 66(2):177–201, Apr 1993.
- J. J. Eggermont, A. M. Aertsen, and P. I. Johannesma. Prediction of the responses of auditory neurons in the midbrain of the grass frog based on the spectro-temporal receptive field. *Hear Res*, 10(2):191–202, May 1983.
- M. Egorova, G. Ehret, I. Vartanian, and K. H. Esser. Frequency response areas of neurons in the mouse inferior colliculus. i. threshold and tuning characteristics. *Exp Brain Res*, 140(2):145–161, Sep 2001.

- G. Ehret. Mice and humans perceive multiharmonic communication sounds in the same way. *Proc Natl Acad Sci U.S.A.*, 99(1):479–482, Jan 2002. doi: 10.1073/pnas.012361999.
- Gunter Ehret and Raymond Romand. *The Central Auditory System*, chapter The auditory midbrain, a "shunting yard" of acoustical information processing, pages 259–316. Oxford University Press, USA, 1996. ISBN 0195096843.
- Monty A Escabi and Heather L Read. Representation of spectrotemporal sound information in the ascending auditory pathway. *Biol Cybern*, 89(5):350–362, Nov 2003. doi: 10.1007/s00422-003-0440-8.
- D. G. Evans, M. Sainio, and M. E. Baser. Neurofibromatosis type 2. *J Med Genet*, 37(12):897–904, Dec 2000.
- Richard A Felix and Christine V Portfors. Excitatory, inhibitory and facilitatory frequency response areas in the inferior colliculus of hearing impaired mice. *Hear Res*, 228(1-2):212–229, Jun 2007. doi: 10.1016/j.heares.2007.02.009. URL <http://dx.doi.org/10.1016/j.heares.2007.02.009>.
- Robert Fettiplace and Carole M Hackney. The sensory and motor roles of auditory hair cells. *Nat Rev Neurosci*, 7(1):19–29, Jan 2006. doi: 10.1038/nrn1828.
- E. Fischler and B. Z. Bobrovsky. Mean time to loose lock of phase tracking by particle filtering. *Signal Processing*, 86:3481–3485, Jun. 2006.
- J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell System Technical Journal*, pages 1493–1509, 1966.
- Gabbiani and Metzner. Encoding and processing of sensory information in neuronal spike trains. *J Exp Biol*, 202:1267–1279, May 1999.
- Robert Galambos and Hallowell Davis. The Response of Single Auditory-Nerve Fibers to Acoustic Stimulation. *J Neurophysiol*, 6(1):39–57, 1943.
- T. J. Gawne and B. J. Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J Neurosci*, 13(7):2758–2771, Jul 1993.
- D. D. Gehr, H. Komiya, and J. J. Eggermont. Neuronal responses in cat primary auditory cortex to natural and altered species-specific calls. *Hear Res*, 150(1-2): 27–42, Dec 2000.
- A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, Sep 1986.

- Patrick Gill, Junli Zhang, Sarah M. N. Woolley, Thane Fremouw, and F. E. Theunissen. Sound representation methods for spectro-temporal receptive field estimation. *J Comput Neurosci*, 21(1):5–20, Aug 2006. doi: 10.1007/s10827-006-7059-4.
- Simon Godsill, Arnaud Doucet, and Mike West. Maximum a posteriori sequence estimation using monte carlo particle filters. *Ann. Inst. Statist. Math.*, 53(1):82–96, 2001.
- T. Gold. Hearing. II. The Physical Basis of the Action of the Cochlea. *Royal Society of London Proceedings Series B*, 135:492–498, December 1948.
- T. Gold and R. J. Pumphrey. Hearing. I. The Cochlea as a Frequency Analyzer. *Royal Society of London Proceedings Series B*, 135:462–491, December 1948.
- J. L. Goldstein. Auditory nonlinearity. *The Journal of the Acoustical Society of America*, 41(3):676–699, 1967. doi: 10.1121/1.1910396.
- J.L. Goldstein and N.Y.S. Kiang. Neural correlates of the aural combination tone 2f1-f2. *Proceedings of the IEEE*, 56(6):981–992, 1968. ISSN 0018-9219.
- Julie A Grace, Noopur Amin, Nandini C Singh, and Frdric E Theunissen. Selectivity for conspecific song in the zebra finch auditory forebrain. *J Neurophysiol*, 89(1):472–487, Jan 2003. doi: 10.1152/jn.00088.2002. URL <http://dx.doi.org/10.1152/jn.00088.2002>.
- J. L. Hall and M. R. Schroeder. Monaural phase effects for two-tone signals. *The Journal of the Acoustical Society of America*, 51(6B):1882–1884, 1972. doi: 10.1121/1.1913047.
- S. Shuichi Haupt, Friederike Spengler, Robert Husemann, and Hubert R Dinse. Receptive field scatter, topography and map variability in different layers of the hindpaw representation of rat somatosensory cortex. *Exp Brain Res*, 155(4):485–499, Apr 2004. doi: 10.1007/s00221-003-1755-3. URL <http://dx.doi.org/10.1007/s00221-003-1755-3>.
- M. H. Hayes, J. S. Lim, and A. V. Oppenheim. Signal Reconstruction from Phase or Magnitude. *IEEE Transactions on Acoustics Speech and Signal Processing*, 28:672–680, 1980.
- Hermann Helmholtz. *On the Sensations of Tone*. Dover Publications, 1853. ISBN 0486607534.
- L. Holmstrom, R. Santiago, and G.G. Lendaris. Designing an adaptive systems identifier through reinforcement learning. *Submitted to IJCNN*, 2005.

- L. Holmstrom, A. Toland, and G. Lendaris. Experience Based Surface Discernment by a Quadruped Robot. In *IEEE Symposium on Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007*, pages 409–414, 2007a.
- Lars Holmstrom, Patrick D Roberts, and Christine V Portfors. Responses to social vocalizations in the inferior colliculus of the mustached bat are influenced by secondary tuning curves. *J Neurophysiol*, 98(6):3461–3472, Dec 2007b. doi: 10.1152/jn.00638.2007. URL <http://dx.doi.org/10.1152/jn.00638.2007>.
- Lars Holmstrom, Sunghan Kim, James McNames, and Christine Portfors. Stimulus design for auditory neuroethology using state space modeling and the extended kalman smoother. *Hear Res*, 247(1):1–16, Jan 2009. doi: 10.1016/j.heares.2008.09.012. URL <http://dx.doi.org/10.1016/j.heares.2008.09.012>.
- Lars A Holmstrom, Lonneke B M Eeuwes, Patrick D Roberts, and Christine V Portfors. Efficient encoding of vocalizations in the auditory midbrain. *J Neurosci*, 30(3):802–819, Jan 2010. doi: 10.1523/JNEUROSCI.1964-09.2010. URL <http://dx.doi.org/10.1523/JNEUROSCI.1964-09.2010>.
- D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *J Physiol*, 160:106–154, Jan 1962.
- Chlo Huetz, Bndicte Philibert, and Jean-Marc Edeline. A spike-timing code for discriminating conspecific vocalizations in the thalamocortical system of anesthetized and awake guinea pigs. *J Neurosci*, 29(2):334–350, Jan 2009. doi: 10.1523/JNEUROSCI.3269-08.2009. URL <http://dx.doi.org/10.1523/JNEUROSCI.3269-08.2009>.
- F. Hutter and R. Dearden. The gaussian particle filter for diagnosis of non-linear systems. In *Proceedings of the 14th International Conference on Principles of Diagnosis (DX’03)*, pages 65–70, 2003.
- Michael Isard and Andrew Blake. Condensationconditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, Aug. 1998.
- Ben James, Brian D.O. Anderson, and Robert C. Williamson. Conditional mean and maximum likelihood approaches to multiharmonic frequency estimation. *IEEE Transactions on Signal Processing*, 42(6):1366–1375, June 1994. doi: 10.1109/78.286953.
- P. Jensfelt, D.J. Austin, O. Wijk, and M. Andersson. Feature based condensation for mobile robot localization. In *Robotics and Automation, 2000. Proceedings. ICRA ’00. IEEE International Conference on*, volume 3, pages 2531–2537, San Francisco, CA, USA, Apr. 2000. IEEE.

- Don H. Johnson. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *The Journal of the Acoustical Society of America*, 68(4):1115–1122, 1980. doi: 10.1121/1.384982.
- B. M. Johnstone and A. J. Boyle. Basilar membrane vibration examined with the Mössbauer technique. *Science*, 158(799):389–390, Oct 1967.
- R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, 82:35–45, 1960.
- Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell. *Principles of Neural Science*. McGraw-Hill Medical, 2000. ISBN 0838577016.
- Y. Katsuki, T. Ssumi, H. Uchiyama, and T. Watanabe. Electric responses of auditory neurons in cat to sound stimulation. *J Neurophysiol*, 21(6):569–588, Nov 1958.
- D. T. Kemp. Stimulated acoustic emissions from within the human auditory system. *The Journal of the Acoustical Society of America*, 64(5):1386–1391, 1978. doi: 10.1121/1.382104.
- M.S. Keshner. 1/f noise. *Proceedings of the IEEE*, 70(3):212–218, March 1982. ISSN 0018-9219.
- N. Y. Kiang and E. C. Moxon. Tails of tuning curves of auditory-nerve fibers. *J Acoust Soc Am*, 55(3):620–630, Mar 1974.
- D. O. Kim, C. E. Molnar, and J. W. Matthews. Cochlear mechanics: Nonlinear behavior in two-tone responses as reflected in cochlear-nerve-fiber responses and in ear-canal sound pressure. *Acoustical Society of America Journal*, 67:1704–1721, May 1980.
- P. J. Kim and E. D. Young. Comparative analysis of spectro-temporal receptive fields, reverse correlation functions, and frequency tuning curves of auditory-nerve fibers. *J Acoust Soc Am*, 95(1):410–422, Jan 1994.
- G. Kitagawa. Monte-carlo filter and smoother for non-gaussian nonlinear state sapace models. *J. Comput. and Graph. Statist.*, 1:1–25, 1996.
- D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma. Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. *J Comput Neurosci*, 9(1):85–111, 2000.
- Achim Klug, Eric E Bauer, Joshua T Hanson, Laura Hurley, John Meitzen, and George D Pollak. Response selectivity for species-specific calls in the inferior colliculus of mexican free-tailed bats is generated by inhibition. *J Neurophysiol*, 88(4):1941–1954, Oct 2002.

- Barbara F. La Scala and Robert R. Bitmead. Design of an extended Kalman filter frequency tracker. *IEEE Transactions on Signal Processing*, 44(3):739–742, March 1996. doi: 10.1109/78.489052.
- Barbara F. La Scala, Robert F. Bitmead, and Matthew R. James. Conditions for stability of the extended Kalman filter and their application to the frequency tracking problem. *Mathematics of Control, Signals, and Systems*, 8(1):1–26, 1995.
- Barbara F. La Scala, Robert R. Bitmead, and Barry G. Quinn. An extended Kalman filter frequency tracker for high-noise environments. *IEEE Transactions on Signal Processing*, 44(2):431–434, February 1996.
- J. Laroche, Y. Stylianou, and E. Moulines. Hnm: a simple, efficient harmonic+noise model for speech. *Applications of Signal Processing to Audio and Acoustics, 1993. Final Program and Paper Summaries., 1993 IEEE Workshop on*, pages 169–172, Oct 1993. doi: 10.1109/ASPAA.1993.379970.
- C. Lee, W. H. Rohrer, and D. L. Sparks. Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, 332(6162):357–360, Mar 1988. doi: 10.1038/332357a0. URL <http://dx.doi.org/10.1038/332357a0>.
- Y. W. Lee and M. Schetzen. Measurement of the Wiener kernels of a non-linear system by cross-correlation. *International Journal of Control*, 2:237–254, 1965.
- Minoo Lenarz, Hubert H Lim, James F Patrick, David J Anderson, and Thomas Lenarz. Electrophysiological validation of a human prototype auditory midbrain implant in a guinea pig model. *J Assoc Res Otolaryngol*, 7(4):383–398, Dec 2006a. doi: 10.1007/s10162-006-0056-5.
- Thomas Lenarz, Hubert H Lim, Guenter Reuter, James F Patrick, and Minoo Lenarz. The auditory midbrain implant: a new auditory prosthesis for neural deafness-concept and device description. *Otol Neurotol*, 27(6):838–843, Sep 2006b. doi: 10.1097/01.mao.0000232010.01116.e9.
- M. C. Liberman. The cochlear frequency map for the cat: labeling auditory-nerve fibers of known characteristic frequency. *J Acoust Soc Am*, 72(5):1441–1449, Nov 1982.
- Hubert H Lim, Thomas Lenarz, Gert Joseph, Rolf-Dieter Battmer, Amir Samii, Majid Samii, James F Patrick, and Minoo Lenarz. Electrical stimulation of the mid-brain for hearing restoration: insight into the functional organization of the human central auditory system. *J Neurosci*, 27(49):13541–13551, Dec 2007. doi: 10.1523/JNEUROSCI.3123-07.2007.
- R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

- Jun S. Liu and Rong Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1033, 1998.
- Robert C Liu and Christoph E Schreiner. Auditory cortical detection and discrimination correlates with communicative significance. *PLoS Biol*, 5(7):e173, Jun 2007. doi: 10.1371/journal.pbio.0050173. URL <http://dx.doi.org/10.1371/journal.pbio.0050173>.
- A. N. Lukashkin and I. J. Russell. A descriptive model of the receptor potential nonlinearities generated by the hair cell mechanoelectrical transducer. *J Acoust Soc Am*, 103(2):973–980, Feb 1998.
- Manuel S. Malmierca. Direct projections from cochlear nuclear complex to auditory thalamus in the rat. *J Neurosci*, 22(24):10891–10897, Dec 2002.
- Manuel S. Malmierca. The inferior colliculus: A center for convergence of ascending and descending auditory information. *Neuroembryol Aging*, 3:215–229, 2004/2005.
- Dimitris G. Manolakis, Vinay K. Ingle, Vinay Ingle, and Stephen M. Kogon. *Statistical And Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering And Array Processing*. Artech House, 2005.
- D. Margoliash. Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *J Neurosci*, 3(5):1039–1057, May 1983.
- D. Margoliash. Preference for autogenous song by auditory neurons in a song system nucleus of the white-crowned sparrow. *J Neurosci*, 6(6):1643–1661, Jun 1986.
- David McAlpine. Neural sensitivity to periodicity in the inferior colliculus: evidence for the role of cochlear distortions. *J Neurophysiol*, 92(3):1295–1311, Sep 2004. doi: 10.1152/jn.00034.2004.
- M. M. Merzenich and M. D. Reid. Representation of the cochlea within the inferior colliculus of the cat. *Brain Res*, 77(3):397–415, Sep 1974.
- N. Metropolis, AW Rosenbluth, and MN Rosenbluth. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087, Jun. 1953.
- Kenneth A. Myers and Byron D. Tapley. Adaptive sequential estimation with unknown noise statistics. *IEEE Transactions on Automatic Control*, 21(4):520–523, Aug. 1976.
- J-P Nadal and N Parga. Nonlinear neurons in the low noise limit: a factorial code maximizes information transfer. *Network-Computation in Neural Systems*, 5:565–581, 1994.

- Kiran Nataraj and Jeffrey J Wenstrup. Roles of inhibition in creating complex auditory responses in the inferior colliculus: facilitated combination-sensitive neurons. *J Neurophysiol*, 93(6):3294–3312, Jun 2005. doi: 10.1152/jn.01152.2004.
- M. Nomoto, N. Suga, and Y. Katsuki. Discharge pattern and inhibition of primary auditory nerve fibers in the monkey. *J Neurophysiol*, 27:768–787, Sep 1964.
- A. L. Nuttall and D. F. Dolan. Intermodulation distortion (f2-f1) in inner hair cell and basilar membrane responses. *J Acoust Soc Am*, 93(4 Pt 1):2061–2068, Apr 1993.
- K.K. Ohlemiller, J.S. Kanwal, J.A. Butman, and N. Suga. Stimulus design for auditory neuroethology: Synthesis and manipulation of complex communication sounds. *Auditory Neuroscience*, 1:19–37, 1994.
- G. S. Ohm. Ueber die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. *Annalen der Physik*, 135:513–565, 1843. doi: 10.1002/andp.18431350802.
- Jules B Panksepp, Kimberly A Jochman, Joseph U Kim, Jamie J Koy, Ellie D Wilson, Qiliang Chen, Clarinda R Wilson, and Garet P Lahvis. Affiliative behavior, ultrasonic communication and social reward are influenced by genetic variation in adolescent mice. *PLoS ONE*, 2(4):e351, 2007. doi: 10.1371/journal.pone.0000351. URL <http://dx.doi.org/10.1371/journal.pone.0000351>.
- Philip J. Parker and Brian D.O. Anderson. Frequency tracking of nonsinusoidal periodic signals in noise. *Signal Processing*, 20(2):127–152, Jun 1990.
- David Pérez-González, Manuel S. Malmierca, and Ellen Covey. Novelty detector neurons in the mammalian auditory midbrain. *Eur J Neurosci*, 22(11):2879–2885, Dec 2005. doi: 10.1111/j.1460-9568.2005.04472.x.
- David Pérez-González, Manuel S. Malmierca, J. M. Moore, O. Hernández, and Ellen Covey. Duration selective neurons in the inferior colliculus of the rat: topographic distribution and relation of duration sensitivity to other response properties. *J Neurophysiol*, 95(2):823–836, Feb 2006. doi: 10.1152/jn.00741.2005.
- B. Philibert, J. Laudanski, and J-M. Edeline. Auditory thalamus responses to guinea-pig vocalizations: a comparison between rat and guinea-pig. *Hear Res*, 209(1-2): 97–103, Nov 2005. doi: 10.1016/j.heares.2005.07.004. URL <http://dx.doi.org/10.1016/j.heares.2005.07.004>.
- R. Plomp. The ear as a frequency analyzer. *The Journal of the Acoustical Society of America*, 36(9):1628–1636, 1964. doi: 10.1121/1.1919256.
- R. Plomp. Detectability threshold for combination tones. *The Journal of the Acoustical Society of America*, 37(6):1110–1123, 1965. doi: 10.1121/1.1909532.

- G. D. Pollak, D. S. Marsh, R. Bodenhamer, and A. Souther. Characteristics of phasic on neurons in inferior colliculus of unanesthetized bats with observations relating to mechanisms for echo ranging. *J Neurophysiol*, 40(4):926–942, Jul 1977.
- George D Pollak, Achim Klug, and Eric E Bauer. Processing and representation of species-specific communication calls in the auditory system of bats. *Int Rev Neurobiol*, 56:83–121, 2003.
- C. V. Portfors and R. A. Felix. Spectral integration in the inferior colliculus of the cba/caj mouse. *Neuroscience*, 136(4):1159–1170, 2005. doi: 10.1016/j.neuroscience.2005.08.031. URL <http://dx.doi.org/10.1016/j.neuroscience.2005.08.031>.
- C. V. Portfors, P. D. Roberts, and K. Jonson. Over-representation of species-specific vocalizations in the awake mouse inferior colliculus. *Neuroscience*, 162(2):486–500, Aug 2009. doi: 10.1016/j.neuroscience.2009.04.056. URL <http://dx.doi.org/10.1016/j.neuroscience.2009.04.056>.
- Christine V Portfors. Combination sensitivity and processing of communication calls in the inferior colliculus of the moustached bat *pteronotus parnellii*. *An Acad Bras Cienc*, 76(2):253–257, Jun 2004. doi: /S0001-37652004000200010. URL <http://dx.doi.org/S0001-37652004000200010>.
- Christine V Portfors. Types and functions of ultrasonic vocalizations in laboratory rats and mice. *J Am Assoc Lab Anim Sci*, 46(1):28–34, Jan 2007.
- Christine V. Portfors and Jeffrey Wenstrup. Excitatory and facilitatory frequency response areas in the inferior colliculus of the mustached bat. *Hear Res*, 168(1-2): 131–138, Jun 2002.
- Christine V. Portfors and Jeffrey Wenstrup. Delay-tuned neurons in the inferior colliculus of the mustached bat: implications for analyses of target distance. *J Neurophysiol*, 82(3):1326–1338, Sep 1999.
- Christine V. Portfors and Jeffrey J. Wenstrup. Responses to combinations of tones in the nuclei of the lateral lemniscus. *J Assoc Res Otolaryngol*, 2(2):104–117, Jun 2001.
- M. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transaction on Acoustics, Speech, and Signal Processing*, 24(3):243–248, 1976. ISSN 0096-3518.
- M. R. Portnoff. Time-scale modification of speech based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29, 1981.
- J. P. Rauschecker and R. V. Shannon. Sending sound to the brain. *Science*, 295(5557):1025–1029, Feb 2002. doi: 10.1126/science.1067796.

- William S. Rhode. Observations of the vibration of the basilar membrane in squirrel monkeys using the m[ö]ssbauer technique. *The Journal of the Acoustical Society of America*, 49(4B):1218–1231, 1971. doi: 10.1121/1.1912485.
- F. Rieke, D. Warland, and W. Bialek. Coding efficiency and information rates in sensory neurons. *Europhys Lett*, 22, April 1993.
- F. Rieke, D. A. Bodnar, and W. Bialek. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc Biol Sci*, 262(1365):259–265, Dec 1995. doi: 10.1098/rspb.1995.0204. URL <http://dx.doi.org/10.1098/rspb.1995.0204>.
- Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: Exploring the Neural Code (Computational Neuroscience)*. The MIT Press, 1999. ISBN 0262681080.
- Dario L Ringach, Robert M Shapley, and Michael J Hawken. Orientation selectivity in macaque v1: diversity and laminar dependence. *J Neurosci*, 22(13):5639–5651, Jul 2002. doi: 20026567. URL <http://dx.doi.org/20026567>.
- Ariel Rokem, Sebastian Watzl, Tim Gollisch, Martin Stemmler, Andreas V M Herz, and Ins Samengo. Spike-timing precision underlies the coding efficiency of auditory receptor neurons. *J Neurophysiol*, 95(4):2541–2552, Apr 2006. doi: 10.1152/jn.00891.2005. URL <http://dx.doi.org/10.1152/jn.00891.2005>.
- E. T. Rolls and M. J. Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J Neurophysiol*, 73(2):713–726, Feb 1995.
- E T Rolls and A Treves. The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network-Comp Neural*, 1(4):407–421, 1990. URL <http://stacks.iop.org/0954-898X/1/407>.
- J. E. Rose, J. F. Brugge, D. J. Anderson, and J. E. Hind. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *J Neurophysiol*, 30(4):769–793, Jul 1967.
- Jerzy E. Rose, Donald D. Greenwood, Jay M. Goldberg, and Joseph E. Hind. Some Discharge Characteristics of Single Neurons in the Inferior Colliculus of the Cat. I. Tonotopical Organization, Relation of Spike-Counts to Tone Intensity, and Firing Patterns of Single Elements. *J Neurophysiol*, 26(2):294–320, 1963.
- M. B. Sachs and N. Y. Kiang. Two-tone inhibition in auditory-nerve fibers. *J Acoust Soc Am*, 43(5):1120–1128, May 1968.
- M. B. Sachs, E. D. Young, and M. I. Miller. Speech encoding in the auditory nerve: implications for cochlear implants. *Ann N Y Acad Sci*, 405:94–113, 1983.

- E. Saldaña, M. Feliciano, and E. Mugnaini. Distribution of descending projections from primary auditory neocortex to inferior colliculus mimics the topography of intracollicular projections. *J Comp Neurol*, 371(1):15–40, Jul 1996. doi: [gt;3.0.CO;2-O](https://doi.org/10.1002/gt.3.0.CO;2-O). URL <http://dx.doi.org/gt;3.0.CO;2-O>.
- Amir Samii, Minoo Lenarz, Omid Majdani, Hubert H Lim, Madjid Samii, and Thomas Lenarz. Auditory midbrain implant: a combined approach for vestibular schwannoma surgery and device implantation. *Otol Neurotol*, 28(1):31–38, Jan 2007. doi: 10.1097/01.mao.0000247819.16325.7d.
- J. F. Schouten, R. J. Ritsma, and B. Lopes Cardozo. Pitch of the residue. *The Journal of the Acoustical Society of America*, 34(9B):1418–1424, 1962. doi: 10.1121/1.1918360.
- S. Schreiber, J. M. Fellous, D. Whitmer, P. Tiesinga, and T. J. Sejnowski. A new correlation-based measure of spike timing reliability. *Neurocomputing*, 52-54:925–931, June 2003.
- C. E. Schreiner and G. Langner. Laminar fine structure of frequency organization in auditory midbrain. *Nature*, 388(6640):383–386, Jul 1997. doi: 10.1038/41106.
- A. Seebeck. Ueber die Definition des Tones. *Annalen der Physik*, 139:353–368, 1844. doi: 10.1002/andp.18441391102.
- K. Sen, F. E. Theunissen, and A. J. Doupe. Feature analysis of natural sounds in the songbird auditory forebrain. *J Neurophysiol*, 86(3):1445–1458, Sep 2001.
- Maoz Shamir and Haim Sompolinsky. Nonlinear population codes. *Neural Comput*, 16(6):1105–1136, Jun 2004. doi: 10.1162/089976604773717559. URL <http://dx.doi.org/10.1162/089976604773717559>.
- Maoz Shamir and Haim Sompolinsky. Implications of neuronal diversity on population coding. *Neural Comput*, 18(8):1951–1986, Aug 2006. doi: 10.1162/neco.2006.18.8.1951. URL <http://dx.doi.org/10.1162/neco.2006.18.8.1951>.
- Maoz Shamir, Kamal Sen, and H. Steven Colburn. Temporal coding of time-varying stimuli. *Neural Comput*, 19(12):3239–3261, Dec 2007. doi: 10.1162/neco.2007.19.12.3239. URL <http://dx.doi.org/10.1162/neco.2007.19.12.3239>.
- Donal G Sinex, Jennifer Henderson Sabes, and Hongzhe Li. Responses of inferior colliculus neurons to harmonic and mistuned complex tones. *Hear Res*, 168(1-2):150–162, Jun 2002.
- Donal G Sinex, Heidi Guzik, Hongzhe Li, and Jennifer Henderson Sabes. Responses of auditory nerve fibers to harmonic and mistuned complex tones. *Hear Res*, 182(1-2):130–139, Aug 2003.

- M. Slaney and R. Lyon. *Visual Representations of Speech Signals*, chapter 5, pages 95–116. John Wiley & Sons Ltd., 1993.
- H. Sompolinsky, H. Yoon, K. Kang, and M. Shamir. Population coding in neuronal systems with correlated noise. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(5 Pt 1):051904, Nov 2001.
- R. B. Stein, A. S. French, and A. V. Holden. The frequency response, coherence, and information capacity of two neuronal models. *Biophys J*, 12(3):295–322, Mar 1972.
- M. Steinschneider, I. O. Volkov, M. D. Noh, P. C. Garell, and M. A. Howard. Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *J Neurophysiol*, 82(5):2346–2357, Nov 1999.
- N. Suga. Philosophy and stimulus design for neuroethology of complex-sound processing. *Philos Trans R Soc Lond B Biol Sci*, 336(1278):423–428, Jun 1992. doi: 10.1098/rstb.1992.0078.
- N. Suga and P. H. Jen. Disproportionate tonotopic representation for processing cf-fm sonar signals in the mustache bat auditory cortex. *Science*, 194(4264):542–544, Oct 1976.
- N. Suga, W. E. O’Neill, and T. Manabe. Harmonic-sensitive neurons in the auditory cortex of the mustache bat. *Science*, 203(4377):270–274, Jan 1979.
- Daniel Suta, Eugen Kvasnk, Jir Popelr, and Josef Syka. Representation of species-specific vocalizations in the inferior colliculus of the guinea pig. *J Neurophysiol*, 90(6):3794–3808, Dec 2003. doi: 10.1152/jn.01175.2002. URL <http://dx.doi.org/10.1152/jn.01175.2002>.
- F. E. Theunissen. Methods for the analysis of auditory processing in the brain. *Ann N Y Acad Sci*, 1016:187–207, Jun 2004. doi: 10.1196/annals.1298.020.
- F. E. Theunissen and A. J. Doupe. Temporal and spectral sensitivity of complex auditory neurons in the nucleus hvc of male zebra finches. *J Neurosci*, 18(10):3786–3802, May 1998.
- F. E. Theunissen, K. Sen, and A. J. Doupe. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J Neurosci*, 20(6):2315–2331, Mar 2000.
- F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network*, 12(3):289–316, August 2001. ISSN 0954-898X.

- W. R. Thurlow, G. W. N. B. Gross, D. T. E. H. Kemp, and K. Lowy. Microelectrode studies of neural auditory activity of cat. i. inferior colliculus. *J Neurophysiol*, 14(4):289–304, Jul 1951.
- A.H. Toland, L.A. Holstrom, and G.G. Lendaris. Effectiveness of a Coupled Oscillator Network for Surface Discernment by a Quadruped Robot based on Kinesthetic Experience. In *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, pages 2171–2176, 2007.
- M. C. van Rossum. A novel spike distance. *Neural Comput*, 13(4):751–763, Apr 2001.
- J. Vermaak, C. Andrieu, A. Doucet, and S. Godsill. Particle methods for bayesian modeling and enhancement of speechsignals. *IEEE Transactions on Speech and Audio Processing*, 10(3):173–185, Mar. 2002.
- J. D. Victor and K. P. Purpura. Nature and precision of temporal coding in visual cortex: a metric-space analysis. *J Neurophysiol*, 76(2):1310–1326, Aug 1996.
- Jonathan D Victor and Keith P Purpura. Metric-space analysis of spike trains: theory, algorithms and application. *Network: Computation in Neural Systems*, 8(2):127–164, 1997. doi: 10.1088/0954-898X_8_2_003. URL http://www.informahealthcare.com/doi/abs/10.1088/0954-898X_8_2_003.
- A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEE Transactions on Information Theory*, 13(2):260–269, Apr. 1967.
- V. Volterra. *Theory of functionals and of integro-differential equations*. Dover Publications, 1930.
- Georg von Békésy. The vibration of the cochlear partition in anatomical preparations and in models of the inner ear. *The Journal of the Acoustical Society of America*, 21(3):233–245, 1949a. doi: 10.1121/1.1906502.
- Georg von Békésy. On the resonance curve and the decay period at various points on the cochlear partition. *The Journal of the Acoustical Society of America*, 21(3):245–254, 1949b. doi: 10.1121/1.1906503.
- Haoran Wang, Shuyin Liang, Jeffrey Burgdorf, Jurgen Wess, and John Yeomans. Ultrasonic vocalizations induced by sex and amphetamine in m2, m4, m5 muscarinic and d2 dopamine receptor knockout mice. *PLoS ONE*, 3(4):e1893, 2008. doi: 10.1371/journal.pone.0001893. URL <http://dx.doi.org/10.1371/journal.pone.0001893>.
- Le Wang, Rajiv Narayan, Gilberto Graa, Maoz Shamir, and Kamal Sen. Cortical discrimination of complex natural stimuli: can single neurons match behavior? *J Neurosci*, 27(3):582–589, Jan 2007. doi: 10.1523/JNEUROSCI.3699-06.2007.

- X. Wang and S. C. Kadia. Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. *J Neurophysiol*, 86(5): 2616–2620, Nov 2001.
- X. Wang, M. M. Merzenich, R. Beitel, and C. E. Schreiner. Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *J Neurophysiol*, 74(6):2685–2706, Dec 1995.
- Xiaoqin Wang, Thomas Lu, Ross K Snider, and Li Liang. Sustained firing in auditory cortex evoked by preferred stimuli. *Nature*, 435(7040):341–346, May 2005. doi: 10.1038/nature03565. URL <http://dx.doi.org/10.1038/nature03565>.
- Ben Warren, Gabriella Gibson, and Ian J Russell. Sex recognition through midflight mating duets in culex mosquitoes is mediated by acoustic distortion. *Curr Biol*, 19(6):485–491, Mar 2009. doi: 10.1016/j.cub.2009.01.059. URL <http://dx.doi.org/10.1016/j.cub.2009.01.059>.
- Ernest Glen Wever. The Physiology of Hearing: The Nature of Response in the Cochlea. *Physiol. Rev.*, 13(3):400–425, 1933.
- Ernest Glen Wever and Charles W. Bray. Action Currents in the Auditory Nerve in Response to Acoustical Stimulation. *Proceedings of the National Academy of Sciences*, 16(5):344–350, 1930. doi: 10.1073/pnas.16.5.344.
- Langford B. White. An iterative method for exact maximum likelihood estimation of the parameters of a harmonic series. *IEEE Transactions on Automatic Control*, 38(2):367–370, Feb 1993.
- N. Wiener. *Nonlinear Problems in Random Theory*. Wiley, New York, 1958.
- Jeffery A. Winer and Christoph E. Schreiner. *The Inferior Colliculus*. Springer, 2005. ISBN 0387220380.
- Sarah M. N. Woolley and John H. Casseday. Processing of Modulated Sounds in the Zebra Finch Auditory Midbrain: Responses to Noise, Frequency Sweeps, and Sinusoidal Amplitude Modulations. *J Neurophysiol*, 94(2):1143–1157, 2005. doi: 10.1152/jn.01064.2004.
- Eric D Young and Barbara M Calhoun. Nonlinear modeling of auditory-nerve rate responses to wideband stimuli. *J Neurophysiol*, 94(6):4441–4454, Dec 2005. doi: 10.1152/jn.00261.2005.

Appendices

Appendix A: State-Space Stimulus Design for Auditory Neuroethology

A.1 Overview

To the best of our knowledge, past efforts in the field of auditory neuroethology have all relied on spectro-temporal representations of vocalizations obtained from the Short-Time Fourier Transform (STFT) for the estimation of signal frequencies and amplitudes. There are, however, significant limitations to this approach. For example there is a time-frequency resolution trade-off in the STFT which limits the accuracy of frequency estimates even in noise-free non-stationary signals [Manolakis et al., 2005]. Another limitation is the difficulty of estimating the frequency, amplitudes, and phases of noisy signals. Even in cases where spectrographic frequency tracking can provide reasonable estimates of the model parameters, an additional optimization step must be taken to fit these parameters in order to faithfully reproduce the original vocalization [Laroche et al., 1993].

The analysis-synthesis methods described here addresses each of these limitations while providing new possibilities for the signal morphing requirements of auditory neuroethological experiments. By using a statistical model-based approach, we show that the Extended Kalman Smoother (EKS) and Particle Filter (PF) methods outperform STFT based frequency tracking methods in synthetic vocalization signals regardless of the noise level, while simultaneously tracking the amplitudes and relative phases of the harmonics of the vocalization. We show that the state-space harmonic tracker accurately analyzes and synthesizes real vocalizations recorded from four species: hummingbird, human, bat, and mouse. All four vocalizations primarily have a multi-harmonic structure, but have different morphologies. The hummingbird recording is the most noise-free of the four and shows a clear harmonic structure throughout the

whole recording. This example shows that the vocalization analysis-synthesis almost perfectly reproduces the original recording. The human vocalization (the phrase *ä-shä*) contains a large number of harmonics in addition to a period of broad-band noise (the “sh”). This example shows that the harmonic model is capable of capturing much of the information present in non-harmonic components of vocalizations while regaining track on the harmonic components following noisy sections. The bat vocalization contains numerous recording artifacts including significant noise, echo, and aliasing. This example shows that the state-space analysis-synthesis method can work as a denoising technique by selectively extracting the harmonic components from the noisy signal and synthesizing the vocalization nearly free of recording artifacts. Finally, the mouse vocalization includes abrupt jumps in frequency, making it a candidate for analysis using the PF. This example shows that the PF is capable of tracking these abrupt jumps better than the EKS.

In the application domain of auditory neurophysiology, these techniques will be useful for de-noising vocalizations that are recorded in noisy field or captive environments. In addition, our synthesis technique will allow for efficient manipulations of the vocalizations (such as decomposing the original signal into its individual harmonics, shifting frequencies either harmonically or an-harmonically, altering temporal cues, etc.) so that specific hypotheses can be tested regarding how vocalizations are encoded and represented in the auditory system.

A.2 Methods

A.2.1 State-Space Signal Modeling

A discrete-time state-space model is a mathematical representation of a physical system that relates a sequence of measurements to a sequence of system states. A

state transition function governs how the states of the system transition through time, possibly under the effect of outside control or a noisy (stochastic) process, and therefore defines the dynamics of the system. A measurement function defines the relationship between these states and possibly noisy (stochastic) measurements of the system.

When using a discrete time state-space representation to model an underlying stochastic process generating an observable signal, the model may be expressed in the following form:

$$\begin{aligned}\mathbf{x}_{n+1} &= f(\mathbf{x}_n) + \mathbf{u}_n \\ \mathbf{y}_n &= h(\mathbf{x}_n) + \mathbf{v}_n\end{aligned}\tag{A.1}$$

where $\mathbf{x}_n$ represents the states of the underlying process, $f(\cdot)$ is the state transition function, $\mathbf{u}_n$ is the process noise, $\mathbf{y}_n$ is the measured signal, $h(\cdot)$ is the measurement function, $\mathbf{v}_n$ is the measurement (recording) noise, and $n = 0, \dots, N$ is the independent variable representing discrete time. In general, both $f(\cdot)$ and $h(\cdot)$ may be nonlinear functions. A visual representation of this process can be seen in (A.2). The horizontal arrows represent the action of the state update function, $f(\cdot)$, and the vertical arrows represent the action of the measurement function, $h(\cdot)$.

$$\begin{array}{ccccccc}\longrightarrow & \mathbf{x}_{n-1} & \longrightarrow & \mathbf{x}_n & \longrightarrow & \mathbf{x}_{n+1} & \longrightarrow \\ & \Downarrow & & \Downarrow & & \Downarrow & \\ & \mathbf{y}_{n-1} & & \mathbf{y}_n & & \mathbf{y}_{n+1} & \end{array}\tag{A.2}$$

Provided that an appropriate state-space representation is chosen to model a given system, algorithms such as the Kalman filter and Particle filter provide a means

of estimating the states of the system given a measured signal generated from the system.

A.2.2 Harmonic Frequency State-Space Signal Model

For modeling a signal that is presumed to be generated by a sum of harmonically related components with independent amplitudes and phase relationships (as is the case for many vocalizations), we used the following signal model, known as the *rectangular model* [Parker and Anderson, 1990]:

$$\mathbf{y}_n = \sum_{k=1}^m \mathbf{a}_{k,n} \cos(k\boldsymbol{\theta}_n) + \mathbf{b}_{k,n} \sin(k\boldsymbol{\theta}_n) + \mathbf{v}_n \quad (\text{A.3})$$

where m is the number of the harmonics, $\boldsymbol{\theta}_n$ is the instantaneous phase of the fundamental frequency component, $\mathbf{a}_{k,n}$ and $\mathbf{b}_{k,n}$ are the amplitudes of the sinusoidal components of the harmonics, and $\mathbf{v}_n$ is a white noise process with zero-mean and variance r . Note how the instantaneous phases $k\boldsymbol{\theta}_n$ of the harmonic components ($k > 1$) are defined to be integer multiples of the instantaneous phase of the fundamental, $\boldsymbol{\theta}_n$. This results in a signal $\mathbf{y}_n$ with components whose frequencies are integer multiples of the fundamental. Also, one should note how the amplitudes $\mathbf{a}_{k,n}$ and $\mathbf{b}_{k,n}$ of the the harmonics are not related, resulting in sinusoidal components of $\mathbf{y}_n$ that have independent amplitudes and phase offsets. At any given time n , the magnitude $A_{k,n}$ and phase offset $\phi_{k,n}$ of the k^{th} component can be calculated by:

$$\begin{aligned} A_{k,n} &= |\mathbf{a}_{k,n} + j\mathbf{b}_{k,n}| \\ \phi_{k,n} &= \angle\{\mathbf{a}_{k,n} + j\mathbf{b}_{k,n}\} \end{aligned} \quad (\text{A.4})$$

The instantaneous quasi-periodic phase $\boldsymbol{\theta}_n$ of the fundamental is approximated as

a first-order numerical integration of the instantaneous fundamental frequency $\mathbf{f}_n$:

$$\boldsymbol{\theta}_{n+1} = \text{mod}_{2\pi}\{\boldsymbol{\theta}_n + 2\pi T_s \mathbf{f}_n\} \quad (\text{A.5})$$

where $T_s = 1/f_s$ is the sampling interval. The modulus operator, $\text{mod}_{2\pi}$, has no effect on the model mathematically, but keeps $\boldsymbol{\theta}_{n+1}$ bounded to $0 \leq \boldsymbol{\theta}_n \leq 2\pi$ and reduces roundoff error in the iterative numerical computations.

The instantaneous fundamental frequency $\mathbf{f}_n$ is modeled as a first-order autoregressive (AR) process whose mean is $\bar{f}$:

$$\mathbf{f}_{n+1} = \alpha (\mathbf{f}_n - \bar{f}) + \bar{f} + \mathbf{u}_{f,n} \quad (\text{A.6})$$

where α is the AR process coefficient, and $\mathbf{u}_{f,n}$ is a white noise process with zero mean and variance q_f . This models a signal whose fundamental frequency has a mean of $\bar{f}$ but may vary from this mean by an amount determined by α and $\mathbf{u}_{f,n}$. A value of $\alpha = 1$ results in a random walk process for $\mathbf{f}_n$ where the variance increases linearly with time. As α decreases from 1, the fundamental frequency $\mathbf{f}_n$ is increasingly biased towards the specified mean frequency $\bar{f}$ of the vocalization. A value of $\alpha = 0$ means that the frequency at the previous time has no bearing on the current time step and results in a white noise process with a mean of $\bar{f}$ and a variance of q_f .

The sinusoidal amplitude coefficients $\mathbf{a}_{k,n}$ and $\mathbf{b}_{k,n}$ are modeled as random walk processes,

$$\mathbf{a}_{k,n+1} = \mathbf{a}_{k,n} + \mathbf{u}_{a,k,n} \quad (\text{A.7})$$

$$\mathbf{b}_{k,n+1} = \mathbf{b}_{k,n} + \mathbf{u}_{b,k,n}$$

where $\mathbf{u}_{.,k,n}$ are mutually uncorrelated white noise processes. By (A.3) and (A.4) it

can be seen that the variance of these processes, represented as $q_{a,k}$ and $q_{b,k}$, determine how quickly the amplitudes and phase relationships of the harmonic components of $\mathbf{y}_n$ may change over time.

By defining the state vector $\mathbf{x}_n$ as:

$$\mathbf{x}_n = \begin{bmatrix} \boldsymbol{\theta}_n & \mathbf{f}_n & \mathbf{a}_{1,n} & \dots & \mathbf{a}_{m,n} & \mathbf{b}_{1,n} & \dots & \mathbf{b}_{m,n} \end{bmatrix}^T \quad (\text{A.8})$$

equations (A.3), (A.5), (A.6), and (A.7) can be rewritten concisely as:

$$\begin{aligned} \mathbf{x}_{n+1} &= f(\mathbf{x}_n) + \mathbf{u}_n \\ &= \begin{bmatrix} \text{mod}_{2\pi}\{\mathbf{x}_{1,n} + 2\pi T_s \mathbf{x}_{2,n}\} \\ \alpha(\mathbf{x}_{2,n} - \bar{f}_n) + \bar{f}_n \\ \mathbf{x}_{3,n} \\ \vdots \\ \mathbf{x}_{2m+1,n} \\ \mathbf{x}_{2m+2,n} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{1,n} \\ \mathbf{u}_{2,n} \\ \mathbf{u}_{3,n} \\ \vdots \\ \mathbf{u}_{2m+1,n} \\ \mathbf{u}_{2m+2,n} \end{bmatrix} \\ \mathbf{y}_n &= h(\mathbf{x}_n) + \mathbf{v}_n \\ &= \sum_{k=1}^m \mathbf{x}_{2+k,n} \cos(k\mathbf{x}_{1,n}) + \mathbf{x}_{2+m+k,n} \sin(k\mathbf{x}_{1,n}) + \mathbf{v}_n \end{aligned} \quad (\text{A.9})$$

where $f(\cdot)$ and $h(\cdot)$ are the state transition and measurement functions, respectively.

To summarize, this state based model describes a signal that is composed of a number of harmonically related components with independent amplitudes and phase offsets. The fundamental frequency of the signal varies over time but has a mean equal to $\bar{f}$. The magnitude by which the frequency can vary in each time step is specified by q_f and its bias towards $\bar{f}$ is specified by α . The amplitudes and phase offsets of the harmonic components also vary over time and the degree by which they

vary is specified by $q_{a,k}$ and $q_{b,k}$.

A.2.3 Vocalization Analysis

Analysis of vocalizations involves estimating the state-space parameter values for the harmonic signal model defined in Appendix A.2.2. It is desired that the estimator provide optimal estimates (in the mean squared error sense) while satisfying the constraints imposed by the statistics of the signal model. Due to the nonlinearity and multi-modality of this estimation problem, both the extended Kalman smoother (EKS) and particle filter methodologies were pursued. Details about both of these methods are included in Appendix B and Appendix C.

A.2.4 Modifying Vocalizations

After completing the analysis phase, the resulting state-space model of the vocalization lends itself to certain ways of “morphing” the signal before synthesis. These modifications include:

- Pitch shifting without changing the signal duration by simply adding a constant value to the frequency estimates. This is considerably less complex and more accurate than comparable methods using spectrographic decompositions of the vocalizations such as phase vocoding [Flanagan and Golden, 1966] — a popular technique in speech and music processing which is limited by a time-frequency tradeoff of the STFT calculation in the analysis phase and requires complicated and/or heuristic algorithms to maintain phase coherence in the synthesized signal.
- Time shifting without affecting the pitch through resampling of the frequency and amplitude estimates and assuming the same sampling frequency. Again,

this is considerably less complex and more accurate than comparable methods such as the phase vocoder.

- Independently change the phase offset of the different harmonic components by adding a constant phase to each component during synthesis.
- Removing frequency modulation while maintaining the original amplitude modulation envelope.
- Removing amplitude modulation while maintaining the original frequency modulation.

Other modifications can be made by filtering or processing the model parameter values in other ways. In general, the signals defined by each model parameter's values are much more narrow in bandwidth than the original vocalization, making filtering or processing of the modeled vocalization more tractable than filtering or processing of the original vocalization.

A.2.5 Synthesizing Vocalizations

Once the states of the harmonic model have been estimated and optionally modified, synthesis of the modeled signal is performed in a two step process. First, the instantaneous quasi-periodic phase θ_n of the fundamental is estimated through a first-order numerical integration of the instantaneous fundamental frequency f_n :

$$\theta_{n+1} = \text{mod}_{2\pi}\{\theta_n + 2\pi T_s f_n\} \quad (\text{A.10})$$

where $T_s = 1/f_s$ is the sampling interval. The modulus operator, $\text{mod}_{2\pi}$, has no affect on the model mathematically, but keeps θ_{n+1} bounded to $0 \leq \theta_n \leq 2\pi$ and

reduces roundoff error in the iterative numerical computations. Second, the observation function can be used to generate the synthesized signal using the instantaneous quasi-periodic phase (calculated in the previous step) and the time-varying amplitudes (calculated in the state estimation process):

$$\mathbf{y}_n = \sum_{k=1}^m \mathbf{a}_{k,n} \cos(k\boldsymbol{\theta}_n) + \mathbf{b}_{k,n} \sin(k\boldsymbol{\theta}_n) \quad (\text{A.11})$$

where m is the number of the harmonics and $\mathbf{a}_{k,n}$ and $\mathbf{b}_{k,n}$ are the amplitudes of the sinusoidal components of the harmonics.

Appendix B: Vocalization Analysis with the Extended Kalman Smoother

B.1 The Kalman Filter

Given a sequence of measurements, the Kalman filter recursively estimates the states of a linear stochastic process such that the mean squared error is minimized [Kalman, 1960]. The Kalman filter can be used to generate both the predicted and filtered estimates of the state variables. The predicted estimates, $\hat{\mathbf{x}}_{n|n-1}$, use only past measurements $\{\mathbf{y}_0, \dots, \mathbf{y}_{n-1}\}$, while the filtered estimates, $\hat{\mathbf{x}}_{n|n}$, also include the current measurement $\{\mathbf{y}_0, \dots, \mathbf{y}_n\}$.

In the context of the harmonic signal model specified in Appendix A.2.2, we are interested in estimating the frequency, amplitude, and phase information of a harmonically related vocalization (the states) by analyzing a noisy recording of the vocalization (the measurements). While the mathematics involved in the derivation of the Kalman filter are deep, an intuitive understanding of how it functions is possible from a higher level description.

The Kalman filter operates through an iterative application of estimation and correction: the filter estimates the states of the process at a specific time based on the dynamics of the state-space model, and corrects these estimates as new information is obtained from noisy measurements (the recorded signal). Based on these two phases of the algorithm, the Kalman recursions can be separated into two groups of equations: the time update equations and the measurement update equations.

The time update equations are responsible for projecting one sample forward (in time) the current state and error covariance estimates to obtain the a priori state and error covariance estimates for the next sample. This can be thought of as the predictive step which uses the dynamics of the state-space model alone to

project one step into the future, while disregarding information contained in the next sample of the recorded signal. The measurement update equations are responsible for incorporating the next sample of the recorded signal into the a priori estimate to obtain an improved a posteriori estimate. This can be thought of as the correction step.

B.2 The Extended Kalman Filter

The extended Kalman filter (EKF) is a generalization of the Kalman filter to the case of a nonlinear state-space model which approximates the nonlinearity through local linearization. For the current application, since the phases of the harmonic components have a nonlinear relationship with the observed vocalization signal shown in (A.3), we had to use the EKF.

The filtered (measurement update) and predicted (time update) estimates are iteratively calculated using the Kalman recursions. The recursions are primed with an initialization step, which is followed by iterations over the time update and measurement update equations for each sample of the recorded signal:

- Initialization

$$\hat{\mathbf{x}}_{0|-1} = \begin{bmatrix} \hat{\mathbf{x}}_{1,0} \\ \hat{\mathbf{x}}_{2,0} \\ \hat{\mathbf{x}}_{3,0} \\ \vdots \\ \hat{\mathbf{x}}_{2m+1,0} \\ \hat{\mathbf{x}}_{2m+2,0} \end{bmatrix} \quad (\text{B.1})$$

$$Q = \begin{bmatrix} q_{1,0} & 0 & \cdots & 0 \\ 0 & q_{2,0} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & q_{2m+2,0} \end{bmatrix} \quad (\text{B.2})$$

$$P_{0|-1} = Q \quad (\text{B.3})$$

In (B.1), an a priori estimate of the initial state vector for the harmonic signal model is specified by the user. Note that $2m + 2$ is the total number of states in the system, where m is the number of harmonics specified for the signal model. In (B.2), the process noise covariance matrix is specified by the user, where q_i is the variance of the associated white noise process $\mathbf{u}_i$ as defined in (A.9). The fact that Q is a diagonal matrix reflects the assumption that the noise processes for each state are independent from one-another (e.g., the stochasticity of the time-varying fundamental frequency's amplitude is uncorrelated with the stochasticity of the harmonic amplitudes). In (B.3), an a priori estimate of the state error covariance matrix is specified by the user. This matrix provides a statistical estimate of the accuracy of the state estimates. For initialization, the state error covariance matrix is set to be equal to Q , indicating that there is initially more uncertainty about state estimates whose noise processes have

higher variances.

- Time-update equations:

$$F_{n-1} = \left. \frac{\partial f_{n-1}(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{n-1|n-1}} \quad (\text{B.4})$$

$$\hat{\mathbf{x}}_{n|n-1} = f(\hat{\mathbf{x}}_{n-1|n-1}) \quad (\text{B.5})$$

$$P_{n|n-1} = F_{n-1}P_{n-1|n-1}F_{n-1}^T + Q \quad (\text{B.6})$$

The time update equations make use of the filtered (i.e., corrected) state estimates from the previous time step to generate the predicted (i.e., uncorrected) state estimates for the current time step. In (B.4), the non-linear state transition function is linearized about the previous step's filtered state estimate $\hat{\mathbf{x}}_{n-1|n-1}$. See (B.14) below for the form of this linearization for the harmonic signal model. In (B.5), the state is projected forward one step in time by applying the non-linear state transition function $f(\cdot)$ to the previous step's filtered state estimate, generating the predicted estimate for the current time step $\hat{\mathbf{x}}_{n|n-1}$. In essence, the state-space model is “rolled ahead” one step in time using only the model's state transition function and no information obtained from the recorded signal. In (B.6), the error covariance estimate is projected forward one step in time using the linearized state transition function and the process noise covariance matrix Q to generate the a priori error covariance estimate $P_{n|n-1}$. Again, this matrix provides a statistical estimate of the accuracy of the state estimates at each time step.

- Measurement update equations:

$$H_n = \left. \frac{\partial h_n(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{n|n-1}} \quad (\text{B.7})$$

$$\hat{\mathbf{y}}_{n|n-1} = h(\hat{\mathbf{x}}_{n|n-1}) \quad (\text{B.8})$$

$$\tilde{\mathbf{y}}_n = \mathbf{y}_n - \hat{\mathbf{y}}_{n|n-1} \quad (\text{B.9})$$

$$r_{e,n} = r + H_n P_{n|n-1} H_n^T \quad (\text{B.10})$$

$$K_n = P_{n|n-1} H_n^T r_{e,n}^{-1} \quad (\text{B.11})$$

$$P_{n|n} = P_{n|n-1} - K_n r_{e,n} K_n^T \quad (\text{B.12})$$

$$\hat{\mathbf{x}}_{n|n} = \hat{\mathbf{x}}_{n|n-1} + K_n \tilde{\mathbf{y}}_n \quad (\text{B.13})$$

The measurement update equations correct the predicted state estimates generated in (B.5) by incorporating information gained from the estimation error of the current measurement. In (B.7), the non-linear measurement function is linearized about the predicted estimate. See (B.15) below for the form of this linearization for the harmonic signal model. In (B.8), the predicted state estimate is processed through the non-linear measurement function $h(\cdot)$ to produce the predicted measurement. This value indicates what we expect the current measurement to be in the absence of process and measurement noise and under the assumption that the state-space model is appropriate for the process generating the measured signal. In (B.9), the *residual* (sometimes called the *innovation*) is calculated. This value indicates the discrepancy between the predicted and actual signal values at the current time step. The residual is used to correct the predicted state estimates, generating the filtered state estimates. In (B.10), the covariance of the residuals at the current time step is estimated (this value is a matrix in general, but is scalar for the harmonic signal model where the measurements are a 1-dimensional signal). This value is used to calculate the Kalman gain in (B.11). The Kalman gain determines how much the residual

is weighted when correcting the predicted estimate $\hat{\mathbf{x}}_{n|n-1}$ to produce the filtered estimate $\hat{\mathbf{x}}_{n|n}$ in (B.13). It is computed such that the a posteriori error covariance calculated in (B.12) is minimized. One way of thinking about the weighting of the residual is that as the measurement covariance r approaches zero (e.g., a high signal to noise ratio for the recorded signal), the actual measurement $\mathbf{y}_n$ is “trusted” more and the predicted estimates $h(\hat{\mathbf{x}}_{n|n-1})$ are trusted less. Alternately, as the predicted error covariance $P_{n|n-1}$ approaches zero, the actual measurement is trusted less and the predicted measurement is trusted more. In this way, the EKF adaptively balances information gained from the dynamics of the state-space model and the recorded measurements to generate the state estimates.

The state transition function $f(\cdot)$ for the rectangular model (A.9) is linear, providing for a constant Jacobian given by:

$$F_n = \left. \frac{\partial f_n(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{n|n}} = \begin{bmatrix} 1 & 2\pi T_s & & \\ & \alpha & \mathbf{0} & \\ & & 1 & \\ \mathbf{0} & & & \ddots \\ & & & & 1 \end{bmatrix} \quad (\text{B.14})$$

The local linearization of the measurement function $h(\cdot)$ for the rectangular model (A.9) is given by:

$$H_n = \left. \frac{\partial h_n(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\hat{\mathbf{x}}_{n|n-1}} \quad (\text{B.15})$$

$$= \begin{bmatrix} \sum_{k=1}^m -k\mathbf{x}_{2+k,n} \sin(k\mathbf{x}_{1,n}) + k\mathbf{x}_{2+m+k,n} \cos(k\mathbf{x}_{1,n}) \\ 0 \\ \cos(\hat{\mathbf{x}}_{1,n|n-1}) \\ \cos(2\hat{\mathbf{x}}_{1,n|n-1}) \\ \vdots \\ \cos(m\hat{\mathbf{x}}_{1,n|n-1}) \\ \sin(\hat{\mathbf{x}}_{1,n|n-1}) \\ \sin(2\hat{\mathbf{x}}_{1,n|n-1}) \\ \vdots \\ \sin(m\hat{\mathbf{x}}_{1,n|n-1}) \end{bmatrix}$$

which must be evaluated at each time step n in (B.7).

B.3 The Extended Kalman Smoother

Unlike the EKF, which does not consider future signal values when estimating the current states, the fixed-interval Kalman smoother is a noncausal estimator that uses the entire record ($n = 0, \dots, N$) to estimate the *smoothed* states, $\hat{\mathbf{x}}_{n|N}$, of a linear stochastic process. The extended Kalman smoother (EKS), like the EKF, uses a local linear approximation of the transition and measurement functions in order to fit within the linear domain of the Kalman smoother recursions.

The application of the smoothing recursions can be thought of as an additional correction to the filtered estimates, $\hat{\mathbf{x}}_{n|n}$, generated by the EKF, in which information about future measurements are incorporated into the state estimates. This generates better state estimates in the mean squared error sense.

There are many mathematically equivalent expressions for the EKS. Here we use a variant similar to that developed in [Bryson and Frazier, 1963]. The first step

is to apply the EKF recursions and store all the filtered state estimates $\hat{\mathbf{x}}_{n|n}$, the Jacobians F_n and H_n , the predicted error covariance matrices $P_{n|n-1}$, and the residual covariances $r_{e,n}$. Next, the EKS recursions are performed using a backward recursive algorithm over the time sequence $n = N, \dots, 0$, following initialization of the *adjoint variable* $\boldsymbol{\psi}$ at time N

$$\boldsymbol{\psi}_{N+1|N} = 0.$$

One can think of $\boldsymbol{\psi}_{n|N}$ as a correction factor for $\hat{\mathbf{x}}_{n|n-1}$ calculated from all future samples $\mathbf{y}_{n+1}, \mathbf{y}_{n+2}, \dots, \mathbf{y}_N$.

Finally, the smoothed state estimates $\hat{\mathbf{x}}_{n|N}$ are computed using the following recursions:

$$\begin{aligned} K_{p,n} &= (F_n P_{n|n-1} H_n^T) r_{e,n}^{-1} \\ \boldsymbol{\psi}_{n|N} &= (F_n - K_{p,n} H_n)^T \boldsymbol{\psi}_{n+1|N} + H_n^T r_{e,n}^{-1} \mathbf{e}_n \\ \hat{\mathbf{x}}_{n|N} &= \hat{\mathbf{x}}_{n|n-1} + P_{n|n-1} \boldsymbol{\psi}_{n|N} \end{aligned}$$

B.4 Performance Assessment of EKS Analysis of Vocalizations

In order to assess the frequency tracking performance of the EKS and spectrographic frequency trackers on vocalization recordings, it was necessary to generate a synthetic “vocal-like” signal with known frequency components prior to the addition of measurement noise. In this way, the mean squared error (NMSE) performance of each method could be calculated across a range of signal to noise ratios (SNR). This type of analysis is impossible with actual recordings because the true frequencies are unknown, especially when the recordings are noisy.

It is important to note that the general stochastic state-based model described in

(A.1) can be used for both analysis and synthesis. In the context of the extended Kalman smoother, the state-based model is leveraged to estimate the states of the stochastic process given the measured signal. By defining initial conditions and the distribution of the process and measurement noise terms, however, one can also synthesize signals by drawing from the noise distributions at each time step and iterating through the state transition and measurement functions. In this fashion, one can generate an ensemble of random signals with identical statistical properties. One can also vary the measurement noise term to generate signals with different SNRs.

B.4.1 Synthetic Signal Generation

Fig. B.1 shows the spectrogram of a hummingbird vocalization. This was used as the basis of this statistical signal model. This spectrogram is plotted on a decibel scale with a range of 60 dBs to clearly show all 6 harmonics present in the signal.

The statistical model used to generate the vocal-like signals was an extension of the rectangular model discussed in Section A.2.2. Where the basic rectangular model describes a signal whose fundamental frequency has random deviations about a mean frequency, the hummingbird vocalization pictured in Fig. B.1 clearly has more structure than this, including quasi-periodic frequency modulations that are squashed at the top. To accommodate this structure in a stochastic vocal-like model, a second rectangular model was used to represent the modulation frequency. Model states corresponding to this modulation signal all include a subscript m .

As in (A.6), the instantaneous modulation frequency $\mathbf{f}_{m,n}$ is modeled as a first-order autoregressive (AR) process whose mean is $\bar{f}_m$:

$$\mathbf{f}_{m,n+1} = \alpha_m (\mathbf{f}_{m,n} - \bar{f}_m) + \bar{f}_m + \mathbf{u}_{f_{m,n}}$$

Figure B.1: Recorded hummingbird vocalization. Spectral power is plotted on a decibel scale with a range of 60 dBs to clearly show all 6 harmonics present in the signal.

The signal $y_{m,n}$ resulting from the synthesis process outlined in [A.2.5](#) is interpreted as a modulating frequency and added to the instantaneous frequency of the first rectangular model, resulting in composite frequency $\mathbf{f}_{M,n}$ with large amplitude modulations and (relatively) small and rapid deviations about these modulations:

$$\mathbf{f}_{M,n} = y_{m,n} + f_n$$

Frequencies that are greater than a specified threshold, $\mathbf{f}_t$, are squashed using a hyperbolic tangent function to generate the squashed modulated frequency $\mathbf{f}_{s,n}$:

$$\begin{aligned} \mathbf{f}_{s,n} &= \mathbf{f}_{M,n} && \text{when } \mathbf{f}_{M,n} \leq \mathbf{f}_t \\ &= \mathbf{f}_t + \gamma \tanh\left(\frac{\mathbf{f}_{M,n} - \mathbf{f}_t}{\gamma}\right) && \text{when } \mathbf{f}_{M,n} > \mathbf{f}_t \end{aligned}$$

Finally, a signal is generated using the squashed modulated frequency by following the synthesis method in [A.2.5](#).

B.4.2 Measurement Noise Models

While the Extended Kalman Smoother assumes that the measurement noise is white and Gaussian, natural signals may contain noise with different statistical characteristics. This opens up the question of how well the EKS algorithm performs on signals with different types of measurement noise. In particular, $1/f$ noise (also known as pink noise) is of interest due to its ubiquity in natural phenomena and its prevalence in natural auditory environments [[Keshner, 1982](#)]. Waterfalls, rustling leaves, and the ocean are some examples of natural processes that generate $1/f$ noise. Where white noise has uniform spectral power across a broad frequency range, $1/f$ noise has

a power spectral density that is proportional to the reciprocal of the frequency:

$$S(f) \propto 1/f$$

creating a spectral envelope that rolls off at approximately 3 dB per octave. Compared to white noise, $1/f$ noise has most of its power at low frequencies.

For the comparison of state-space and spectrographic frequency tracking algorithms, it is desired to apply these techniques to synthetic vocalizations in both white and $1/f$ noise. To make the frequency tracking comparison as equal as possible, the $1/f$ noise applied to each simulated vocalization was generated by filtering the same white noise sequence applied to each simulated vocalization. To accomplish this, each white noise sequence was generated with additional samples at the beginning to account for the transient required by the $1/f$ filter. An AR filter was then applied to generate the (approximate) $1/f$ noise, and the transient was trimmed from both sequences. This resulted in two simulated vocalizations: one in white measurement noise, and another in $1/f$ measurement noise.

B.4.3 Assessment of State-Space Tracking on Real Vocalizations

A mean squared error metric cannot be used to assess the tracking performance of the tracker on an actual recorded vocalization, y_n , because the true frequency components of this signal are unknown. Instead, residual analysis was performed. In this analysis, the tracker is first used to estimate the frequency, phase, and amplitude of the harmonic components of y_n . These estimates are then used to synthesize the signal by following the synthesis method in [A.2.5](#). The residuals, $\tilde{y}_n$, are given by the

difference between the original vocalization and the synthesized vocalization:

$$\tilde{\mathbf{y}}_n = \mathbf{y}_n - \hat{\mathbf{y}}_n$$

where $\hat{\mathbf{y}}_n$ is the synthesized signal. Good tracking of the states in the harmonic model would result in $\tilde{\mathbf{y}}_n$ approximating the measurement noise of the recording. In other words, everything has been captured by the harmonic model except for the random measurement noise. If white, Gaussian measurement noise is assumed, the “whiteness” of the residuals can be assessed by examining the spectrogram of $\tilde{\mathbf{y}}_n$ and looking for uniform power distribution in the frequency domain. Non-white sections of the residual spectrogram indicate periods where the harmonic model did a poor job of tracking the vocalization.

To provide a comparison of the state-space tracking results with the best practices in the field, a spectrographic frequency tracker was implemented. In our spectrographic approach, the dominant frequency components of the signal are found by following the peaks of the spectral estimates at each time sample in the spectrogram. This algorithm was chosen for its straightforward implementation, its limited number of tunable parameters, and its common usage in this field [Ohlemiller et al., 1994, DiMattina and Wang, 2006]. Many extensions and refinements to this algorithm exist for the frequency tracking problem for different signal classes and applications. They are all limited, however, by the time-frequency tradeoff inherent in the STFT calculation. Nonetheless, for the purpose of this experiment, care was taken to tune the existing algorithm parameters in order to optimize the performance on each of the signals.

The four primary tunable parameters for the spectrographic frequency tracking algorithm include:

- Spectrogram window type
- Spectrogram window length
- Number of samples in the frequency domain
- Frequency range for the peak detection algorithm

In each application of the algorithm, a Blackman window was used which minimizes the spectral leakage of the STFT estimates at the expense of main lobe width. This bias provided a good tradeoff for the spectral peak detection algorithm and increased the performance over using, for example, a rectangular window. The number of samples in the frequency domain was also fixed for each spectrogram calculation at 2^{11} by zero padding (a technique used to increase the number of samples in the frequency domain when windowing is used [[Manolakis et al., 2005](#)]). Sampling the estimated periodogram more frequently than this in the frequency domain did not provide better estimates of the spectral peaks and resulted in increased computation time and memory requirements. The frequency range for the peak detection algorithm was determined by inspecting the spectrogram of the signal and specifying a range that conservatively included the entirety of the fundamental frequency over the course of the signal.

The window length has the largest influence on the performance of the spectrographic frequency tracking algorithm and was tuned for each analyzed signal. When estimating the spectral content of a signal with the STFT, there is a time-frequency tradeoff to consider. Short windows provide for high resolution spectral estimates in the time domain, but poor resolution in the frequency domain. Conversely, long windows provide for high resolution estimates in the frequency domain, but poor resolution in the time domain. Since the frequency tracking algorithm only cares about the peaks of the spectral estimates at each time sample, short windows are

practical for signals with little or no noise and can result in accurate tracking of the spectral peaks. For signals that are increasingly noisy, however, longer windows are required in order to smooth over the effects of high amplitude noise contributions whose transient power eclipses that of the fundamental frequency component. This, however, reduces the time resolution of the algorithm making it impossible to track rapid changes in the spectral peaks.

B.5 Results

B.5.1 Comparison of EKS and Spectrographic Performance

Generating the Synthetic Signal Ensemble

An ensemble of stochastic synthetic signals was generated using the parameters specified in Tables B.1 and B.2. The values of r were chosen in order to vary the signal to noise ratio (SNR). SNR values are all stated on a decibel scale.

A typical stochastic synthetic signal with SNR equal to 5 is shown in Fig. B.2. Note that while the details of the temporal characteristics are clearly different than that of the original Hummingbird vocalization, the general characteristics, including the periodogram (pictured in the left axis), are very similar.

The SNR range for the frequency tracking experiment was specified to be 0 to 35 dB. For each SNR value tested, 1000 signals were generated in the absence of simulated measurement noise. For each of these signals, random white Gaussian noise was generated and scaled to produce the appropriate SNR. The same white Gaussian noise was filtered to approximate $1/f$ (pink) noise and again scaled to produce the appropriate SNR. For each of the 1000 noise-free signals generated at each SNR, this process resulted in two separate signals: one with white gaussian recording noise, and one with $1/f$ recording noise.

Table B.1: Model parameters for the rapidly changing frequency component of synthetic hummingbird vocalizations

Description	Symbol	Value
Duration		0.1 s
Number of Harmonics	m	5
Mean Frequency	$\bar{f}$	13 kHz
AR Coefficient	α	0.9999
Initial State Estimates		
Phase	$\hat{\theta}_0$	0 Radians
Frequency	$\hat{f}_0$	13 kHz
Cosine Amplitudes	$\hat{a}_{k,0}$	$[10^0 \ 10^{-1} \ \dots \ 10^{-5}]$
Sine Amplitudes	$\hat{b}_{k,0}$	$\hat{a}_{k,0}$
Noise Process Variances		
Measurement	r	Variable
Frequency	q_f	10^7
Cosine Amplitudes	q_a	$8 \cdot [10^1 \ 10^0 \ \dots \ 10^{-4}]$
Sine Amplitudes	q_b	q_a

Table B.2: Model parameters for the modulating frequency component of synthetic hummingbird vocalizations

Description	Symbol	Value
Mean Frequency	$\bar{f}^m$	1 kHz
AR Coefficient	α^m	0.999
Initial State Estimates		
Phase	$\hat{\theta}_0^m$	0 Radians
Frequency	$\hat{f}_0^m$	1 kHz
Cosine Amplitude	$\hat{a}_{k,0}^m$	1000
Sine Amplitude	$\hat{b}_{k,0}^m$	$\hat{a}_{k,0}^m$
Noise Process Variances		
Frequency	q_f^m	10^9
Cosine Amplitudes	q_a^m	$4 \cdot 10^6$
Sine Amplitudes	q_b^m	q_a^m

Figure B.2: One realization of a stochastic synthetic hummingbird vocalization. Spectral power is plotted on a decibel scale with a range of 60 dBs to clearly show all 6 harmonics present in the signal.

For each synthetic noisy vocalization, the state-space and spectrographic frequency tracking algorithms were used to estimate the fundamental frequency of the underlying harmonic processes. Each algorithm was optimized using the tunable parameters in order to maximize the performance, on average, for the SNR value of the signal. The tuning parameters are collected in Table B.3 and Table B.4. The variable parameters are all scaled logarithmically. For each simulation, the normalized mean squared error was calculated between the true fundamental frequency $\mathbf{f}_n$ and the estimated fundamental frequency $\hat{\mathbf{f}}_n$:

$$\text{NMSE} = \frac{1}{\sigma_{\mathbf{f}_n}^2 N} \sum_N (\mathbf{f}_n - \hat{\mathbf{f}}_n)^2 \quad (\text{B.16})$$

where $\sigma_{\mathbf{f}_n}^2$ is the variance of the true frequency. Because the performance metric has been normalized, any NMSE greater than 1 can be improved upon by estimating the fundamental frequency to be $\bar{\mathbf{f}}_n$ for each sample, indicating a poor estimate.

Fundamental Frequency Tracking Performance

Fig. B.3 displays the results of the comparison between the fundamental frequency tracking performance of the EKS and spectrographic frequency trackers when the simulated vocalizations are contained in white Gaussian measurement noise. The bold solid and dashed lines indicate the median NMSE values of the EKS and spectrographic frequency trackers, respectively. The shaded areas represent the 95% expectation range of the normalized tracking error over the ensemble for each method. Median was the chosen metric for this figure due to the presence of outliers resulting, typically, from signals where the synthetic vocalization amplitude drops to or near zero for a sustained period of time, making tracking particularly difficult for both methods. Across the whole SNR range considered, the EKS algorithm outperforms

Table B.3: User specified model parameters for tracking synthetic hummingbird vocalizations

Description	Symbol	Value
Number of Harmonics	m	5
Mean Frequency	$\bar{f}$	13.5 kHz
AR Coefficient	α	0.999
Initial State Estimates		
Phase	$\hat{\theta}_0$	0 Radians
Frequency	$\hat{f}_0$	15 kHz
Cosine Amplitudes	$\hat{a}_{k,0}$	$[10^0 \ 10^{-1} \ \dots \ 10^{-5}]$
Sine Amplitudes	$\hat{b}_{k,0}$	$\hat{a}_{k,0}$
Noise Process Variances		
Measurement	r	Variable
Frequency	q_f	Variable
Cosine Amplitudes	q_a	Variable
Sine Amplitudes	q_b	q_a

Table B.4: Variable parameters for comparison of harmonic and spectrographic performance test

Description	Start Value	End Value
Harmonic Tracking Algorithm Noise Process Variances		
Measurement	$2.5 \cdot 10^{-4}$	2.5
Frequency	10^7	$10^{8.5}$
Cosine Amplitudes	$[10^1 \ 10^0 \ \dots \ 10^{-4}]$	$[10^2 \ 10^1 \ \dots \ 10^{-3}]$
Sine Amplitudes	$[10^1 \ 10^0 \ \dots \ 10^{-4}]$	$[10^2 \ 10^1 \ \dots \ 10^{-3}]$
Spectrographic Tracking Algorithm		
Window Length	$3.3 \cdot 10^{-4} \text{ s}$	$1.3 \cdot 10^{-3} \text{ s}$

Figure B.3: Fundamental frequency tracking performance on an ensemble of 1000 synthetic hummingbird vocalizations for the spectrographic and EKS frequency tracking algorithms using white, Gaussian measurement noise. The bands indicate the upper and lower bounds of the 95% expectation range of the normalized tracking error over the ensemble.

the spectrographic algorithm at the task of accurately tracking the fundamental frequency of the synthetic signals. For an SNR equal to 0 dB (equal signal power and noise power), the EKS harmonic tracker has a median NMSE = 0.21, while the spectrographic tracker has a median NMSE = 1.97. For the entire range of SNRs, the EKS harmonic tracker has median NMSE values nearly an order of magnitude lower than the spectrographic tracker.

Fig. B.4 displays the results of the comparison between the fundamental frequency tracking performance of the EKS and spectrographic frequency trackers when the simulated vocalizations are contained in $1/f$ measurement noise. For low noise conditions (SNR > 20 dB) the performance of the two tracking methods is similar, as expected. For high noise scenarios, however, the EKS harmonic tracker performs better on the

Figure B.4: Fundamental frequency tracking performance on an ensemble of 1000 synthetic hummingbird vocalizations for the spectrographic and EKS frequency tracking algorithms using $1/f$ (pink) measurement noise. The bands indicate the upper and lower bounds of the 95% expectation range of the normalized tracking error over the ensemble.

$1/f$ measurement noise vocalizations than the white measurement noise vocalizations and the spectrographic tracker performs worse. For an SNR equal to 0 dB (equal signal power and noise power), the EKS harmonic tracker has a median NMSE = 0.13, while the spectrographic tracker has a median NMSE = 5.32. This result can be explained by the fact that $1/f$ noise with the same power as white noise (to maintain a constant SNR) has significantly more power in the lower frequencies. Since these low frequency components often eclipse the power of the fundamental frequency of the vocalization, the spectrographic tracker often tracks the $1/f$ noise instead of the fundamental frequency. The EKS harmonic tracker retains tracking on the fundamental in most of these cases, however, because abrupt shifts in frequency estimates are considered statistically improbable by the state-space model.

Fig. B.5 provides a closeup of the frequency estimates provided by these two methods in a high noise (SNR equals 0 dB) simulation using white Gaussian noise. The solid black line is the true fundamental frequency of the synthetic signal. The lighter solid line indicates the frequency estimates provided by the EKS tracker. The dashed line indicates the frequency estimates provided by the spectrographic tracker. As can be seen, the EKS harmonic tracker closely follows the contour of the true frequency of the signal. The spectrographic estimates, however, follow the peak of the spectrogram at each point in time, which leads to excessive variance in the estimates when there is significant power in the measurement noise. Spectrographic peak tracking is also susceptible to errors like those encountered at ~ 7.8 ms in Fig. B.5 where the signal power drops away momentarily and the tracker follows a component of the noise instead. The EKS harmonic tracker is not as susceptible to these events because they are improbable with respect to the statistical state-space model driving the frequency estimates.

B.5.2 Harmonic Tracking of Hummingbird Vocalization

The EKS harmonic tracker was used to estimate the frequencies, amplitudes, and phase offsets of the harmonic components of a recorded hummingbird vocalization. This is the same recording that is used as the basis for the synthetic signals described in Section B.4.1 and pictured in Fig. B.1. Table B.5 indicates the parameters used to initialize the EKS harmonic tracker for this signal. The power spectrogram on a linear scale covering the frequency range of the fundamental and the first two harmonics (for visibility) is pictured in Fig. B.6, along with the frequency estimates provided by the tracker. Fig. B.7 shows the estimated amplitude and phase offsets of the harmonic components of the signal. Fig. B.8 displays the spectrogram of the first three harmonics of the signal that was synthesized using the estimated frequencies, ampli-

Figure B.5: An example comparison of fundamental frequency tracking by the EKS and spectrographic methods on a noisy synthetic vocalization (SNR = 0 dB). Note how the EKS approach provides accurate estimates of the true frequency while the spectrographic approach jumps around excessively following the measurement noise.

Figure B.6: EKS harmonic frequency estimates for the hummingbird vocalization. Here, a linear spectral power scale was used and only the fundamental and the first two harmonics are pictured for clarity.

tudes, and phase offsets provided by the EKS harmonic tracker. Fig. B.9 displays the spectrogram of the residuals between the original recording and the synthesized recording. Note how there is virtually no residual spectral power, indicating both that the original vocalization had little measurement noise and that the signal model captured nearly all of the information about the original vocalization.

B.5.3 Morphing the Hummingbird Vocalization

By modifying the estimated model parameters prior to the synthesis stage, morphed vocalizations can be generated. Fig. B.10 shows the results of an example of several possible applications of this procedure to the analyzed hummingbird vocalization. Several modifications described in A.2.4 were performed to generate the resulting signals:

Figure B.7: EKS amplitude and phase offset estimates for the hummingbird vocalization. Each line ($k = 1$ through 6) corresponds to one of the harmonic components of the signal.

Figure B.8: The modeled hummingbird vocalization, synthesized using the state-space variables estimated during the analysis phase.

Table B.5: User specified model parameters for tracking humming bird vocalization

Description	Symbol	Value
Number of Harmonics	m	6
Mean Frequency	$\bar{f}$	12.75 kHz
AR Coefficient	α	0.999
Initial State Estimates		
Phase	$\hat{\theta}_0$	0 Radians
Frequency	$\hat{f}_0$	12.6 kHz
Cosine Amplitudes	$\hat{\mathbf{a}}_{k,0}$	$[0 \ 0 \ 0 \ 0 \ 0 \ 0]$
Sine Amplitudes	$\hat{\mathbf{b}}_{k,0}$	$\hat{\mathbf{a}}_{k,0}$
Noise Process Variances		
Measurement	r	0.0025
Frequency	q_f	$10^{8.5}$
Cosine Amplitudes	q_a	$[10^1 \ 10^0 \ \dots \ 10^{-4}]$
Sine Amplitudes	q_b	q_a

Figure B.9: The residuals between the original hummingbird vocalization and the modeled hummingbird vocalization.

- A. The original vocalization.
- B. The unmodified, synthesized vocalization.
- C. Time stretched by 20% without affecting the frequency.
- D. The fundamental frequency upshifted by 50% and the top 3 harmonics removed.
- E. The amplitude modulation removed.
- F. The frequency modulation removed.

These figures are plotted using the same frequency range and logarithmic spectral power scale as Fig. B.1 to make all of the harmonics visible. One should note that even though the model assumes a harmonic structure in the vocalization for the

parameter estimation phase, the synthesis phase can generate signals where the sinusoidal components aren't harmonically related. This is true of Fig. B.10.D, where the fundamental frequency is not harmonically related to the other components of the signal.

B.5.4 Harmonic Tracking of Human Vocalization

The EKS harmonic tracker was used to estimate the frequencies, amplitudes, and phase offsets of the harmonic components of the recorded human phrase ä-shä. This phrase was chosen due to the harmonically related ä component, with a broad band noise component in the middle. This phrase poses two challenges to the EKS harmonic tracker: 1) whether broad band noise can be analyzed and synthesized using a signal model which assumes a series of harmonically related signal components and 2) whether the EKS harmonic tracker can quickly regain track on the harmonically related section of the signal following the broad band noise. The spectrogram of the vocalization is shown in Fig. B.11. Table B.6 indicates the parameters used to initialize the EKS harmonic tracker for this signal. The frequency estimates provided by the tracker are pictured in Fig. B.12. Fig. B.13 displays the spectrogram of the signal that was synthesized using the estimated frequencies, amplitudes, and phase offsets provided by the harmonic tracker. Fig. B.14 displays the spectrogram of the residuals between the original vocalization and the synthesized vocalization. Note that the harmonic model accounts for nearly all of the power in the original signal, resulting in very little residual power. This indicates that the EKS harmonic tracker successfully analyzed and synthesized the broad band noise section of the vocalization and quickly regained track of the following harmonically related section.

Figure B.10: An example of a morphed hummingbird vocalization. Spectral power is plotted on a logarithmic scale to show all harmonics. A. The original vocalization. B. The unmodified, synthesized vocalization. C. Time stretched by 20% without affecting the frequency. D. The fundamental frequency upshifted by 50% and the top 3 harmonics removed. E. The amplitude modulation removed. F. The frequency modulation removed.

Table B.6: User specified model parameters for tracking human vocalization

Description	Symbol	Value
Number of Harmonics	m	40
Mean Frequency	$\bar{f}$	70 Hz
AR Coefficient	α	0.9993
Initial State Estimates		
Phase	$\hat{\theta}_0$	0 Radians
Frequency	$\hat{f}_0$	130 Hz
Cosine Amplitudes	$\hat{a}_{k,0}$	0
Sine Amplitudes	$\hat{b}_{k,0}$	$\hat{a}_{k,0}$
Noise Process Variances		
Measurement	r	0.01
Frequency	q_f	10^{10}
Cosine Amplitudes	q_a	$[10^6 \dots 10^6]$
Sine Amplitudes	q_b	q_a

Figure B.11: Recorded human phrase ä-shä. Spectral power is plotted on a linear scale.

Figure B.12: EKS harmonic frequency estimates for the human vocalization. 40 harmonics were used in the state-space harmonic model.

Figure B.13: The modeled human vocalization, synthesized using the state-space variables estimated during the analysis phase.

Figure B.14: The residuals between the original human vocalization and the modeled human vocalization.

B.5.5 Harmonic Tracking of Bat Vocalization

The EKS harmonic tracker was used to estimate the frequencies, amplitudes, and phase offsets of the harmonic components of a recorded bat vocalization. The spectrogram of the vocalization is pictured in Fig. B.15. This vocalization was chosen partly due to its harmonic structure, but also because there is a large amount of measurement noise present in the signal as a result of the recording environment. The artifacts present include recording noise, echo, and digital aliasing. The desire here was to use the EKS harmonic tracker to extract the harmonic content out of the noisy signal such that a clean signal can be synthesized that is void of the recording artifacts. Table B.7 lists the parameters used for this application. The frequency estimates provided by the tracker are pictured in Fig. B.16. Fig. B.17 displays the spectrogram of the signal that was synthesized using the estimated frequencies, amplitudes, and phase offsets provided by the EKS harmonic tracker. Note how the synthesized vocalization contains the harmonic content of the vocalization and effectively removes much of the signal power contained in the the recording artifacts. Fig. B.18 displays the spectrogram of the residuals between the original vocalization and the synthesized vocalization. While the residual process is approaching a white noise process, there are clearly some narrow-band components present. Some of the residuals are the very echoes and anti-aliasing effects that we were attempting to remove from the recording. Others, however, are artifacts that resulted from the autoregressive frequency tracking parameter, α , being set aggressively low. This is particularly evident where there is residual power present at the peaks of the frequency modulation resulting from setting α to bias the frequency estimates closer to the mean frequency, $\bar{f}$, of 25 kHz. This was necessary to keep the EKS algorithm from losing track in the presence of the noise artifacts and, in particular, the echo

Table B.7: User specified model parameters for tracking bat vocalization

Description	Symbol	Value
Number of Harmonics	m	5
Mean Frequency	$\bar{f}$	25 kHz
AR Coefficient	α	0.997
Initial State Estimates		
Phase	$\hat{\theta}_0$	0 Radians
Frequency	$\hat{f}_0$	25 kHz
Cosine Amplitudes	$\hat{a}_{k,0}$	[0.01...0.01]
Sine Amplitudes	$\hat{b}_{k,0}$	$\hat{a}_{k,0}$
Noise Process Variances		
Measurement	r	0.01
Frequency	q_f	$10^{8.5}$
Cosine Amplitudes	q_a	[100...100]
Sine Amplitudes	q_b	q_a

present in the room.

Figure B.15: Recorded bat vocalization. Spectral power is plotted on a linear scale.

Figure B.16: EKS harmonic frequency estimates for the bat vocalization. 5 harmonics were used in the state-space harmonic model.

Figure B.17: The modeled bat vocalization, synthesized using the state-space variables estimated during the analysis phase. Note how many of the recording artifacts present in Fig. B.15, including recording noise, echo, and digital aliasing, have been removed.

Figure B.18: The residuals between the original bat vocalization and the modeled bat vocalization.

Appendix C: Vocalization Analysis with the Particle Filter

C.1 Overview

As discussed in Appendix B, when the process model and measurement model of a state-space system are both linear, the *optimal* state estimate can be obtained using the Kalman filter [Kalman, 1960]. If the process model or measurement model is nonlinear, however, different estimation methods must be employed. Approaches to this problem such as the extended Kalman filter and extended Kalman smoother rely on estimates of the mean and covariance alone to characterize the marginal posterior distribution of the state given a sequence of observations, $p(\mathbf{x}_n|\mathbf{y}_{0:n})$ where $\mathbf{y}_{0:n} = \{\mathbf{y}_0, \dots, \mathbf{y}_n\}$. For some nonlinear processes the marginal posterior distribution is not well characterized by these statistics and may be multi-modal. In these cases, more complete descriptions of the posterior distribution are necessary in many estimation applications. This is true of the marginal posterior distribution of frequencies using the harmonic signal model specified in Appendix A.2.2, where the distribution of frequency states is multimodal with maxima at sub- and super-harmonics of not only the true fundamental frequency, but the harmonics as well. For harmonic signals with excessive noise, background artifacts, abrupt jumps, or periods of silence, it is possible for an EKS harmonic tracker to loose track and lock onto one of these sub- or super-harmonics, generating inaccurate estimates of the signal model parameters.

Monte Carlo methods are a possible alternative for estimating the states of such signals [Metropolis et al., 1953]. These methods can be used to estimate point statistics of an unknown or analytically unsolvable distribution up to a normalizing constant based on a sequence of sufficient random samples drawn from the distribution. Sequential Monte Carlo (SMC) methods apply this method sequentially to ease memory

and computational requirements of the algorithm.

In general these methods cannot be applied directly to state-space models because the posterior distribution $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$ is unknown and there is no means to draw independent samples from it. One can, however, draw the random samples from an *importance distribution*, $q(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$. Choice of the importance distribution may be arbitrary, as long as its support is larger than $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$. By appropriately weighting these random samples, the posterior distribution can be represented as a weighted combination of point masses. The expected value of any statistic of the state can then be estimated as a weighted average

$$\mathbb{E}[g(\mathbf{x}_{0:n})] \approx \frac{1}{N_p} \sum_{i=1}^{N_p} w_n^{(i)} g(\mathbf{x}_{0:n}^{(i)}) \quad (\text{C.1})$$

where N_p represents the total number of random samples, $w_n^{(i)}$ the importance weight of the i^{th} particle at time n , and $\mathbf{x}_{0:n}^{(i)}$ the i^{th} random sample drawn from the importance distribution $q(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$, and $g(\cdot)$ is an arbitrary function of the state that one wishes to estimate. The estimate converges as $\mathcal{O}(N_p^{-1})$ by the law of large numbers. The sequential Monte Carlo method is better known as a *particle filter* (PF) where a *particle* represents a random state trajectory, $\mathbf{x}_{0:n}^{(i)}$ [Djurić et al., 2003]. Many research groups have applied SMC methods to various applications such as fault detection [Hutter and Dearden, 2003, de Freitas, 2002], target localization [Jensfelt et al., 2000], computer vision [Isard and Blake, 1998], harmonic tracking for audio signals [Dubois and Davy, 2005], and speech recognition [Vermaak et al., 2002].

Asymptotically as the number of particles becomes large enough the approximation (C.1) converges to the true value. However, the computational requirements are proportional to the number of particles N_p . This is problematic because particle filters are plagued by two curses of dimensionality; the number of particles needed

for reasonable approximations scales with both *the state dimension* ℓ and *duration of the observation sequence* n at the time the state estimate is needed.

The marginal particle filter (MPF) in [Casella and Robert, 1996, Liu and Chen, 1998, Doucet et al., 2000] can partly overcome the first curse, in special cases. Suppose that we can partition the state vector,

$$\mathbf{x}_n = \begin{bmatrix} \mathbf{x}_n^L \\ \mathbf{x}_n^N \end{bmatrix} \quad (\text{C.2})$$

such that the system is linear when the nonlinear portion of the state is known. Here $\mathbf{x}_n^L$ and $\mathbf{x}_n^N$ represent the linear and nonlinear parts of the state, respectively. Then, the process can be modeled as

$$\mathbf{x}_{n+1}^L = F_n(\mathbf{x}_n^N)\mathbf{x}_n^L + \mathbf{u}_n^L \quad (\text{C.3})$$

$$\mathbf{x}_{n+1}^N = f_n(\mathbf{x}_n^N, \mathbf{u}_n^N) \quad (\text{C.4})$$

$$\mathbf{y}_n = H_n(\mathbf{x}_n^N)\mathbf{x}_n^L + \mathbf{v}_n \quad (\text{C.5})$$

where $F_n(\mathbf{x}_n^N)$ is a time-varying state-transition matrix and $H_n(\mathbf{x}_n^N)$ is an observation matrix, both of which are functions of the nonlinear portion of the state, $\mathbf{x}_n^N$. With this partition, the linear Kalman filter recursions can be used to *optimally* estimate the linear portion of the state for each particle, $\mathbf{x}_n^{L,(i)}$. This can dramatically reduce the number of particles needed to represent the remaining nonlinear portion of the posterior state distribution, $p(\mathbf{x}_{0:n}^N | \mathbf{y}_{0:n})$, since the dimension of the nonlinear portion of the state is less than the total state vector. If the linear process noise $\mathbf{u}_n^L$ and the measurement noise $\mathbf{v}_n$ are Gaussian, then the conditional linear posterior distribution

$p(\mathbf{x}_{0:n}^L | \mathbf{y}_{0:n}, \mathbf{x}_{0:n}^N)$ is also Gaussian. Then, the total posterior distribution is given by

$$p(\mathbf{x}_{0:n} | \mathbf{y}_{0:n}) = p(\mathbf{x}_{0:n}^L | \mathbf{y}_{0:n}, \mathbf{x}_{0:n}^N) p(\mathbf{x}_{0:n}^N | \mathbf{y}_{0:n}) \quad (\text{C.6})$$

This approach is sometimes called the *Rao-Blackwellized particle filter* or the *marginalized particle filter* (MPF) since the nonlinear portion of the state-space is estimated from the marginal distribution. The MPF leverages the traditional Kalman filter to optimally estimate the linear portion of the state and reduces the memory and computational requirements by reducing the dimension of the state vector that is estimated with a particle filter.

This, however, does not solve the second curse of dimensionality; the number of particles required to accurately estimate an expected value of the distribution still increases exponentially with the duration of the observations n at the time the estimate is needed. Practically this causes a phenomenon called *sample degeneracy*, where all but one of the importance weights $w_n^{(i)}$ become very small after a few iterations. This results in a large population of particles that contribute very little to the estimate in (C.1). Resampling schemes have been proposed to address this problem which essentially use a bootstrap approach to sample the posterior distribution with replacement. However, this approach generates many duplicate particles of the most probable state trajectories and results in less coverage of the set of all possible state trajectories. This is called *sample impoverishment*. If the true posterior distribution makes an abrupt change or develops a new prominent mode in a region not covered by the particle population, a typical PF may be slow to adapt to the abrupt change or fail to lock on to the new prominent mode.

Another problem is that estimates of expected values (C.1) do not always provide acceptable measures of multi-modal posterior distributions. For example, the mean

of the distribution may be an improbable state trajectory that is located between modes that correspond to probable state trajectories. In these cases an alternative, such as the median or most probable state trajectory, may be the preferred estimate.

This problem can be solved with a variation of particle filters developed for maximum *a posteriori* (MAP) estimation. These produce an estimate of the state trajectory that approximately maximizes the posterior distribution,

$$\hat{\mathbf{x}}_{0:n} \approx \arg \max_{\mathbf{x}_{0:n}} p(\mathbf{x}_{0:n} | \mathbf{y}_{0:n}) \quad (\text{C.7})$$

while the canonical PF produces an estimate of the mean state trajectory,

$$\hat{\mathbf{x}}_{0:n} \approx \mathbb{E}[\mathbf{x}_{0:n} | \mathbf{y}_{0:n}] \quad (\text{C.8})$$

MAP estimation is suitable for multi-modal distributions because it essentially selects the state estimates corresponding to the tallest mode of the distribution. MAP estimates minimize the most probable error, whereas mean estimates minimize mean square error (MSE).

It is difficult to obtain a MAP estimate from typical particle filters that represent the posterior as weighted point masses. The MAP estimate is not equivalent to choosing the state trajectory with the largest importance weight, because the weights are dependent on both the target distribution $p(\mathbf{x}_{0:n} | \mathbf{y}_{0:n})$ and the importance distribution $q(\mathbf{x}_{0:n} | \mathbf{y}_{0:n})$ that the particles are actually drawn from [Cappé et al., 2007]. However, a reasonable MAP estimator can be obtained by simply selecting the state trajectory with the largest posterior probability

$$\hat{\mathbf{x}}_{0:n} = \arg \max_i p(\mathbf{x}_{0:n}^{(i)} | \mathbf{y}_{0:n}) \quad (\text{C.9})$$

Both the sample degeneracy and sample impoverishment problems associated with the second curse of dimensionality can be addressed by combining this estimator with the Viterbi algorithm [Viterbi, 1967]. This approach increases the total number of candidate state trajectories exponentially with time n , though at a computational cost of N_p^2 per time step as compared to N_p for the typical particle filter. Since the MAP state trajectory may sequentially choose states from any particle in the population, even particles that are in unlikely regions of state-space may, in an instant, become relevant and contribute to the MAP sequence. This alleviates the need for particle resampling and allows the estimator to better adapt to abrupt changes in the posterior distribution. Thus, it does not suffer from either the sample degeneracy or sample impoverishment problems of standard particle filters.

Here we describe an optimal method and a fast approximate method for combining the advantages of marginalized particle filters (MPF) and MAP particle filters (MAM-PF) based on the Viterbi algorithm. These algorithms can be applied to any stochastic process that can be expressed as (C.3)–(C.5). Combining the MPF and MAM-PF methods introduces new problems that are not encountered in using the methods independently. We fully describe these problems and how they may be overcome.

C.2 Notation

We have adopted the notation used in [Cappé et al., 2007] with minor modifications. We use boldface to denote random processes, normal face for deterministic parameters and functions, upper case letters for matrices, lower case letters for vectors and scalars, superscripts in parenthesis for particle indices, upper-case superscripts for nonlinear/linear indication, and subscripts for time indices. For example, the nonlinear portion of the state vector for the i^{th} state trajectory (i.e., particle) is denoted as $\mathbf{x}_n^{\text{N},(i)}$ where n represents the discrete time index and (i) denotes the i^{th} particle.

The unnormalized particle weights are denoted as $\tilde{w}^{(i)}$ and the normalized particle weights as $w^{(i)}$. The state trajectories before resampling are denoted as $\tilde{\mathbf{x}}_n^{(i)}$ and as $\mathbf{x}_n^{(i)}$ after resampling.

C.3 Recursions for Typical Particle Filtering Methods

The newly proposed MAM-PF is a combination of the MAP and marginalized PFs with an additional adaptive measurement noise variance estimation procedure. In order to clarify the algorithm of this proposed hybrid, we first provide pseudo-code to describe each of the PF algorithms in detail so that the integration of marginalization and MAP estimation can be presented more clearly.

C.3.1 Standard Resampling Particle Filter (PF)

The standard PF algorithm, Algorithm 1, is given below. This algorithm incorporates the stratified resampling scheme in which resampling is performed when the estimated number of effective particles drops below a user-defined threshold, N_t [Kitagawa, 1996]. The number of effective particles is given by

$$\hat{N}_e = \frac{1}{\sum_{i=1}^{N_p} \left(w_n^{(i)}\right)^2}$$

where $w_n^{(i)}$ is a normalized importance weight of the i^{th} particle and N_p is the number of particles. The normalized importance weights w_n are calculated at each time index from the unnormalized importance weights $\tilde{w}_n$, which are themselves derived from the normalized importance weights from the previous time index. In Algorithm 1, N_T refers to the number of observed samples. Here we assume that the importance

density has been chosen such that it can be factored as

$$q(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) = q_n(\mathbf{x}_n|\mathbf{x}_{0:n-1}, \mathbf{y}_{0:n})q(\mathbf{x}_{0:n-1}|\mathbf{y}_{0:n-1}) \quad (\text{C.10})$$

so that the particle weights can be calculated recursively.

Then, the weight update recursion can be written as,

$$\tilde{\mathbf{w}}_n^{(i)} = \mathbf{w}_{n-1}^{(i)} \frac{p(\mathbf{y}_n|\mathbf{x}_n^{(i)})p(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(i)})}{q_n(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(i)}, \mathbf{y}_{0:n})} \quad (\text{C.11})$$

where $\tilde{\mathbf{w}}_n^{(i)}$ represents an unnormalized importance weight at time n and $\mathbf{w}_{n-1}^{(i)}$ a normalized importance weight at time $n-1$. The most common choice for the marginal importance density is the prior probability given by the process model,

$$q_n(\mathbf{x}_n|\mathbf{x}_{0:n-1}, \mathbf{y}_{0:n}) = p(\mathbf{x}_n|\mathbf{x}_{n-1}). \quad (\text{C.12})$$

The weight update recursion, then, is simplified as follows,

$$\begin{aligned} \tilde{\mathbf{w}}_n^{(i)} &= \mathbf{w}_{n-1}^{(i)} \frac{p(\mathbf{y}_n|\mathbf{x}_n^{(i)})p(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(i)})}{q_n(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(i)}, \mathbf{y}_{0:n})} \\ &= \mathbf{w}_{n-1}^{(i)} p(\mathbf{y}_n|\mathbf{x}_n^{(i)}) \end{aligned} \quad (\text{C.13})$$

where the current weight is the previous weight multiplied by the likelihood function, $p(\mathbf{y}_n|\mathbf{x}_n^{(i)})$.

C.3.2 Maximum A Posteriori Particle Filter (MAP-PF)

The MAP-PF approach based on the Viterbi algorithm provides a MAP estimate that avoids the sample degeneracy and sample impoverishment problems of the standard PF [Godsill et al., 2001, Cappé et al., 2007, Viterbi, 1967]. The algo-

Algorithm 1 Standard Resampling Particle Filter (PF)

Initialization

```
for  $i = 1, \dots, N_p$  do
  Sample  $\tilde{\mathbf{x}}_0^{(i)} \sim \pi_0(\tilde{\mathbf{x}}_0|\mathbf{y}_0)$ 
  Calculate initial importance weights:
     $\tilde{w}_0^{(i)} = \pi_0(\mathbf{x}_0)p(\mathbf{y}_0|\tilde{\mathbf{x}}_0^{(i)})$ 
end for
for  $i = 1, \dots, N_p$  do
   $w_0^{(i)} = \tilde{w}_0^{(i)} / \sum_j \tilde{w}_0^{(j)}$ 
end for
```

Sequential Importance Sampling

```
for  $n = 1, \dots, N_T$  do
  if  $N_e \geq N_t$  then
    for  $i = 1, \dots, N_p$  do
       $\mathbf{x}_{n-1}^{(i)} = \tilde{\mathbf{x}}_{n-1}^{(i)}$ 
    end for
  else
    Set  $k = 1$  and  $\tau = 1/N_p$ 
    for  $i = 1, \dots, N_p$  do
      while  $w_{n-1}^{(k)} \leq \tau$  do
         $k = k + 1$ 
      end while
       $\mathbf{x}_{n-1}^{(i)} = \tilde{\mathbf{x}}_{n-1}^{(k)}$ 
       $w_{n-1}^{(i)} = 1/N$ 
       $\tau = \tau + 1/N$ 
    end for
  end if
  for  $i = 1, \dots, N_p$  do
     $\tilde{\mathbf{x}}_n^{(i)} \sim q_n(\tilde{\mathbf{x}}_n^{(i)}|\mathbf{x}_{n-1}^{(i)}, \mathbf{y}_{0:n})$ 
     $\tilde{w}_n^{(i)} = w_{n-1}^{(i)} \frac{p(\mathbf{y}_n|\tilde{\mathbf{x}}_n^{(i)})p(\tilde{\mathbf{x}}_n^{(i)}|\mathbf{x}_{n-1}^{(i)})}{q_n(\tilde{\mathbf{x}}_n^{(i)}|\mathbf{x}_{n-1}^{(i)}, \mathbf{y}_{0:n})}$ 
  end for
  for  $i = 1, \dots, N_p$  do
     $w_n^{(i)} = \frac{\tilde{w}_n^{(i)}}{\sum_j \tilde{w}_n^{(j)}}$ 
  end for
end for
```

Algorithm 2 MAP Particle Filter (MAP-PF)

Initialization

```
for  $i = 1, \dots, N_p$  do
  Sample  $\mathbf{x}_0^{(i)} \sim \pi_0(\mathbf{x}_0)$ 
   $\alpha_0^{(i)} = \pi_0(\mathbf{x}_0^{(i)})p(\mathbf{y}_0|\mathbf{x}_0^{(i)})$ 
   $\mathbf{z}_0^{(i)} = \mathbf{x}_0^{(i)}$ 
end for
 $i^* = \underset{i}{\operatorname{argmax}} \alpha_0^{(i)}$ 
 $\hat{\mathbf{x}}_0 = \mathbf{x}_0^{(i^*)}$ 
```

Sequential MAP Estimation

```
for  $n = 1, \dots, N_T$  do
  for  $i = 1, \dots, N_p$  do
     $\mathbf{x}_n^{(i)} \sim q_n(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(i)}, \mathbf{y}_{0:n})$ 
  end for
  for  $i = 1, \dots, N_p$  do
     $k^* = \underset{k}{\operatorname{argmax}} \alpha_{n-1}^{(k)}p(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(k)})$ 
     $\mathbf{z}_{0:n}^{(i)} = \left\{ \mathbf{z}_{0:n-1}^{(k^*)}, \mathbf{x}_n^{(i)} \right\}$ 
     $\alpha_n^{(i)} = \alpha_{n-1}^{(k^*)}p(\mathbf{x}_n^{(i)}|\mathbf{x}_{n-1}^{(k^*)})p(\mathbf{y}_n|\mathbf{x}_n^{(i)})$ 
  end for
   $i^* = \underset{i}{\operatorname{argmax}} \alpha_n^{(i)}$ 
   $\hat{\mathbf{x}}_{0:n} = \mathbf{z}_{0:n}^{(i^*)}$ 
end for
```

rithm is memory-efficient because at each sample time the Viterbi algorithm discards $(N_p^2 - N_p)$ possible trajectories and only retains the N_p most probable ones. The algorithm for the MAP-PF is given in Algorithm 2.

Unlike the standard particle filter (PF), the MAP-PF does not calculate or track importance weights for each particle. These are unnecessary because the MAP-PF does not estimate the mean or other moments of the state posterior distribution, as in (C.1). Instead the MAP-PF simply tracks the posterior probability of each state trajectory. These are represented by the coefficients $\alpha_n^{(i)}$, which are computed recursively for each particle instead of the importance weights.

C.3.3 The Marginalized Particle Filter (MPF)

The marginalized particle filter (MPF) can be applied to special state-space models in which a portion of the state-space is nonlinear and the other portion can be modeled as a linear process if the nonlinear portion of the state vector is known. When this partition can be performed, the linear portion of the state can be sequentially estimated using the Kalman filter, and particle filtering can be used to estimate the nonlinear portion of the state vector. The MPF reduces the variance of the posterior distribution estimation by providing *optimal* estimates for the linear portion of the state-space while reducing the dimensionality of the nonlinear portion of the state that is estimated with a particle filter.

The MPF recursions are listed in Algorithm 3 for the state model in (C.3)-(C.5). The state update probability $p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(i)})$ is obtained from the nonlinear state update model in (C.4).

Algorithm 3 Marginalized Particle Filter (MPF)

Initialization

for $i = 1, \dots, N_p$ **do**
 Sample $\mathbf{x}_0^{N,(i)} \sim \pi_0(\mathbf{x}_0^N)$
 $\hat{\mathbf{x}}_{0:-1}^{L,(i)} = E[\mathbf{x}_0^{L,(i)} | \mathbf{x}_0^{N,(i)}]$
 Initial Weight Calculation:
 $\tilde{w}_0^{(i)} = p(\mathbf{y}_0 | \mathbf{x}_0^{N,(i)}, \mathbf{x}_0^L)$
end for

Marginalized Sequential Estimation

for $n = 0, \dots, N_T$ **do**
 if $N_e \leq N_t$ **then**
 Resample (see Algorithm 1)
 end if
 for $i = 1, \dots, N_p$ **do**
 Particle Propagation
 $\mathbf{x}_n^{N,(i)} \sim q_n(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(i)}, \mathbf{y}_{0:n})$
 Kalman Filtering
 Measurement Update:
 $\mathbf{R}_{e,n}^{(i)} = H_n(\mathbf{x}_n^{N,(i)}) \mathbf{C}_{n|0:n-1}^{(i)} H_n(\mathbf{x}_n^{N,(i)})^T + R_v$
 $\mathbf{K}_n = \mathbf{C}_{n|0:n-1} H_n(\mathbf{x}_n^{N,(i)})^T (\mathbf{R}_{e,n}^{(i)})^{-1}$
 $\hat{\mathbf{y}}_{n|0:n-1}^{(i)} = H_n(\mathbf{x}_n^{N,(i)}) \hat{\mathbf{x}}_{n|0:n-1}^{L,(i)}$
 $\hat{\mathbf{x}}_{n|0:n}^{L,(i)} = \hat{\mathbf{x}}_{n|0:n-1}^{L,(i)} + \mathbf{K}_n [\mathbf{y}_n - \hat{\mathbf{y}}_{n|0:n-1}^{(i)}]$
 $\mathbf{C}_{n|0:n} = [I - \mathbf{K}_n H_n(\mathbf{x}_n^{N,(i)})] \mathbf{C}_{n|0:n-1}^{(i)}$
 Time Update:
 $\hat{\mathbf{x}}_{n+1|0:n}^{L,(i)} = F_n(\mathbf{x}_n^{N,(i)}) \hat{\mathbf{x}}_{n|0:n}^{L,(i)}$
 $\mathbf{C}_{n+1|0:n}^{(i)} = F_n(\mathbf{x}_n^{N,(i)}) \mathbf{C}_{n|0:n}^{(i)} F_n(\mathbf{x}_n^{N,(i)})^T + \mathbf{Q}_u^L$
 Weight Update
 $p(\mathbf{y}_n | \mathbf{y}_{0:n-1}, \mathbf{x}_{0:n}^{N,(i)}) \sim \mathcal{N}(\hat{\mathbf{y}}_{n|0:n-1}^{(i)}, \mathbf{R}_{e,n}^{(i)})$
 $\tilde{w}_n^{(i)} = w_{n-1}^{(i)} \frac{p(\mathbf{y}_n^{(i)} | \mathbf{y}_{0:n-1}, \mathbf{x}_n^{N,(i)}) p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(i)})}{q_n(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{0:n-1}^{N,(i)}, \mathbf{y}_{0:n})}$
 end for
 for $i = 1, \dots, N_p$ **do**
 $w_n^{(i)} = \frac{\tilde{w}_n^{(i)}}{\sum_j \tilde{w}_n^{(j)}}$
 end for
end for

Algorithm 4 Optimal MAM-PF (Part 1 of 2)

Initialization

```
for  $i = 1, \dots, N_p$  do
  Sample  $\mathbf{x}_0^{N,(i)} \sim \pi_0(\mathbf{x}_0^N)$ 
   $\hat{\mathbf{x}}_{0:-1}^{L,(i)} = \mathbb{E} \left[ \mathbf{x}_0^{L,(i)} | \mathbf{x}_0^{N,(i)} \right]$ 
end for
for  $i = 1, \dots, N_p$  do
   $\alpha_0^{(i)} = \pi_0(\mathbf{x}_0^{N,(i)}) p(\mathbf{y}_0 | \mathbf{x}_0^{N,(i)}, \mathbf{x}_0^{L,(i)})$ 
   $\mathbf{z}_0^{(i)} = \mathbf{x}_0^{(i)}$ 
end for
 $i^* = \operatorname{argmax} \alpha_0^{(i)}$ 
 $\hat{\mathbf{x}}_0 = \mathbf{x}_0^{i^*}$ 
```

C.4 Optimal MAP Adaptive Marginalized Particle Filter (Optimal MAM-PF)

One can apply the Viterbi algorithm to obtain the MAP state trajectory within the canonical particle filter framework [Cappé et al., 2007, Godsill et al., 2001]. However, the Viterbi algorithm does not guarantee the true MAP state trajectory when the state is marginalized [Cemgil and Kappen, 2003]. Estimation of the MAP state trajectory with the marginalized state has not been previously described.

The MAM-PF is a hybrid particle filtering method which leverages the advantages of the MAP-PF and MPF algorithms. The MAP-PF portion of the algorithm permits the particles to densely cover the nonlinear portion of the state-space. While this prevents the sample impoverishment problem that would normally be caused by resampling, it also means that the likelihood function $p(\mathbf{y}_n | \mathbf{y}_{0:n-1}, \mathbf{x}_n^{N,(i)})$ must be evaluated for particles $\mathbf{x}_n^{N,(i)}$ whose values may be far away from probable values. In this case some of the variation in $\mathbf{y}_n$ caused by the true underlying state would not be accounted for and the residual variance $\mathbf{y}_n - \hat{\mathbf{y}}_{n|0:n-1}$ will be much larger than the measurement noise $R_{\mathbf{v},n}$ or the prediction error $R_{\mathbf{e},n}$ provided by the Kalman filter

Algorithm 5 Optimal MAM-PF (Part 2 of 2)

```

for  $n = 1, \dots, N_T$  do
  for  $i = 1, \dots, N_p$  do
    Particle Propagation
     $\mathbf{x}_n^{N,(i)} \sim q_n(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(i)}, \mathbf{y}_n)$ 
    Marginalized Sequential Estimation
    for  $k = 1, \dots, N_p$  do
      Measurement Update
       $\hat{\mathbf{y}}_{n|0:n-1} = H_n(\mathbf{x}_n^{N,(i)}) \hat{\mathbf{x}}_{n|0:n-1}^{L,(k)}$ 
       $\mathbf{e}_n = \mathbf{y}_n - \hat{\mathbf{y}}_{n|0:n-1}$ 
       $\mathbf{R}_{v,n} = \left[ \mathbf{e}_n \mathbf{e}_n^T - H_n(\mathbf{x}_n^{N,(i)}) \mathbf{C}_{n|0:n-1}^{(k)} H_n(\mathbf{x}_n^{N,(i)})^T \right]_+$ 
       $\hat{\mathbf{R}}_{v,n}^{(i,k)} = \beta \hat{\mathbf{R}}_{v,n-1}^{(k)} + (1 - \beta) \mathbf{R}_{v,n}$ 
       $\mathbf{R}_{e,n}^{(k)} = H_n(\mathbf{x}_n^{N,(i)}) \mathbf{C}_{n|0:n-1}^{(k)} H_n(\mathbf{x}_n^{N,(i)})^T + \hat{\mathbf{R}}_{v,n}^{(k)}$ 
       $\mathbf{K}_n^{(k)} = \mathbf{C}_{n|0:n-1}^{(k)} H_n(\mathbf{x}_n^{N,(i)})^T (\mathbf{R}_{e,n}^{(k)})^{-1}$ 
       $\hat{\mathbf{x}}_{n|0:n}^{L,(k)} = \hat{\mathbf{x}}_{n|0:n-1}^{L,(k)} + \mathbf{K}_n^{(k)} [\mathbf{y}_n - \hat{\mathbf{y}}_{n|0:n-1}^{(k)}]$ 
       $\mathbf{C}_{n|0:n}^{(k)} = \left[ I - \mathbf{K}_n^{(k)} H_n(\mathbf{x}_n^{N,(i)}) \right] \mathbf{C}_{n|0:n-1}^{(k)}$ 
      Time Update
       $\hat{\mathbf{x}}_{n+1|0:n}^{L,(i,k)} = F_n(\mathbf{x}_n^{N,(i)}) \hat{\mathbf{x}}_{n|0:n}^{L,(k)}$ 
       $\mathbf{C}_{n+1|0:n}^{(i,k)} = F_n(\mathbf{x}_n^{N,(i)}) \mathbf{C}_{n|0:n}^{(k)} F_n(\mathbf{x}_n^{N,(i)})^T + \mathbf{Q}_u^L$ 
    end for
    MAP Estimation
     $k^* = \underset{k}{\operatorname{argmax}} \alpha_{n-1}^{(k)} p(\mathbf{y}_n | \mathbf{x}_n^{N,(i)}, \hat{\mathbf{x}}_{n|0:n-1}^{L,(k)}) p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(k)})$ 
     $\alpha_n^{(i)} = \alpha_{n-1}^{(k^*)} p(\mathbf{y}_n | \mathbf{x}_n^{N,(i)}, \hat{\mathbf{x}}_{n|0:n-1}^{L,(k^*)}) p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(k^*)})$ 
     $\mathbf{C}_{n+1|0:n}^{(i)} = \mathbf{C}_{n+1|0:n}^{(i,k^*)} \ \& \ \hat{\mathbf{R}}_{v,n}^{(i)} = \hat{\mathbf{R}}_{v,n}^{(i,k^*)}$ 
     $\hat{\mathbf{x}}_{n+1|0:n}^{L,(i)} = \hat{\mathbf{x}}_{n+1|0:n}^{L,(i,k^*)}$ 
     $\hat{\mathbf{x}}_n^{(i)} = \left[ \hat{\mathbf{x}}_{n|0:n}^{L,(k^*)}, \mathbf{x}_n^{N,(i)} \right]^T$ 
     $\mathbf{z}_{0:n}^{(i)} = \left[ \mathbf{z}_{0:n-1}^{(k^*)}, \hat{\mathbf{x}}_n^{(i)} \right]$ 
  end for
  Update MAP State Estimate
   $i^* = \underset{i}{\operatorname{argmax}} \alpha_n^{(i)}$ 
   $\hat{\mathbf{x}}_{0:n} = \mathbf{z}_{0:n}^{i^*}$ 
end for

```

recursions. This underestimation of the prediction error causes the likelihood function to have a distribution that is too narrow, which in turn distorts the posterior distribution and ultimately leads to suboptimal particle selection. This is a critical problem that has not been addressed previously and only occurs when attempting to use both marginalization and MAP estimation with the Viterbi algorithm.

One elegant solution to this problem is to continuously estimate the prediction error covariance from the residuals for each particle. The particle filter algorithms described below use an adaptive covariance estimation method first proposed in [Myers and Tapley, 1976]. To ensure that the estimated covariance matrix is positive semi-definite, an eigenvalue decomposition of the covariance matrix is performed and all non-positive eigenvalues are eliminated. We denote this operation as $[\mathbf{R}]_+$.

A second critical issue that occurs in merging marginalization and MAP estimation is that the likelihood function $p(\mathbf{y}_n | \mathbf{y}_{0:n-1}, \mathbf{x}_n^{\text{N},(i)})$ must be handled carefully because it is only conditioned on the nonlinear portion of the state vector. This distribution can be obtained from the Kalman filter recursions, as was done for the MPF

$$p(\mathbf{y}_n | \mathbf{y}_{0:n-1}, \mathbf{x}_n^{\text{N},(i)}) = p(\mathbf{y}_n | \hat{\mathbf{x}}_{n|0:n-1}^{\text{L},(i)}, \mathbf{x}_n^{\text{N},(i)}) \quad (\text{C.14})$$

$$\sim \mathcal{N}(\hat{\mathbf{y}}_{n|0:n-1}^{(i)}, \mathbf{R}_{\mathbf{e},n}^{(i)}) \quad (\text{C.15})$$

However, during the maximization over all past trajectories it is crucial to recognize that this includes the linear portion of the state-space, $\hat{\mathbf{x}}_{n|0:n-1}^{\text{L},(i)}$, unlike the MAP-PF.

Algorithm 4 gives a complete account of the Optimal MAM-PF recursions. The covariance coefficient β is a user-specified parameter that controls the memory of the recursion for first order recursive estimation of the adaptive signal prediction error covariance, $\mathbf{R}_{\mathbf{e},n}$.

C.5 Fast MAP Adaptive Marginalized Particle Filter (Fast MAM-PF)

Algorithm 6 Fast MAM-PF (Part 1 of 2)

Initialization

```

for  $i = 1, \dots, N_p$  do
  Sample  $\mathbf{x}_0^{N,(i)} \sim \pi_0(\mathbf{x}_0^N)$ 
   $\hat{\mathbf{x}}_{0:-1}^{L,(i)} = \mathbb{E} \left[ \mathbf{x}_0^{L,(i)} | \mathbf{x}_0^{N,(i)} \right]$ 
end for
for  $i = 1, \dots, N_p$  do
   $\alpha_0^{(i)} = \pi_0(\mathbf{x}_0^{N,(i)}) p(\mathbf{y}_0 | \mathbf{x}_0^{N,(i)}, \mathbf{x}_0^{L,(i)})$ 
   $\mathbf{z}_0^{(i)} = \mathbf{x}_0^{(i)}$ 
end for
 $i^* = \operatorname{argmax} \alpha_0^{(i)}$ 
 $\hat{\mathbf{x}}_0 = \mathbf{x}_0^{i^*}$ 

```

A key computational disadvantage of the Optimal MAM-PF algorithm is that the linear Kalman filter recursions must be applied N_p times for each particle, which results in N_p^2 Kalman filter recursions for each time update of the state estimate. This ensures that the maximization over all possible previous trajectories correctly accounts for the effect of the linear state estimates on the likelihood function, which is given by $p(\mathbf{y}_n | \mathbf{x}_n^{N,(i)}, \hat{\mathbf{x}}_{n|0:n-1}^{L,(k)})$ for the i^{th} particle. In most cases the likelihood function does not strongly affect the selection of the previous trajectory and this term can therefore be eliminated from the MAP estimation step. Specifically,

$$k^* = \operatorname{argmax}_k \alpha_{n-1}^{(k)} p(\mathbf{y}_n | \mathbf{x}_n^{N,(i)}, \hat{\mathbf{x}}_{n|0:n-1}^{L,(k)}) p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(k)}) \quad (\text{C.16})$$

$$\approx \operatorname{argmax}_k \alpha_{n-1}^{(k)} p(\mathbf{y}_n | \mathbf{x}_n^{N,(i)}, \hat{\mathbf{x}}_{n|0:n-1}^{L,(i)}) p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(k)}) \quad (\text{C.17})$$

$$= \operatorname{argmax}_k \alpha_{n-1}^{(k)} p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(k)}) \quad (\text{C.18})$$

This approximation sacrifices the asymptotic optimality of the Optimal MAM-PF,

Algorithm 7 Fast MAM-PF (Part 2 of 2)

```

for  $n = 1, \dots, N_T$  do
  for  $i = 1, \dots, N_p$  do
    MAP Estimation
    Particle Propagation
     $\mathbf{x}_n^{N,(i)} \sim q_n(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(i)}, \mathbf{y}_n)$ 
     $k^* = \underset{k}{\operatorname{argmax}} \alpha_{n-1}^{(k)} p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(k)})$ 
    Marginalized Sequential Estimation
    Measurement Update
     $\hat{\mathbf{y}}_{n|0:n-1} = H_n(\mathbf{x}_n^{N,(i)}) \hat{\mathbf{x}}_{n|0:n-1}^{L,(k^*)}$ 
     $\mathbf{e}_n = \mathbf{y}_n - \hat{\mathbf{y}}_{n|0:n-1}$ 
     $\mathbf{R}_{v,n} = \left[ \mathbf{e}_n \mathbf{e}_n^T - H_n(\mathbf{x}_n^{N,(i)}) \mathbf{C}_{n|0:n-1}^{(k^*)} H_n(\mathbf{x}_n^{N,(i)})^T \right]_+$ 
     $\hat{\mathbf{R}}_{v,n} = \beta \hat{\mathbf{R}}_{v,n-1}^{(k^*)} + (1 - \beta) \mathbf{R}_{v,n}$ 
     $\mathbf{R}_{e,n} = H_n(\mathbf{x}_n^{N,(i)}) \mathbf{C}_{n|0:n-1}^{(k^*)} H_n(\mathbf{x}_n^{N,(i)})^T + \hat{\mathbf{R}}_{v,n}$ 
     $\mathbf{K}_n = \mathbf{C}_{n|0:n-1}^{(k^*)} H_n(\mathbf{x}_n^{N,(i)})^T (\mathbf{R}_{e,n})^{-1}$ 
     $\hat{\mathbf{x}}_{n|0:n}^L = \hat{\mathbf{x}}_{n|0:n-1}^{L,(k^*)} + \mathbf{K}_n \mathbf{e}_n$ 
     $\mathbf{C}_{n|0:n} = \left[ \mathbf{I} - \mathbf{K}_n H_n(\mathbf{x}_n^{N,(i)}) \right] \mathbf{C}_{n|0:n-1}^{(k^*)}$ 
    Time Update
     $\hat{\mathbf{R}}_{v,n}^{(i)} = \hat{\mathbf{R}}_{v,n}$ 
     $\hat{\mathbf{x}}_{n+1|0:n}^{L,(i)} = F_n(\mathbf{x}_n^{N,(i)}) \hat{\mathbf{x}}_{n|0:n}^L$ 
     $\mathbf{C}_{n+1|0:n}^{(i)} = F_n(\mathbf{x}_n^{N,(i)}) \mathbf{C}_{n|0:n} F_n(\mathbf{x}_n^{N,(i)})^T + \mathbf{Q}_u^L$ 
     $\alpha_n^{(i)} = \alpha_{n-1}^{(k^*)} p(\mathbf{y}_n | \mathbf{x}_n^{N,(i)}, \hat{\mathbf{x}}_{n|0:n-1}^{L,(k^*)}) p(\mathbf{x}_n^{N,(i)} | \mathbf{x}_{n-1}^{N,(k^*)})$ 
     $\hat{\mathbf{x}}_n^{(i)} = \left[ \hat{\mathbf{x}}_{n|0:n}^L, \mathbf{x}_n^{N,(i)} \right]^T$ 
     $\mathbf{z}_{0:n}^{(i)} = \left[ \mathbf{z}_{0:n-1}^{(k^*)}, \hat{\mathbf{x}}_n^{(i)} \right]$ 
  end for
  Update MAP State Estimate
   $i^* = \underset{i}{\operatorname{argmax}} \alpha_n^{(i)}$ 
   $\hat{\mathbf{x}}_{0:n} = \mathbf{z}_{0:n}^{i^*}$ 
end for

```

but substantially reduces the computational burden because the selection of the best past trajectory for a particle no longer requires the linear state estimates or Kalman

filter recursions for each possible past trajectory. Rather, the best past trajectory can be determined before the Kalman filter recursions. These recursions can then be calculated once per particle instead of N_p times per particle. Algorithm 6 gives a complete description of the Fast MAM-PF algorithm.

C.6 Performance Assessment of Particle Filter Analysis of Vocalizations

One shortcoming of the EKS methodology for vocalization analysis described in Appendix B occurs when vocalizations include abrupt jumps in frequency. Under these conditions, the EKS algorithm may lock onto a sub- or super-harmonic of the fundamental due to the multi-modal nature of the marginal posterior distribution of the fundamental frequency. Many species, however, emit vocalizations containing such changes in frequency. Several examples of mouse vocalization exhibiting these behavior are pictured in Fig. C.1. Here we assess whether PF methods are a better solution for analyzing this class of vocalizations.

Synthetic Signal Generation

As with the performance comparison of the EKS and spectrographic trackers found in B.5.1, an ensemble of synthetic vocalizations was used to assess the tracking capabilities of the Fast MAM-PF algorithm described in C.5. Synthetic signals with known frequency were used such that the normalized mean squared error (NMSE) of the frequency estimates of each method could be calculated across a range of signal to noise ratios (SNRs). Furthermore, the true signal in the absence of measurement noise is known for each synthetic vocalization, enabling the calculation of the NMSE of the reconstructed signal following the analysis phase. This type of analysis is impossible with actual recordings because the true frequencies and measurement noise are not known.

Parameters values for the stochastic signal model defined in A.2.2 were determined in order to generate an ensemble of mouse vocalizations modeled after those pictured in Fig. C.1, but with significant variety to test the tracking capabilities of the different methods under a range of conditions. Note how these two examples of real vocalizations have three distinct sections, with abrupt jumps in frequency between them. To achieve the jumps in frequency in the synthetic vocalizations, each vocalization was comprised of three individually generated sections which were then concatenated to form a composite signal. The parameter values for the generation of these vocalizations are defined in Table C.1 and Table C.2, where “...” is used to denote a range of values that were uniformly drawn from for each simulation, and commas are used to denote the different values for the 3 different sections of the vocalization. Fig. C.2 shows 6 vocalizations generated by this stochastic synthesis method. The top, middle, and bottom rows of this figure display signals with SNRs of 25 dB, 12.5 dB, and 0 dB, respectively. It is understood that these are not exact reproductions of the types of mouse vocalizations shown in Fig. C.1. Instead, they are designed to have a range of spectro-temporal characteristics that are likely to be encountered in real mouse vocalizations.

Results

In order to assess the tracking performance of the EKS, Fast MAM-PF, and spectrographic trackers, synthetic vocalizations were created and measurement noise was added to create 3 different SNR classes: 25 db, 12.5 db, and 0 db. For each of these noise classes, 500 synthetic vocalizations were created and analyzed. Even though there was significant variability among the vocalizations, an attempt was made to optimize the tracking parameters to work well for all of the signals. The measurement noise tracking parameter was modified for each SNR class. These tracking param-

Figure C.1: Real mouse vocalizations used as models for the synthetic ensemble used in the assessment of the particle filter tracking performance. Each of these vocalizations include abrupt jumps in frequency, which pose a problem for frequency tracking methods based on the extended Kalman smoother.

Table C.1: Model parameters for the rapidly changing frequency component of synthetic mouse vocalizations

Description	Symbol	Value
Duration		0.01, 0.01, 0.02 s
Number of Harmonics	m	2
Mean Frequency	$\bar{f}$	17, 25, 15 kHz
AR Coefficient	α	0.9999
Initial State Estimates		
Phase	$\hat{\theta}_0$	0 Radians
Frequency	$\hat{f}_0$	15... 25, 25... 35, 15... 20 kHz
Cosine Amplitudes	$\hat{a}_{k,0}$	[0.1 1.3], [1 0.4], [0.1 1.2]
Sine Amplitudes	$\hat{b}_{k,0}$	$\hat{a}_{k,0}$
Noise Process Variances		
Measurement	r	Variable
Frequency	q_f	10^8
Cosine Amplitudes	q_a	[10 20]
Sine Amplitudes	q_b	q_a

Figure C.2: Examples of synthetic mouse vocalizations used to assess the performance of the EKS, Fast MAM-PF, and spectrographic frequency tracking algorithms on signals with abrupt changes in frequency. The top, middle, and bottom rows correspond to signals with SNRs of 25 dB, 12 dB, and 0 dB, respectively.

Table C.2: Model parameters for the modulating frequency component of synthetic mouse vocalizations

Description	Symbol	Value
Mean Frequency	$\bar{f}^m$	100...400 Hz
AR Coefficient	α^m	0.9999
Initial State Estimates		
Phase	$\hat{\theta}_0^m$	0, 0, $\pi/2$ Radians
Frequency	$\hat{f}_0^m$	100...400 Hz
Cosine Amplitude	$\hat{a}_{k,0}^m$	3000, 5000, 3000
Sine Amplitude	$\hat{b}_{k,0}^m$	$\hat{a}_{k,0}$
Noise Process Variances		
Frequency	q_f^m	10^6
Cosine Amplitudes	q_a^m	10^4
Sine Amplitudes	q_b^m	q_a^m

eters are shown in Table C.3. Because computational expense of the Fast MM-PF algorithm scales exponentially with the number of particles, there were practical considerations that limited the number of vocalizations that were tracked at each SNR and the number of particles used in each simulation.

Fig. C.3 displays the tracking results for three representative synthetic vocalizations using each of the tracking algorithms. From top to bottom, this figure displays signals with SNRs of 25 dB, 12.5 dB, and 0 dB, respectively. In each of these examples, the spectrographic tracker performed the worst. Because it tracks the peak in the spectral power at each time sample of the calculated spectrogram, it would lock onto the first harmonic instead of the fundamental when the first harmonic was dominant. The one parameter that can be used to control this aspect of our implementation of this algorithm is the frequency range over which the peak detection occurs. This value was set to span 5 kHz to 45 kHz in order to cover the full range of the true fundamental of the ensemble of synthesized vocalizations, and significantly overlaps

Table C.3: User specified EKS and Fast MAM-PF model parameters for tracking the synthetic mouse vocalizations. The only parameter modified for the 25, 12.5, and 0 SNR cases was the measurement noise.

Description	Symbol	Value
Number of Harmonics	m	2
Mean Frequency	$\bar{f}$	N/A
AR Coefficient	α	1
Number of Particles	N_p	1000
Initial State Estimates		
Phase	$\hat{\theta}_0$	0 Radians
Frequency	$\hat{f}_0$	True initial frequency for EKS 10. . . 50 kHz for Fast MAM-PF
Cosine Amplitudes	$\hat{\mathbf{a}}_{k,0}$	[0.1 1.3]
Sine Amplitudes	$\hat{\mathbf{b}}_{k,0}$	$\hat{\mathbf{a}}_{k,0}$
Noise Process Variances		
Measurement	r	0.1000, 1.0000, 10.0000
Frequency	q_f	10^{11}
Cosine Amplitudes	q_a	[500 500]
Sine Amplitudes	q_b	q_a

the frequency range of the first harmonic as well. This explains why, for example, the spectrographic estimate jumps from the first harmonic down to the fundamental at about 0.007 s in the top example in Fig. C.3.

While the EKS algorithm generates smooth estimates in each of these three examples, it loses track during the first abrupt frequency jump and tracks a sub-harmonic of the fundamental during the middle section of each vocalization. It does manage to regain track after the second frequency jump in each case, but the NMSE of the frequency estimates and the reconstructed signal are both significantly impacted by the loss of track. The Fast MAM-PF algorithm, on the other hand, is capable of tracking the frequency jumps in each of these examples. While this algorithm provided smooth estimates of the frequency in the nearly noise-free example seen in the top panel of Fig. C.3, there is significantly more variability in the frequency estimates under the noisy conditions. Attempts were made to find better tracking parameters for the Fast MAM-PF algorithm but there were always trade-offs which made further tuning difficult or impossible. For example, increasing the variance of the frequency noise may improve the ability of the algorithm to track a large jump in frequency, but it may also make the algorithm track sub-optimally on sections of the vocalization that have little frequency variability.

Fig. C.4 shows the frequency tracking performance of the three algorithms across the whole synthetic ensemble of vocalizations. The median NMSE for the Fast MAM-PF algorithm's frequency estimates were about an order of magnitude lower than than the median NMSE of the EKS algorithm. Both the EKS and spectrographic approaches resulted in median NMSE values greater than 1, indicating that an estimate equal to the mean of the fundamental frequency would have been an improvement. Similarly, Fig. C.5 shows that the estimates provided by the Fast MAM-PF algorithm resulted in a more accurate reconstruction of the original signal with the measure-

Figure C.3: Tracking of synthetic mouse vocalizations. The top, middle, and bottom rows display examples of tracking signals with SNRs of 25 dB, 12 dB, and 0 dB, respectively. Note how the the EKS tracker locks onto a sub-harmonic of the fundamental in each of these cases. fNMSE indicates the NMSE of the frequency estimates. yNMSE indicates the NMSE of the reconstructed signal, relative to the synthesized signal in the absence of measurement noise.

Figure C.4: Frequency tracking performance of the EKS, Fast MAM-PF, and spectrographic trackers across abrupt frequency jumps. The median value of the NMSE for 500 simulation along with the 95% expectation range is shown. The Fast MAM-PF tracker outperforms both the spectrographic and EKS trackers due to its ability to regain track on the fundamental after abrupt frequency jumps. The error bars indicate the 95% expectation range of the collected data.

ment noise removed. Because the spectrographic algorithm only provided estimates of the fundamental frequency and not the phase and amplitudes of the whole harmonic stack, the signal could not be regenerated and it is not included in this comparison.

Most of the error incurred in the application of the EKS algorithm on this ensemble of signals was due to the loss of track after the first frequency jump. To better characterize this, the frequency tracking and signal reconstruction performance was assessed for the portion of each signal preceding the first frequency jump. These results are displayed in Figs. C.6 and C.7. These results make it clear that the EKS algorithm, on average, outperformed the Fast MAM-PF algorithm across the whole stimulus ensemble. Because PF methods generate asymptotically optimal state estimates as the number of particles goes to infinity, this result is likely a result of sample impoverishment. Practical limitations due to the computational requirements

Figure C.5: Accuracy of signal reconstruction for each synthetic mouse vocalization across abrupt frequency jumps. The median value of the NMSE for 500 simulation along with the 95% expectation range is shown. The Fast MAM-PF tracker outperforms both the spectrographic and EKS trackers due to its ability to regain track on the fundamental after abrupt frequency jumps. The error bars indicate the 95% expectation range of the collected data.

Figure C.6: Frequency tracking performance of the EKS, Fast MAM-PF, and spectrographic trackers on the section of each synthetic mouse vocalization before the first abrupt frequency jump. This allows for a comparison of the tracking performance without the large error that can be introduced by loss of track during the jump. The median value of the NMSE for 500 simulation along with the 95% expectation range is shown. The spectrographic tracker performs particularly poorly, because it tracks the first harmonic, which has more power than the fundamental during this section. The EKS tracker performs better than the PF across the full range of SNRs.

of increasing the number of particles made a statistical assessment of this difficult.

Figure C.7: Accuracy of signal reconstruction for each synthetic mouse vocalization before the first abrupt frequency jump. The median value of the NMSE for 500 simulation along with the 95% expectation range is shown. This assessment indicates how well the signal models are estimated (not just the frequency) by reconstructing the vocalization from the estimated parameter values and comparing the result to the original vocalization before the addition of measurement noise. Here again, the EKS tracker outperforms the Fast MAM-PF tracker across the full range of SNRs. Signal reconstruction using the spectrographic tracker is beyond the scope of this research and is not used in this assessment.

Appendix D: Neural Data Analysis and Visualization

To analyze and visualize data collected by the proprietary software suite (called Batlab) installed in the neurophysiological laboratory, a custom Matlab package was written (called Bat2Matlab) allowing for the import of the binary microelectrode recording (MER) signals and the experimental metadata. This tool provides tunable settings for spike detection from the MER signals as well as processing of automated Batlab tests for generating frequency tuning curves, peristimulus histograms (PSTHs), calibrated intensity curves for vocalizations in units of decibels of sound pressure level (dB SPL), spectrographic representations of the vocalizations, and modeling as described in Section 3.7. A typical plot of a single test is pictured in Figure D.1. This figure represents the data collected from the ICC of a mustache bat to 20 presentations of a conspecific vocalization. The top panel shows the MER signals (each color represents the recording from one presentation) along with asterisks where the spikes were detected. The next panel shows the PSTH, which is a histogram showing where the spikes occurred over the ensemble of presentations, as well as the waveform of the vocalization stimulus. The next panel shows a spectrographic representation of the vocalization stimulus. Finally, the bottom panel shows the intensity of the vocalization in dB SPL.

Figure D.1: Visualizing data in Bat2Matlab