BIOLOGICALLY-INFORMED COMPUTATIONAL MODELS OF HARMONIC SOUND DETECTION AND IDENTIFICATION

by

Kevin J Kostlan

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March 2021

© 2021 Kevin J Kostlan
All rights reserved

Abstract

Harmonic sounds or harmonic components of sounds are often fused into a single percept by the auditory system. Although the exact neural mechanisms for harmonic sensitivity remain unclear, it arises presumably in the auditory cortex because subcortical neurons typically prefer only a single frequency. Pitch sensitive units and *harmonic template units* found in awake marmoset auditory cortex are sensitive to temporal and spectral periodicity, respectively. This thesis is a study of possible computational mechanisms underlying cortical harmonic selectivity.

To examine whether harmonic selectivity is related to statistical regularities of natural sounds, simulated auditory nerve responses to natural sounds were used in principal component analysis in comparison with independent component analysis, which yielded harmonic-sensitive model units with similar population distribution as real cortical neurons in terms of harmonic selectivity metrics. This result suggests that the variability of cortical harmonic selectivity may provide an efficient population representation of natural sounds.

Several network models of spectral selectivity mechanisms are investigated. As a side study, adding synaptic depletion to an integrate-and-fire model could explain the observed modulation-sensitive units, which are related to pitch-sensitive units but cannot account for precise temporal regularity. When a feed-forward network is trained to detect harmonics, the result is always a *sieve*, which is excited by integer multiples of the fundamental frequency and inhibited by half-integer multiples. The sieve persists over a wide variety of conditions including changing evaluation criteria, incorporating Dale's principle, and adding a hidden layer. A recurrent network trained by Hebbian learning produces harmonic-selective by a novel dynamical mechanism that could be explained by a Lyapunov function which favors inputs that match the learned frequency correlations. These model neurons have sieve-like weights like the harmonic template units when probed by random harmonic stimuli, despite there being no sieve pattern anywhere in the network's weights.

Online stimulus design has the potential to facilitate future experiments on nonlinear sensory neurons. We accelerated the sound-from-texture algorithm to enable online adaptive experimental design to maximize the activities of sparsely responding cortical units. We calculated the optimal stimuli for harmonic-selective units and investigated model-based information-theoretic method for stimulus optimization.

Primary Reader and Advisor: Dr. Kechen Zhang

Secondary Reader: Dr. Xiaoqin Wang

Acknowledgements

This dissertation is the result of what has been by far my deepest exploration into scientific research. Extensive work with my thesis advisor, Kechen Zhang, taught me the necessary skills to not only to ask the most interesting questions but also to communicate in an efficient and elucidating manor.

Johns Hopkins introduced me to the world of medical research. It is a place where patience and patients are front-and-center to effective discovery and where topics are explored through a biological lens. Here, academic rigor is held to the highest standards possible.

Through these years, my advisor taught me how the biologists approach problems and communicate as well as how to navigate the world of journal publishing. He gives us freedom to explore our own directions without compromising the academic guidance and feedback. He is an excellent writer and is aware of the best way to communicate difficult scientific concepts to the readers. However, he carefully explains our many writing mistakes and has us correct them, which ensures that we become better writers, even though it takes a significant amount of his time.

My thesis committee, Kechen Zhang, Xiaoqin Wang, and Mounya Elhiali, have repeatably made themselves available for our meetings despite their busy schedule and given very helpful feedback as I progressed from GBO to proposal to permission to write to the public defense.

I also want to thank my two student collaborators, Lixia Gao and Darik Gamble, for working with me and sharing their data to facilitate the process of me designing algorithms for them. They have been good at answering my questions about their data that they shared with me and helping me frame their research into the larger scientific question they are answering.

Professor Emeritus Eric Young has been instrumental for transitioning me to this new university during my time as a masters student. He has extensive experience in mentoring students as well as post-docs in the world of academic research. It was sad to see him retire and his former lab enclosed with plywood construction walls as it was being renovated and replaced.

The administrative director, Hong Lan, also deserves a special mention. She is a very likable person who at one of the PhD Halloween parties earned herself a cheer from the crowd for her hard work helping all of us. She personally has been very helpful for ensuring I and other students have all the steps ahead of us laid out and that the progress is as smooth as possible, sometimes working after-hours to ensure that our needs are met.

Finally, I had support from friends and family throughout this long journey. Many of these friends went down the same road as me, getting their committee together and meeting with them for each milestone. They acted as guiding stars whom I could follow in their footsteps, and explained to me the biggest pitfalls and best approaches to walking down this path.

Table of Contents

Abstract ii
Acknowledgements iv

List of Tables

List of Figures

1. Introduction 1
  1.1 Harmonic sounds and perception 1
  1.2 Ecology and acoustic environments 4
  1.3 Subcortical anatomy and harmonic processing 5
  1.4 Cortical anatomy and harmonic processing 7
  1.5 Computational models of pitch perception and harmonic processing 12
    Time-domain models 13
    Frequency-domain models 14
  1.6 Objectives of this dissertation 17
    1: Learning harmonic selectivity from statistical structure 17
    2a: Learning harmonic selectivity with feedforward and recurrent networks 17
    2b Analysis of recurrent networks by Lyapunov functions 18
    3: Optimal stimulus design 19
2. Learning harmonic structure from natural sound statistics 21
  2.1 Introduction 21
  2.2 Methods 25
    Sound Stimuli 25
    Marmoset vocalizations 25
    Generic bank 25
    Vowels 26
    Harmonics, mistuned harmonics, and white noise 26
    Generating the training data 27
    Time windows 30
    Weight Training 31
    Harmonicity Detection 33
    Population statistics 35
  2.3 Results 35
    Sparsity 42
    Harmonic selectivity 43
  2.4 Discussion 50
    Both ICA and PCA can yield realistic levels of harmonic selectivity 51
    ICA with AN model shows more realistic tuning curve width and BF range 52
    Biological relevance of PCA and ICA learning 52
    Relation with harmonic selectivity in auditory cortex 53
3. Training harmonicity selectivity in feedforward neural networks 56
  3.1 Introduction 56
    The subcortical levels are monospectral 56
    Can temporal cortical models be used? 57
    Is a sieve trivial? 58

3.2 Methods........58
Spectro-temporal subcortical models........58
Temporal cortical modelling........71
Training stimuli........77
Spectral subcortical models........77
Noise model........78
Neural network topology........79
Neural network model........80
Supervised training and testing........81
Effective linear weights........82
Linearized weights around an input........82
RSS weights........83
Facilitation and periodicity indexes........83
Sieveness and linclassness........83
3.3 Results........84
Sieves are robust........84
Dale's principle splits the sieve........87
RSS weights are also sieve-like........88
If it can be solved, it is a sieve........89
3.4 Discussion........91
Optimal linear classifiers are sieves........91
Optimal classifiers and log likelihood........92
Locality constraints as biological budgeting........92
Breaking down a sieve into excitation and inhibition........92
Future experiments........93
4. Training harmonicity selectivity in recurrent neural networks........94
4.1 Introduction........94
The abundance of auditory cortical recurrent networks........94
Training of recurrent networks........95
4.2 Methods........97
Neural network topology........97
Neural network and noiseless model........98
Subcortical input........98
Hebbian training and testing........100
Eigen-analysis of linearized dynamics over time........102
4.3 Results........102
Weights are an eggcrate........102
Response to a fixed stimulus over time........105
Response to tones, harmonics, and mistuned harmonics........108
Random harmonic stimuli and regression weights........110
Reponses to a rise-fall stimuli and eigen analysis in the hysteresis loop........115
Pattern completion........116
Oscillation........119
Sensitivity to parameter changes........120
Secondary parameter effects........121
Eggcrate FI is limited, unless variation in population is accounted for........122

Covariance learning as an alternate rule 122
Response of a combined feedforward and recurrent network 123
4.4 Discussion...123
Input dynamic range 123
Egg-crates as an alternate to sieves...124
Experimental considerations...125
5. Analysis of harmonically selective recurrent networks 126
5.1 Introduction...126
The richness of recurrent networks...126
5.2 Analytical two-unit phase-plane analysis 127
ReLU phase-plane analysis..127
Zrect phase-plane analysis 134
5.3 Converting neural networks into a two-unit networks 140
Phase offsets..142
5.4 Lyapunov function analysis 143
Gains and Lyapunov functions 143
Lyapunov functions as objective functions..144
5.5 Bf-agnostic harmonicity detection..145
A highly nonlinear problem 145
Solving the agnostic detection of harmonics problem 148
5.6 Optimal hysteresis parameters for autocorrelated time-series 152
5.7 Other potential recurrent network advantages 157
Locality constraints as biological budgeting..157
Robustness to damage..157
6. Optimizing stimulus design for neural response and model identification159
6.1 Introduction...159
The space of perception 159
Model-free approach vs model-informed approach...160
6.2 Online, model-free determination of optimal stimulus 160
Accelerating the inverse problem 163
Filling the wave in under the envelopes 164
Results...167
6.3 Computational model-based optimal stimulus design for recurrent networks 167
Reproduction of DiMattina and Zhang, 2011 168
Optimal stimulus design for recurrent networks..169
Egg-crate weight parameterizations...172
Optimal stimulus design for recurrent networks, results 172
6.4 Discussion...175
7. Conclusion...177
7.1 Temporal harmonic selectivity mechanisms 177
7.2 Subcortical models..178
7.3 Natural sound statistics generate some harmonically selective units.....................179
7.4 Training harmonic template units produces sieves 180
7.5 Recurrent network harmonic selectivity 181
7.6 Recurrent network fingerprints 181
7.7 Variable *f0* harmonicity detection..182

7.8 Model-free stimulus design with sound textures 183
7.9 Model-based stimulus design 184
REFERENCES 185
Curriculum Vitae 201

List of Tables

Table 3.1: Model IC unit parameters 69
Table 3.2: Parameter values and ranges used for the integrate-and-fire model 73
Table 3.3: Summary of training conditions, accuracy, and performance 91
Table 4.1: Effects of changing individual parameters 121
Table 5.1: Parameters for the ReLU phase-plane analysis 128
Table 5.2: Exact values of ReLU equilibrium points 131
Table 5.3: Parameters for the Zrect phase-plane analysis 134
Table 5.4: Exact values of Zrect equilibrium points 138
Table 5.5: Four Lyapunov functions for four commonly used neural network gains 143
Table 5.6: The training and test performance data of the various networks. 152
Table 5.7: Hysteresis switching rules 153
Table 6.1: All the statistics and their dimensionality 163

List of Figures

Figure 1.1: The most important connections in the auditory ascending pathway 5
Figure 1.2: The core and belt regions of the Marmoset auditory cortex............................ 8
Figure 1.3: A marmoset cortical harmonic template unit 11
Figure 1.4: Two marmoset cortical harmonic template units 12
Figure 1.5: Modelling harmonically selective units with a spectral sieve........................16
Figure 2.1: Examples of sounds used in the training dataset...29
Figure 2.2: The effect of different window sizes on the trained weight vectors................31
Figure 2.3: Top raw-sound weight vectors for three soundbanks....................................36
Figure 2.4: Weight vectors of hand-selected “non-harmonic environmental sounds” 38
Figure 2.5: The top six weight vectors for PCA and ICA on the AN model....................41
Figure 2.6: Examples of a harmonic template unit weight vectors..................................44
Figure 2.7: The most harmonically selective PCA and ICA weight vectors 45
Figure 2.8: Cumulative distributions of the FI and PI 47
Figure 2.9: Cumulative distributions of the FI and PI under a randomized threshold.......49
Figure 3.1: PSTH’s of an AN fiber unit at 2000 Hz 62
Figure 3.2: PSTH’s of an AN fiber unit at 5000 Hz 62
Figure 3.3: Rate-level functions in comparison to experimental data 63
Figure 3.4: Steady-state response map of a model high spontaneous AN unit.................63
Figure 3.5: Primary-like and type 2 CN unit response maps..65
Figure 3.6: Instantaneous firing rates of model cochlear nucleus neurons.......................68
Figure 3.7: Model inferior colliculus neurons 70
Figure 3.8: Modulation-sensitive neurons in the marmoset A1.......................................71
Figure 3.9: Demonstration of the “synaptic vitality” at work...72
Figure 3.10: The integrate-and-fire model’s results 76
Figure 3.11: Example training stimuli and subcortical input..78
Figure 3.12: Diagram of the simplest feed-forward training network 80
Figure 3.13: Diagram of the sign-constrained network that respects Dale’s principle......80
Figure 3.14: The sign-constrained network represented differently.................................80
Figure 3.15: How a sieve network detects harmonics 82
Figure 3.16: The weight vector for the simplest harmonic training case..........................84
Figure 3.17: The weight vector for 1-norm-constrained training case 87
Figure 3.18: The weight vector for the Dale’s principle training case 88
Figure 3.19: RSS linear weights showing a sieve-like structure 89
Figure 3.20: A one-hidden-layer network..90
Figure 3.21: A two-hidden-layer network 90
Figure 4.1: A cat auditory cortical dye-injection experiment...95
Figure 4.2: A sign-agnostic recurrent network 97
Figure 4.3: A recurrent network respecting Dale’s principle 98
Figure 4.4: Subcortical inputs for stimuli for fixed BW and for fixed Q 99
Figure 4.5: How training produces an egg-crate network..101
Figure 4.6: Finding the point in the approximate center of the hysteresis loop..............102
Figure 4.7: The recurrent weights for a fixed bandwidth sign-agnostic case 103
Figure 4.8: The recurrent weights for a fixed Q-factor sign-agnostic case 103
Figure 4.9: The recurrent weights for the Dale’s principle case.....................................104

Figure 4.10: Presenting a fixed stimulus to the network 105
Figure 4.11: The network state over time to a fixed harmonic stimulus 106
Figure 4.12: The network response over time to a fixed harmonic stimulus...................107
Figure 4.13: Stimuli and subcortical input to assess harmonic selectivity 108
Figure 4.14: Response maps to tones, harmonics, and mistuned stimuli 108
Figure 4.15: The FI and PI of each unit 109
Figure 4.16: The histogram of FI and PI values 110
Figure 4.17: A sample of random harmonic stimuli 111
Figure 4.18: The RHS linear and quadratic weights for a single unit.............................112
Figure 4.19: Integer-unit RHS weights and eigenanalysis..113
Figure 4.20: Halfinteger-unit RHS weights and eigenanalysis.......................................114
Figure 4.21. The state and response to a rise-fall stimuli 115
Figure 4.22. Eigen analysis in the middle of the hysteresis loop....................................116
Figure 4.23. Stimuli with missing components 117
Figure 4.24. The network state for stimuli with missing components............................118
Figure 4.25: Selected neural states for stimuli with missing components.......................119
Figure 4.26: Oscillation under a different parameter set 120
Figure 4.27: The covariant learning rule...122
Figure 5.1: Two unit ReLU phase-planes for various parameters...................................133
Figure 5.2: Two unit Zrect phase-planes for various parameters 139
Figure 5.3: A recurrent network approximation of max-of-sieves 149
Figure 5.4: The sieve layer in the winner-take all network 151
Figure 5.5: The recurrent network responses to a fixed harmonic stimulus...................152
Figure 5.6: Accuracy as a function of g..155
Figure 5.7: Accuracy as a function of m and p 155
Figure 5.8: Accuracy of the optimal hysteresis parameters when p is small..................156
Figure 6.1: Schematic of how the sound textures are generated.....................................161
Figure 6.2: An example of summary statistics for a sound texture 162
Figure 6.3: Sampling from arbitrary distributions 165
Figure 6.4: The process of calculating the fine structure..166
Figure 6.5: Generating the soundwave the fine structure and envelope 167
Figure 6.6: The maximum information stimulus design figure reproduction.................168
Figure 6.7: The equilibrium state and sensitivity given a tone stimulus 173
Figure 6.8: The equilibrium state and sensitivity given harmonic/mistuned stimuli.......173
Figure 6.9: The optimal stimulus when given an identity covariance matrix of θ 174
Figure 6.10: The optimal stimuli to deduce the individual elements of θ 175

CHAPTER 1:

Introduction

The auditory system detects pressure changes down to a few dozen μPa and ranges from 20-20kHz. It converts the two waveforms the ears pick up into a rich auditory scene with multiple auditory objects, many of which feature a highly salient *pitch*. Pitch is vital to source-separate an acoustic scene (Assmann & Summerfield, 1990; de Cheveigné et al., 1995), and is to a significant degree innate in humans (Clarkson & Rogers, 1995; Rogers et al., 1996). Pitch perception has been found in birds (Cynx & Shapiro, 1986), cats (Evans & Whitfield, 1964; Heffner & Whitfield, 1976), macaques (Tomlinson & Schwarz, 1988), and marmosets (Osmanski & Wang, 2011). The mechanism and utility of pitch perception and harmonic processing is explored in this thesis from a computational modelling perspective. Sounds must be segregated and classified. Numerous studies under diverse experimental conditions found that binaural cues help with sound segregation, see (Schwartz et al., 2012) for a review. However, monaural cues stimuli are also very important for this task and will be the focus of this thesis.

1.1 Harmonic sounds and perception

Non-inanimate sounds such as animal vocalizations (Agamaite et al., 2015; Singh & Theunissen, 2003) including human speech (Houtgast & Steeneken, 1973; Rosen, 1992), and music (Peretz & Zatorre, 2005) typically have information distributed across a wide range of frequencies. These frequencies are often *harmonically related*: these sounds have spectral energy at small integer multiples of a fundamental frequency (f_0) and the waveforms repeat with a fixed time interval $t_0 = 1/f_0$ (Von Helmholtz, 1912). Harmonic sounds are allowed to have missing components. For example, a “perfect fifth” has a 3:2 ratio of frequencies and so is missing the fundamental frequency. The underlying physics of resonance in string and wind instruments (of which vocal chords could be considered a hybrid) is responsible for making harmonicity so common (Campbell et al., 2004). Given their ubiquity, the auditory system must develop mechanisms for processing harmonic sounds. Furthermore, cortical associative learning is likely to produce harmonic associations.

Marmoset calls are often harmonic and serve different roles: “Twitters” are bird-chirp-like and may represent territory claims; whistling “phee” calls as well as rapidly “warbling” trill calls are used for within-group contact (Agamaite et al., 2015; Bezerra & Souto, 2008; Epple, 1968; Pistorio et al., 2006). Hybrids and other types of calls are made as well, and each type of call has a variety of uses (Pistorio et al., 2006), thus marmosets likely benefit from mechanisms to detect harmonicity and make for a good model species for electrophysiological experiments.

Harmonics are highly perceptually salient. These sounds are fused into a single percept with a *pitch* at f_0 and this pitch is present even when the component at f_0 is removed, a phenomenon called the residue pitch or “missing fundamental” (Jackson & Moore, 2013; Licklider, 1956; Schouten, 1968). A mistuning of a component of 1-3% is usually enough to have it perceived as a separate sound (Moore et al., 1986) and the threshold for detection of mistuning a component for a 200Hz fundamental is about 2Hz from the fundamental to the 10th harmonic (Moore et al., 1985). However, the relative accuracy degrades at much higher frequencies where phase-locking is weak (Hartmann et al., 1990).

Conversely, non-harmonic sounds (apart from pure tones) have much weaker pitch than harmonic sounds. Spectrally delocalized “broadband” sounds such as waterfalls, consonants, and white noise lack precise spectral structure and as expected have absent or very weak pitch. However, inharmonic sounds with strong peaks in their spectral structure still have significantly degraded pitch despite having well-concentrated spectral energy (Boer, 1956; Schouten et al., 1962).

The perceptual segregation and classification of harmonic sounds is useful for segregating different sound sources (Bregman, 1994; Darwin & Carlyon, 1995; Houtsma & Smurzynski, 1990) or for processing vocal communication sounds in noisy environments (Bates et al., 2011; de Cheveigné et al., 1995; Feng et al., 2006). The segregation of multiple voices "cocktail party" but not the perception of single voices is degraded when the sounds are made inharmonic by shifting the carrier frequencies (Popham et al., 2018). The pitch-shift of the out-of-place component is exaggerated even if the harmonic complex is mistuned as a whole and regardless of a ±500ms delay of the oddball component (Brunstrom & Roberts, 2001). Neural responses in the macaque cortex are enhanced to the mistuned component for tone-sensitive neurons (Fishman & Steinschneider, 2010). The accuracy of f0 discrimination degrades for harmonic stimuli that are missing more than the first 10 components, but if the odd components are mistuned 3% the performance is nearly equal to that with those components removed (which is equivalent to doubling the f0 value) (Bernstein & Oxenham, 2008). This is further evidence that components with even slight mistuning are ignored for the purposes of building up an object.

Musical perception also relies on harmonic processing. *Octave equivalence* is the process in which tones octave apart "rhyme" with each-other (Borra et al., 2013; Deutsch & Boulanger, 1984). So-called *consonant chords*, which have better alignment of their harmonics, are musically preferred to *dissonant chords* (Krumhansl, 1979; Malmberg, 1918; McDermott et al., 2010).

1.2 Ecology and acoustic environments

An acoustic environment is a random variable that the auditory system analyzes. Summary statistics such as mean and standard deviation are very concise but miss much of the statistical structure. On the other hand, neurons are likely to be efficient encoders over the stimuli they tend to encounter and in doing so preserve most of the information (Day & Delgutte, 2016; Lewicki, 2002; Simoncelli & Olshausen, 2001; Smith & Lewicki, 2006) and achieve this in part by redundancy reduction (Barlow, 2001; Barlow & others, 1961). Also, a sparse-encoding computational method reproduced experimentally measured STRF’s in the inferior colliculus (IC) (Carlson et al., 2012).

Principle component analysis (PCA) minimizes the reconstruction square error when the number of stored components is limited. This can be approximated biologically with Sanger’s rule, which is a modified Hebbian rule: $\dot{w_i} \sim \langle y_i (x - \sum_{j \le i} y_j w_j) \rangle$ (Oja, 1992); a real biological system is a mixture of “hardwiring” and plasticity.

Independent Component Analysis (ICA), which tries to find statistically independent components in the data, is another compression method. In practice, algorithms approximate ICA by maximizing kurtosis and/or by enforcing a sparsity constraint. ICA has had a lot of success including reproducing Gabor filters in the primary visual cortex as well as gammatone-like filters in the auditory nerve (AN) (Lewicki, 2002) and cortical receptive fields (Blättler & Hahnloser, 2011). However, studies investigating if harmonic selectivity can arise from this are lacking.

Both PCA and ICA yield an optimized population of neurons. Summary statistics can be compared to what was found in (Feng & Wang, 2017) in terms of mistuning sensitivity and harmonic preference over tones. Similar in vivo and in silico statistics make it more likely that harmonic selectivity follows the efficient coding hypothesis. Furthermore, information-based *algorithms are more generalizable than traditional curve-fitting methods (the latter performs poorly with extrapolation).*

1.3 Subcortical anatomy and harmonic processing

The auditory system performs remarkably sophisticated subcortical processing. Several layers of high-speed subcortical nuclei analyze information both in the time and frequency domain before it reaches the (much slower) cortex. See Figure 1.1 for an overview of the pathways.

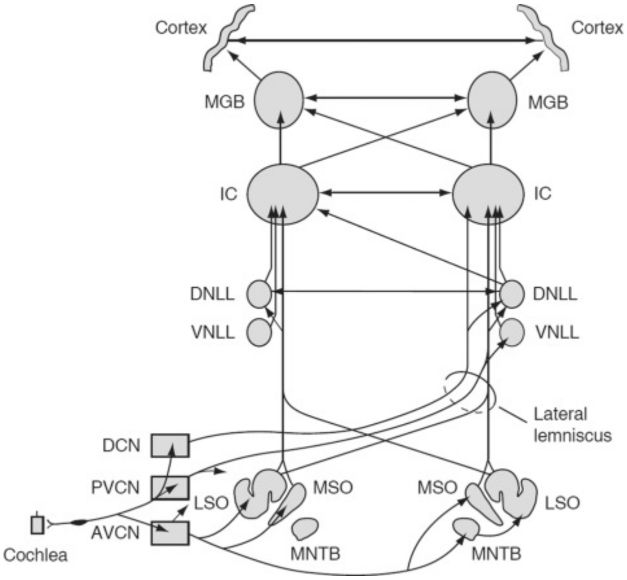


Figure 1.1: The most important connections in the auditory ascending pathway. DCN, PVCN, AVCN: Cochlear nucleus (dorsal, posterior ventral, anterior ventral). LSO, MSO: Lateral and medial superior olivary complex. DNLL, VNLL: Dorsal and ventral nucleus of the lateral lemniscus. IC: The inferior colliculus. MGB: The medial geniculate body (the thalamus). *Figure from (Pickles, 2013).*

Throughout the system there is a *tonotopy* in which regions sensitive to different frequencies are arranged in distinct *laminae* (Abeles & Goldstein, 1972; Fishman et al., 1998, 2013; Kalluri et al., 2008; Sadagopan & Wang, 2008; Schwarz & Tomlinson, 1990). At higher levels, particularly in the cortex, the response is more likely to be *sparse* and unable to be driven by simple stimuli such as tones (Hubel et al., 1959). Also, at higher and higher levels the sound is as well as progressively less phase-locked to temporal stimuli from several kHz at the AN to below 100Hz at the cortex; the *numerous* studies supporting this picture are summarized in (Pickles, 2015).

At the cortical level it is *very important* that the animal is *awake* during electrophysiological experiments as even light anesthesia drastically changes the response properties of neurons, in particular removing (Decharms & Merzenich, 1996) the sustained response found in awake (Bieser & Müller-Preuss, 1996; Chimoto et al., 2002; Wang et al., 2005) animals. Any study under anesthesia thus introduces a large amount of experiential error if used to model awake animals; this thesis focuses on awake Marmoset studies to compare against.

The first step, mechanical transduction, essentially transforms mechanical vibrations in air to hair-cell vibrations and then to neural signals in the auditory nerve (AN). From then on there is little, if any, mechanical cues to all but the loudest sounds. This *transduction* process filters sounds into relatively constant Q-factor bands with a much sharper high-frequency cutoff than low-frequency cutoff (Yang et al., 1992). This is followed by compression and rectification and low-pass filtering (Shamma et al., 1986; Shamma & Morrish, 1987). This process produces nonlinearity in the cochlea which causes harmonic "distortion products" (Robles et al., 1991). However, the response to tones is always single-peaked and purely excitatory in the auditory nerve, with the "best frequency" (BF) designated as the peak of the neuron (Kiang et al., 1967).

The cochlear nucleus processes nearby frequencies at once. Surround inhibition sharpens the input and improves spike timings which may help with pitch perception (Oertel et al., 1990; Rhode, 1995; Winter & Palmer, 1995). There are hints of harmonic processing at the cochlear nucleus level in which some neurons have a much smaller secondary peak that tends to be harmonically related to the BF (Marsh et al., 2006).

There is also a limited degree of across-spectral processing and temporal-structure processing in the inferior colliculus (IC), the main processing center below the cortex. The IC integrates nearby spectral bands with center-surround patterns; with a strong non-linearity that complicates predicting complex spectra responses from pure tunes (Ehret & Merzenich, 1988). Multipeaked neurons have been found in the IC (Portfors & Wenstrup, 2002). The IC also has neurons sensitive to modulation frequency, with bandpass neurons as well as other response types; these neurons seem to form an axes perpendicular to the tonotopy (Langner et al., 2002). Modulation is potentially a mechanism to detect pitch that can complement raw frequency mechanisms.

However, the overall behavior of the ascending IC pathway was found to be mono-spectral: the units tend to have single-peaked receptive fields or in the type O neurons nearby doublets without receptive fields that span multiple octaves (Kostlan, 2015). They responded to harmonics similarly to a superposition of the individual tones and the type I neurons could be used to construct a cortical neuron that would be receptive to harmonics (Kostlan, 2015). It is possible that some of the reports of subcortical multipeaked tuning are instead a harmonic distortion product nonlinearity generated by the speakers or signal processing.

1.4 Cortical anatomy and harmonic processing

The thalamus is the last step before the cortex, and has been found to contain multipeaked neurons (Villa, 1990) but due to the heavy recurrent connections between it and the cortex it has been suggested to combine them as one “thalamocortical” unit (Pickles, 2013).

Thus, the different frequency laminae from the IC must be combined in order the for auditory system to extract pitch as per the mechanisms suggested in (Goldstein, 1973) and

(Cohen et al., 1995). The auditory cortex is likely the main place where said integration happens. Lesions here impair pitch perception in a missing fundamental task but not the non-missing fundamental control task (Zatorre, 1988). Lesions also impair sound discrimination of vocalizations and vowel-like sounds but not tones and other simple sounds (Hefner & Heffner, 1986; Kudoh et al., 2006; Whitfield, 1980). In the primate auditory cortex, shown in figure 1.2, there is a central "core" region which includes the primary cortex "A1" and the primary-like rostral region R. Surrounding this is a "belt" which has its own sub-regions (Kaas & Hackett, 2000).

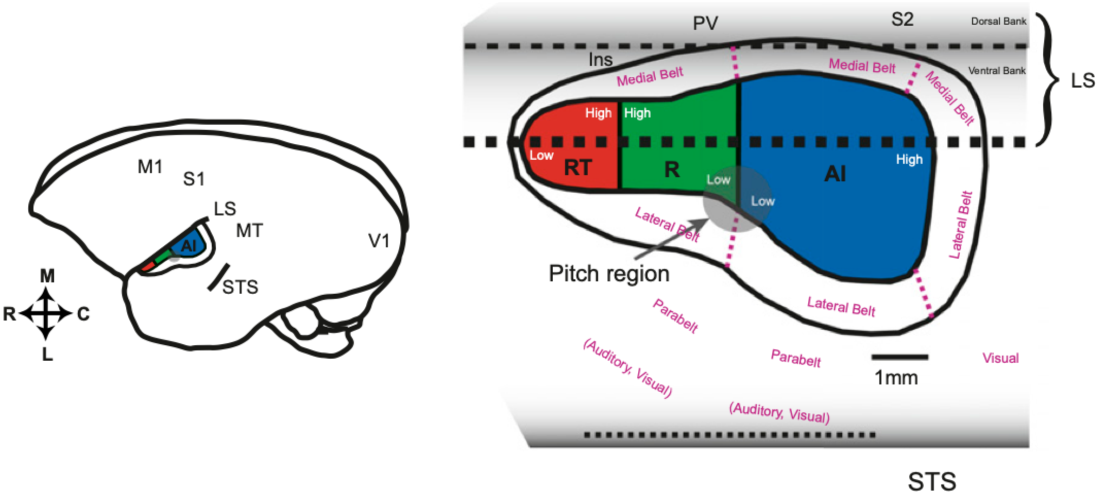


Figure 1.2: The core and belt regions of the Marmoset auditory cortex. *Figure from (Feng & Wang, 2017).*

Neurons in the core region typically have a tonotopy and are sensitive to individual spectral components and/or specific modulation frequencies (Fishman et al., 2013; Schwarz & Tomlinson, 1990; Fishman et al., 1998; Kalluri et al., 2008).

At low frequencies, below 1kHz, a *pitch* region of the marmoset cortex contains neurons which are selective to click trains *of a particular frequency and temporal regularity* (Bendor & Wang, 2005, 2010). Other neurons in this region are receptive to modulation frequencies but not the precise timing (Gao et al., 2016). Similar “pitch centers” were found in humans and macaques (Norman-Haignere et al., 2013, p.; Patterson et al., 2002; Penagos et al., 2004).

At higher frequencies, spectral integration is in the cortex is occurring. Although the majority of neurons are single-peaked, multipeaked neurons have been found in songbirds (Lewicki & Konishi, 1995; Margoliash, 1983), bats (Fitzpatrick et al., 1993; Suga et al., 1983), cats (Abeles & Goldstein, 1972; Matsubara & Phillips, 1988; Phillips & Irvine, 1981; Qin et al., 2005; Sutter & Schreiner, 1991), marmosets (Aitkin & Park, 1993; Kadia & Wang, 2003; Sadagopan & Wang, 2009), and macaques (Rauschecker et al., 1997). Also, *distant inhibition*, where energy far from BF suppresses the neuron, was found in marmosets (Kadia & Wang, 2003), bats (Kanwal et al., 1999), cats (Sutter et al., 1999), and gerbils (Kurt et al., 2008; Moeller et al., 2010). Broad receptive fields were found to extend up to 5 octaves in the rat A1 (Kaur et al., 2004). Anatomical evidence for distant connections exists: there have been long-range recurrent connections within various cortexes including A1 (Gilbert, 1998; Gilbert & Wiesel, 1979; McGuire et al., 1991; Moeller et al., 2010; Wallace et al., 2002). In the cat A1 supergranular layers, these connections can span two octaves and be a few mm long (Kadia et al., 1999; Matsubara & Phillips, 1988; Ojima et al., 1991; Reale et al., 1983; Wallace et al., 1991; Winer, 1992). It has been shown in the cat cortex that these connections are periodic along the tonotopic axis (Wallace et al., 1991). Indeed, the auditory cortex of many species has widely-separated connections that are much more distant than basic center-surround patterns are.

In some cases these peaks are harmonically related, for example a bat cortical unit that is sensitive to the first several overtones of its BF (Suga et al., 1979, 1983). However, multipeaked units are not always harmonic (Kanwal et al., 1999). Harmonic sensitivity was found in A1 and the anterior auditory field of ferrets (Kalluri et al., 2008) and in evoked field potentials in A1 of macaques (Fishman et al., 1998, 2013). Octave two-tone facilitation was found in multi-unit recordings in the macaque A1 (Brosch et al., 1999) and the cat A1 (Brosch & Schreiner, 2000). Octave-separated tones also produce better firing synchrony between to multi-unit clusters in the macaque cortex than other ratios (Brosch et al., 2013), and Multispectral activity with octave spacing was found in the human auditory cortex by FMRI (Moerel et al., 2015).

Multipeaked units in the aforementioned studies often had non-linear facilitation, where the response to two or more tones placed in the different peaks was greater than the sum of the responses to the single tones.

Spectrally-dense stimuli with multiunit recordings revealed that the frequencies of ~3, 5, 10, and 20 kHz (approximate powers of two) were overrepresented in the cat auditory cortex (Noreña et al., 2008). A similar pattern has been found in bats (based around their sonar fundamental) (Suga et al., 1983). This is a different property than facilitation to octave-spaced tone-pairs but it also suggests octave-based organization. Thus the auditory cortex is a hotbed of harmonic spectral integration.

In the marmoset core auditory cortex, spectrally-based harmonic processing is also common. This occurs for *resolved* harmonics, which ranges up to the fourth component at low frequencies through the sixteenth harmonic component at higher frequencies (Osmanski et al., 2013). Although 80% of marmoset cortical units sampled in (Kadia & Wang, 2003) are single-peaked, about half of these had significant two-tone facilitation effects. About 20% of cortical neurons sampled were multipeaked neurons, within which there was a disproportionate representation of small-integer ratios (such as 3/2 or 2/1) between the BF of these neurons and faciliatory and inhibitory frequencies (Kadia & Wang, 2003). Some of these neurons were in the marmoset vocalization range of 4-16 kHz (Agamaite et al., 2015; Pistorio et al., 2006) with a 4-8 kHz fundamental (DiMattina & Wang, 2006). In (Feng & Wang, 2017), about 25% of the neurons sampled were estimated to be *harmonic temple units* which a sieve-like receptive field that is excited by integer multiples of their best fundamental frequency Bf_0. These units are shown in figure 1.3 and figure 1.4 The criteria was neurons that fire to their best harmonic stimuli at least three times as much as to when said stimuli is fully mistuned and at least twice as much to tones of any frequency.

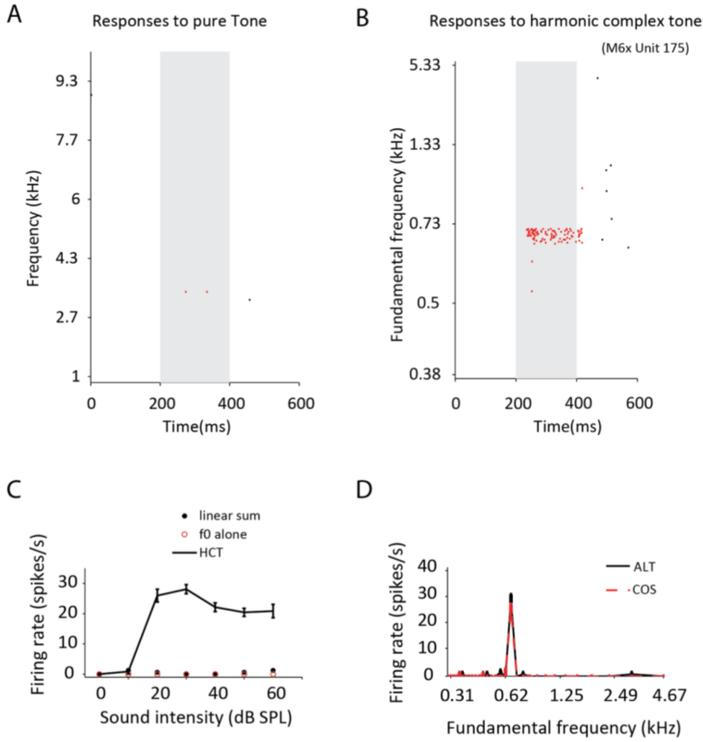


Figure 1.3: A marmoset cortical harmonic template unit. The unit responds to harmonics far more than to tones. *Figure from (Feng & Wang, 2017)*.

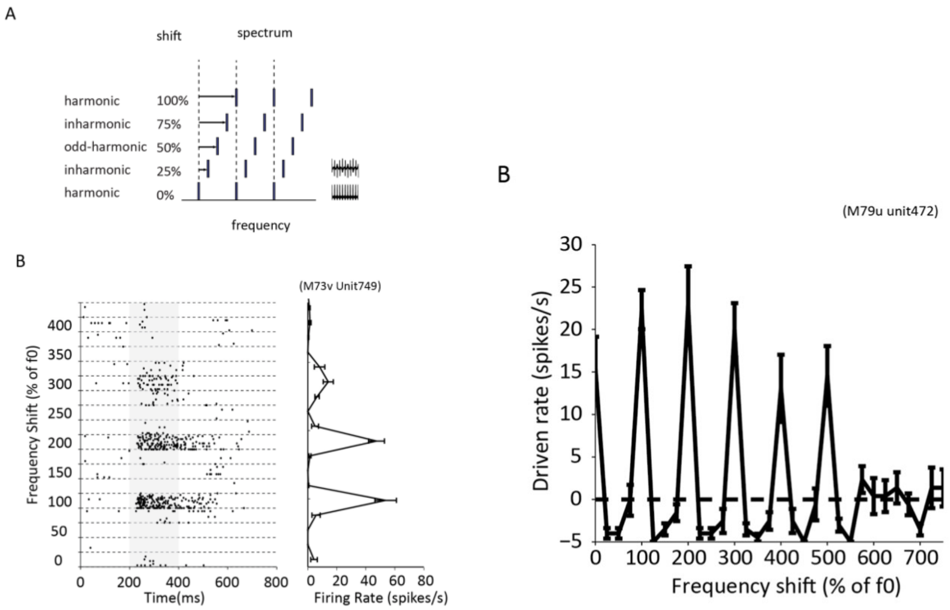


Figure 1.4: Two marmoset cortical harmonic template units. The unit responds to harmonics far more than to mistuned harmonic complexes. *Figures from (Feng & Wang, 2017)*

These template units will tend to extract harmonic sounds from noisy auditory scenes as a single object or percept, and units with inhibitory harmonic selectivity could be removing unwanted harmonic background stimuli (Wang, 2013).

1.5 Computational models of pitch perception and harmonic processing

A computational pitch model is a function that converts the sound waveform into an estimate of the perceived pitch and pitch salience. Pitch is nontransitive, enabling Shepard tones which are a loop of sounds of endlessly "increasing" pitch (Braus, 1995). There are several models of extracting the pitch from the waveform (de Cheveigné et al., 1995). However, the neuronal mechanisms that produce pitch perception are not fully understood.

Time-domain models

Pitch can be estimated by looking for temporal regularity of the sound. As pitch is driven by periodicity, a simple model is to compute autocorrelations: $r(\tau) = \frac{\int \left(x(t)x(t-\tau)\right)^2 dt}{\int x^2 dt}$, where the pitch is $\frac{1}{argmax_{\tau > \epsilon}(r)}$ (De Cheveigne, 2005). It has been proposed (Licklider, 1956) that the auditory system computes $r(\tau)$ across various τ and BF values (i.e. a bank of bandpass filters) in its pitch processing, which could be summed over BF (Meddis & Hewitt, 1991). This would be implemented with delay lines combined with coincidence detectors (Licklider, 1956).

Temporal models should produce similar phase results to human perception. Bohr's phase rule indicates that phase of spectrally resolved components usually has little effect on the sound's pitch, however for highly unresolved components ALT phase will tend to sound an octave higher than SIN or COS phases (De Boer, 1976). This property is also a feature of some autocorrelation models (De Cheveigne, 2005). However, it is possible to generate click trains with the same autocorrelation function but different pitches by making sequences of hybrid randomized and fixed intervals (Pressnitzer et al., 2002), so a simple autocorrelation function isn't sufficient in all cases. Several modifications of these models have been developed. For example, $r'(\tau) = \frac{1-\int \left(x(t)-x(t-\tau)\right)^2 dt}{\int x^2 dt}$ gives the *cancellation model* which can be implemented with inhibitory neurons acting as gatekeepers (De Cheveigné, 1998). The strobed temporal integration model correlates the signal to click trains of various frequencies (that have been filtered) (Patterson et al., 1995). These models may account for some phase (in)sensitivities in pitch perception.

A related method is to compute interspike-interval (ISI) histograms, which requires a model of spikes generation or experimental data such as the cat AN in (Cariani & Delgutte, 1996). The autocorrelation function, if preceded with a spike generator, is equivalent to an *all-order* ISI histogram because it is agnostic to how many spikes lie between two spikes separated the interval τ (De Cheveigne, 2005). In contrast, a first-order histogram only counts spikes if they have no spike in-between them. First-order ISI's have been used in pitch perception models (Cariani & Delgutte, 1996; Rhode, 1995). However, when there is a bank of neurons all-order statistics have much higher sample sizes so low-order ISI statistics may be more prone to error (De Cheveigne, 2005).

The mean rate of click trains has also been found to affect perception: in which the perceived pitch drops to a click train if clicks are randomly removed, even though the *location* of the peak of the ISI histogram and auto-correlogram is unaffected (Carlyon, 1996).

Various neural-network implementations have been developed that attempt to more closely approximate the underlying neural processes. A feed-forward coincidence detector network was developed to explain periodicity sensitivity in the cortex (Huang & Rinzel, 2016). Balanced excitation and inhibition can detect signal periodicity in the inter-click-interval (ICI) ranges observed of the periodicity sensitive units with the inhibitory time-constant controlling the best-ICI (Bürck & van Hemmen, 2009). It's also possible to compute the pitch through *instantaneous* correlations across different pairs of BF AN fiber activities with no delay line: the rectification steps produce harmonic distortions that in turn produce positive correlations between harmonically-related AN fiber spikes (Shamma & Klein, 2000).

Frequency-domain models

Pitch can also be estimated by the spectral components of a sound. Ohms pitch law is one of the earliest (1843) models which simply assumes pitch to be the location of the spectral energy

(Whitfield, 1980). However, this fails to detect the missing fundamental among other inaccuracies, so there are several pattern-matching models to improve upon this. In Goldstein's theory of pitch perception each fundamental frequency is given a template pattern, which is a list of frequencies that includes all the resolved (separated by at least the local cochlear bandwidth) components (Duifhuis et al., 1982). In this model sounds are Fourier-transformed into a list of components. The pitch of a sound is the maximum likelihood estimator overall all templates, assuming additive gaussian noise to the location of the components in said sound (Duifhuis et al., 1982). The Wightman pitch perception model also computes the Fourier transform of the sound, but it then taking it's absolute value and computing the maximum of *its* Fourier transform (Wightman, 1973). This works well for harmonic complexes with many components, but doesn't work well for pure tones. The Terhardt model is similar, except that it uses the estimated cochlear loudness instead of the Fourier transform and uses templates to match against (Terhardt, 1974). These spectral-based pitch models essentially measure modulation spectra in the power-spectrum.

Various neural network models of frequency-based harmonics were developed. A model that is trained on banks sounds (which get stored in long-term memory) and then template-matched with short-term memory from the sound played was developed (McLachlan, 2011). Pitch was also extracted by a sparse-coding compression applied to the responses of a bank of AN fibers; said selectivity was *mostly* attributed to spectral-domain cues (Barzelay et al., 2017).

Models can be derived directly from experimental data. Random harmonic stimuli (RHS) was played to marmoset harmonically-selective units and the spectral weights were inferred from a linear regression model (Feng & Wang, 2017); see figure 1.5. In some cases these weights formed a sieve-like pattern of excitation at integer multiples of Bf0 and inhibition at half-integer multiples (Feng & Wang, 2017). Similar results were achieved by modelling the response to RHS stimuli using a fixed number of Gaussian-tunes excitatory and inhibitory units (Feng & Wang, 2017).

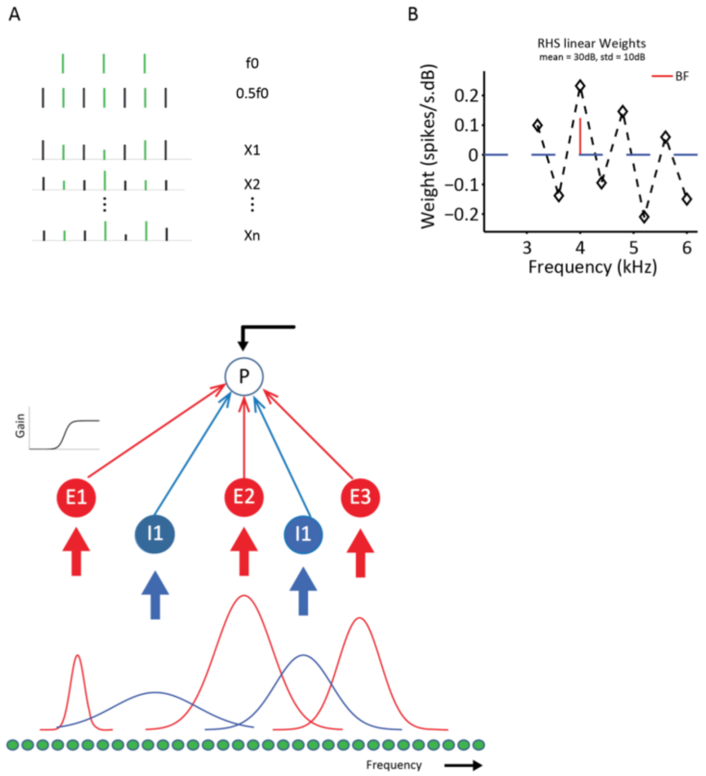


Figure 1.5: Modelling harmonically selective units with a spectral sieve. The sieve can be modelled as linear weights (above) or as a neural network with separate excitatory and inhibitory neurons with Gaussian receptive fields (below). The former is more relevant to this chapter. *Figure from (Feng & Wang, 2017).*

Because pitch perception in marmosets appears similar to humans (Song et al., 2016) a similar spectral sieving process is likely happening in humans as well.

1.6 Objectives of this dissertation

The two main objectives are developing models to elucidate harmonic processing as well as developing stimuli that are most useful in experiments. This breaks down into three main directions of exploration:

1: Learning harmonic selectivity from statistical structure

Neural computation evolved in an ecosystem and it thus must efficiently handle the statistics of said ecosystem. Can cortical harmonic selectivity can deduced from the statistics of natural sounds based on information-theoretical principles of coding? Applying independent component analysis (ICA) with harmonic sounds has been able to reproduce harmonic selectivity (Terashima & Hosoya, 2009), but this study used a Fourier transform rather than a subcortical model.

Although no comprehensive inferior colliculus (IC) model has been developed to feed in to the cortex, a generic, realistic model has been developed for the AN (Zilany et al., 2013), the ascending pathway is mostly mono-spectral all the way up to the marmoset IC (Kostlan, 2015; Pickles, 2013). Given that we want to use the realistic range of Q-factors and represent the information that the input network has, we decided to use an AN model instead of a Fourier transform.

2a: Learning harmonic selectivity with feedforward and recurrent networks

Subcortical models need to be explored in order to find out what model is best to feed into the cortex. Developing a comprehensive subcortical model is very difficult, but it is worthwhile to develop more limited models to help inform what simplifications must be made for feeding into the cortex.

Supervised learning models, unlike PCA or ICA, can be used to force networks to have harmonic selectivity. It is possible that a sieve is the most selective. However, the general problem is non-linear and there are numerous network topologies that are biologically realistic. Finally, real neural weights almost always obey a sign restriction known as Dale's principle in which neurons only output all-excitatory or all-inhibitory efferents, i.e the downstream synapses of a neuron are chemically alike (Eccles et al., 1954; Strata & Harvey, 1999). It is an open question as to whether a sieve is a robust pattern that survives these complications.

We focus on the spectral domain in order to focus our analysis on harmonic template units. To account for the limited resolution of the auditory system, we apply a Gaussian convolution to approximate the receptive fields of type I neurons in the IC (which are the most spectrally selective IC units).

Recurrent networks may make another mechanism of harmonic selectivity. They make up much of the cortex and can exhibit several behaviors such hysteresis, oscillations, and pattern completion/separation not found in feedforward networks (Khalil & Grizzle, 2002). Training is often done with some kind of Hebbian learning, of which there are many realistic models. These effects may be useful for some types of problems such as time-series stimuli with autocorrelation.

2b Analysis of recurrent networks by Lyapunov functions

Recurrent networks have rich theory that can be analyzed in the context of harmonic processing; various behaviors are discussed in (Amit & Amit, 1992). Any neural network with a Lyapunov function, of which discrete or continuous Hopfield networks are (Hopfield, 1982, 1984) or any symmetric network (Cohen & Grossberg, 1983) is guaranteed to settle into a stable equilibrium

(Dale’s principle introduces an asymmetric term but it is still possible to have a Lyapunov function in some cases). Lyapunov functions can be used to optimize problems such as the travelling salesman problem (Hopfield & Tank, 1985). Setting the Lyapunov function to an engineering objective function has been used as a control-systems tool (Ngo et al., 2005; Tee et al., 2009). For a recurrent network, we can reverse this problem and ask what objective function is solved by a harmonically-selective network, and how it relates to log-likelihood.

3: Optimal stimulus design

Stimuli almost always are tailored to experiments, and this process can be improved further with computational methods.

A simple method is to find the stimuli that maximizes the response of the neuron, in order to inform us about what the neuron’s receptive field is. Neurons get sparser at higher and higher levels in the auditory system (Pickles, 2013). The stimuli that best drive them get more and more complex, raising the question of how to find and describe them. The spike-triggered average for white noise has been used for the AN but it usually fails at higher levels (Jane & Young, 2000). Random spectral shapes have been used to deduce IC response properties (Slee & Young, 2013) but only give spectral information and explore only a small part of stimulus space. So-called sound textures have been developed that provide a relatively compact way of representing a rich repertoire of sounds: In (McDermott & Simoncelli, 2011) a *sound texture* is a collection of summary statistics. However, the model is slow and there the space is large, so improvement here is warranted.

If a model is available, stimuli can be designed to best reduce uncertainty in the models parameters. This can be extended to multiple models. This is different from optimizing response:

for a neuron with a BF at 1000Hz will response the most to a tone at 1000Hz but tones near the edges of the receptive field best reduce uncertainty about its BF. Optimal stimulus design has proven successful *in silico* for feed-forward networks (DiMattina & Wang, 2006) and two-unit recurrent ones (Doruk & Zhang, 2019). This is adapted to harmonic Hebbian networks.

CHAPTER 2:

Learning harmonic structure from natural sound statistics

2.1 Introduction

How are harmonic structures of natural sounds encoded in the auditory system? At low-frequencies the *pitch center* in the marmoset auditory cortex (close to the anterolateral border of fields A1 and R) contains *modulation sensitive neurons* and *pitch sensitive neurons*, the latter of which is sensitive to the temporal regularity of click trains (Bendor & Wang, 2005). The range of modulation-sensitivities can be described with simple integrate and fire models with synaptic depletion (Gao et al., 2016). However, the pitch-sensitive neurons lack a supporting computational model. Harmonic selectivity was also observed at higher frequencies in in the core region of marmoset auditory cortex (Feng & Wang, 2017), in which *harmonic template neurons* fired to a best harmonic stimuli at least three times as much as to when said stimuli is 50% mistuned and at least twice as much to any tone. This selectivity appears to originate from alternating excitatory and inhibitory receptive fields at different frequencies, according to a linearized weight model based on harmonic stimuli with per-component randomized intensities (Feng & Wang, 2017). Crucially, this can be thought of as a *weight vector* which tends to be excited by integer multiple frequencies of a Bf0 and inhibited by half-integer multiples of Bf0. A harmonic sound fits into this *spectral sieve* and optimally drives the neuron.

A natural question to ask is if the neurons, as a population of *weight vectors* can be derived from the stimulus statistics of natural sounds. It is likely that sensory neurons efficiently encode environmental stimuli (Simoncelli & Olshausen, 2001). Efficiency can be boosted through redundancy reduction (Barlow, 2001; Barlow & others, 1961). The optimal weight-vector population depends on what properties we care about most. Minimizing the RMS decoding error for a given number of units yields principal component analysis, but it cannot explain the Gabor-like simple cell receptive fields in visual area V1 (Field, 1994). However, Gabor-like filters can be obtained by a sparse coding strategy in which a constraint on the total activity or an L1-norm is added (Olshausen & Field, 1996). Similar results have been obtained by several other methods, including independent component analysis (Bell & Sejnowski, 1997; Van Hateren & van der Schaaf, 1998), which is related to a sparseness constraint (Olshausen & Field, 1996), stacked denoising autoencoders (Vincent et al., 2010) and direct optimization of Shannon mutual information (Huang et al., 2017).

Audition has seen similar success in using this principle to predict neural receptive fields. Independent component analysis has been used to reproduce auditory-nerve-like filters when trained on tooth-tapping sounds (Bell & Sejnowski, 1995, 1997). A related method that assumes a stimulus distribution $p(x) \propto \exp(-|x|^q)$ with the exponent inferred from the data (q<2 for almost all classes of natural sounds) reproduced gammatone-like filters when applied to natural sounds (Lewicki, 2002). The method has been extended to include the time domain in (Klein et al., 2003; Smith & Lewicki, 2006). In (Blättler & Hahnloser, 2011) sparse STRFs are demonstrated on bird songs, even when a given song is highly overrepresented in the training set (i.e. a self-vocalized song). This has been applied this to human speech in the gerbil IC (Carlson et al., 2012).

There are other paradigms besides sparsity. STRFs that maximize efficient coding were observed to change with respect to different stimuli (Zhao & Zhaoping, 2011), which help bridge the world of experimental STRFs with information-theoretically optimal models. The sustained response neurons in the IC could be modelled by looking for rewarding varying responses; the model produced sparse responses *despite* having no sparsity constraint or penalty in it (Carlin & Elhilali, 2013). Maximal causal analysis can be used to build spectro-temporal receptive fields from a collection of nonnegative *generative fields* and a sparse model was found to have a better fit to the ferret A1 (Sheikh et al., 2019). Finally, second-order STRFs can model responses to European starling songs in the starling caudo-medial nidopallium, which is loosely analogues to the auditory belt cortex (Kozlov & Gentner, 2016). Despite the success of these extensions and alternatives, sparsity remains a powerful tool for generating a population of neural responses from the stimuli.

The success of sparse coding extends to harmonic selectivity. Adding an L1-norm penalty to PCA created harmonic-selective receptive fields (Terashima & Hosoya, 2009) using Fourier transformed sounds, although only highly harmonic soundbanks (speech and piano) were able to produce multipeaked harmonic units (Terashima & Hosoya, 2009). However, a Fourier transform is only a simplified approximation of the auditory periphery that has a fixed frequency resolution bandwidth. Will this result still hold given a more realistic model?

We apply an auditory nerve (AN) model that includes realistic Q-factors and compression and two-tone nonlinearities (Zhang et al., 2001; Zilany et al., 2013), which were not considered in existing studies. In the existing models, some form of harmonic sensitivity might be implied in some units, but to examine whether it is sufficient to explain the biological data, they need to be systematically compared against the harmonic selectivity measured in auditory cortex. Inclusion of the AN model is justified because all auditory neurons downstream derive their inputs ultimately from AN responses, so it allows to analyze the harmonic sensitivity of the resultant learned models and compare them directly with known cortical neurons in marmoset auditory cortex.

In addition to the biological justification, there are reasons that these properties are relevant for ICA models of sound processing in general and harmonic processing in particular. ICA models such as (Lewicki, 2002) depend on the kurtosis of the sound statistics. However, the compressive non-linearity of the AN could reduce, or even *reverse* said kurtosis, which could change the ICA's behavior fundamentally. For the distortion products of the AN are at sums and differences of frequencies (Zilany et al., 2013), which could create harmonics in non-harmonic sounds, fill in missing fundamentals, or even interfere with the weaker components of a harmonic sound. For capturing any of this potential phenomenology, a quantitively accurate AN model is very important. The AN model used with a two-layer model (with a 1-norm activity penalty and a 2-norm reconstruction-error penalty) produced a variety of compact spectro-temporal receptive fields (STRFs) in the first layer and as well as excitation-inhibition second-layer features (Mlynarski & McDermott, 2017).

We consider the feasibility of producing harmonic selectivity given the finite tuning widths of the AN and the rectifying nature of the cortex. In sum, we will apply PCA and ICA to AN model outputs. We compare the results with existing neurophysiological data and make several experimental predictions with regard to harmonics sensitivity of auditory neurons.

Further realism can be accomplished by adding auditory nerve (AN) models. An AN model used with a two-layer model (with a 1-norm activity penalty and a 2-norm reconstruction-error penalty) produced a variety of compact spectro-temporal receptive fields (STRFs) in the first layer and as well as excitation-inhibition second-layer features (Mlynarski & McDermott, 2017).

2.2 Methods

Sound Stimuli

We have compiled banks of sounds from several sources. There are a wide variety of types of sounds available when making a sound bank and both the temporal and spectral structures should be considered.

Marmoset vocalizations

The three major marmoset call types (see Figure 2.1 for various sound examples) typically have a second harmonic at twice the f_0. The auditory cortex of marmosets appears to integrate spectral and temporal cues from their calls to form higher-level objects (Wang et al., 1995). We used an in-house marmoset call bank which was recorded from a marmoset colony in the Johns Hopkins University (Agamaite et al., 2015). Marmoset vocalizations were sampled at 50 kHz. There were 2292 "phee" calls (average 1.35 s long) with energy concentrated to the 7-9 kHz region, 1919 "trill" calls (average 0.51 s long) with energy concentrated around 6-8 kHz, and 1676 "twitter" calls (average 1.21 s long) with energy primarily in the 6-10kHz range. These vocalization samples also had significant natural background noise below 0.75 kHz.

Generic bank

We also have a generic soundbank that is a compilation of many different sounds from a variety of sources (Free Sound Effects Archive, 2016, http://www.grsites.com/archive/sounds/). The generic bank has 1108 sounds total and has a wide range of durations (4.85 seconds on average). Most of the sounds in bank were sampled to 11 kHz with a few at 22 kHz and 44 kHz. The bank covers a wide-range of sounds: bird calls, mammalian and other animal vocalizations, human speech and other sounds, industrial machinery, vehicles, musical instruments, and uncategorized "miscellaneous" sounds. Each category is approximately equally represented; many "uncategorized" sounds fit into one of the aforementioned categories. The generic bank has very few if any marmoset vocalizations (the individual sounds are not labeled). Most of the spectral energy is in the range of 0.02-5 kHz.

Vowels

We have a sound bank of vowels in human English speech (Hillenbrand et al., 1995). Each sound is a single vocalization of a word starting with h and ending with d with a vowel or vowel-with-transition in the middle, such as "heard". Many words are nonsense words such as "hawd". There are 540, 576, and 552 man, woman, and child voices respectively, and multiple different speakers. All sounds were originally sampled to 16kHz.

Harmonics, mistuned harmonics, and white noise

Finally, we also employ three types of artificially generated sounds. Each tone within a harmonic or mistuned complex has equal amplitude and random phase. Harmonic f_0 values are randomly distributed, uniformly on a log scale, from 50 Hz to 5000 Hz. Each complex includes between 2 and 25 components, which covers most harmonic sounds; 2 is the bare minimum and human vowels can have 40 substantial components; see (Schnupp et al., 2011) for a discussion on various sound stimuli. The only difference between mistuned harmonics and harmonics is that the former has a per-sound frequency shift of each component. The shift ranged from zero to the sound's f_0, uniformly distributed. White noise is modelled as a random soundwave where each element is sampled from a gaussian distribution with zero mean and a fixed variance; each element is independent from every other element. Sounds were set to 25 dB total energy except the harmonics/mistuned harmonics which were set to 25 dB per component.

Generating the training data

We have used two types of training data, namely, the direct training data of the raw sound waves themselves (despite the term “raw” we do allow resampling, but perform no other processing on them), and the AN training data. Samples from these banks are shown in figure 2.1 The natural sounds were resampled if necessary to make the sampling rate at 50 kHz for the direct data, and 100 kHz for the AN data, and all sounds were normalized to 25 dB. For the real banks, sounds were randomly selected for training, and windows were randomly selected within each sound; it was possible for overlapping sound-windows combinations to be selected but the fraction of overlapping windows was small for the generic bank and marmoset calls. The time interval of each sound snippet was always set to 10 ms, preceded by a 5 ms ramp to avoid artifacts in the AN model; no ramp was necessary for the direct training data. We always picked 10000 samples from each sound bank. The start time of each snippet was chosen randomly with a uniform distribution, and we allowed different snippets to have partial overlaps. The total window size was 100 seconds and the two banks had 5612 seconds and 6111 seconds of audio, respectively, for a coverage fraction of only around 1.7%, and the chance for overlapping windows was only about 3×10^{-4}. The coverage fraction was larger in the vowel bank as there were only 271 seconds of audio, but it is unlikely this significantly changed the overall pattern of the weight vectors. Each file was equally likely to be selected in the generic bank but the selection probabilities were weighted in the marmoset calls and vowels so that there was a 1/3 probability of being in each category. The synthetic sounds were time-invariant and required no resampling since they were generated at the same sample frequency they were used.

The AN training data involved more processing to generate. The AN model includes surround inhibition, saturation, and other nonlinearities (Zhang et al., 2001; Zilany et al., 2013). This computational model is written in C++ and integrated with MATLAB. Roughly speaking, the raw sound waves are passed through a gammatone filter followed by several non-linearities such as power-law adaptation, compression, and an output Boltzmann sigmoid model (Zilany et al., 2013). The raw sound waves were sampled at 100 kHz and fed into the model. For simplicity, no head-related-transfer-function was used here. The AN model was exposed to the given sound for a period of 30 ms prior to the time window that was extracted for training, allowing the response to stabilize, sounds were ramped for the first 5 ms of this padding. The expected spiking rate of a high-spontaneous fiber was used; no randomness or spike generation was in the model (we removed the randomness in the code). We used 240 model AN units logarithmically spaced from 0.125 kHz to 20 kHz, which was the range of best-frequencies allowed by the model. Then we linearly interpolated the results to create a total of 720 units. This interpolation, reduced the number of AN model computations we needed to perform and store but it only generated at most a 1.2% RMS error in the firing rates after the average firing rate was subtracted out. Finally, the responses were averaged over time within the 10ms long window; just as the raw sound waves only had a temporal dimension, the AN model only had a frequency dimension after the averaging.

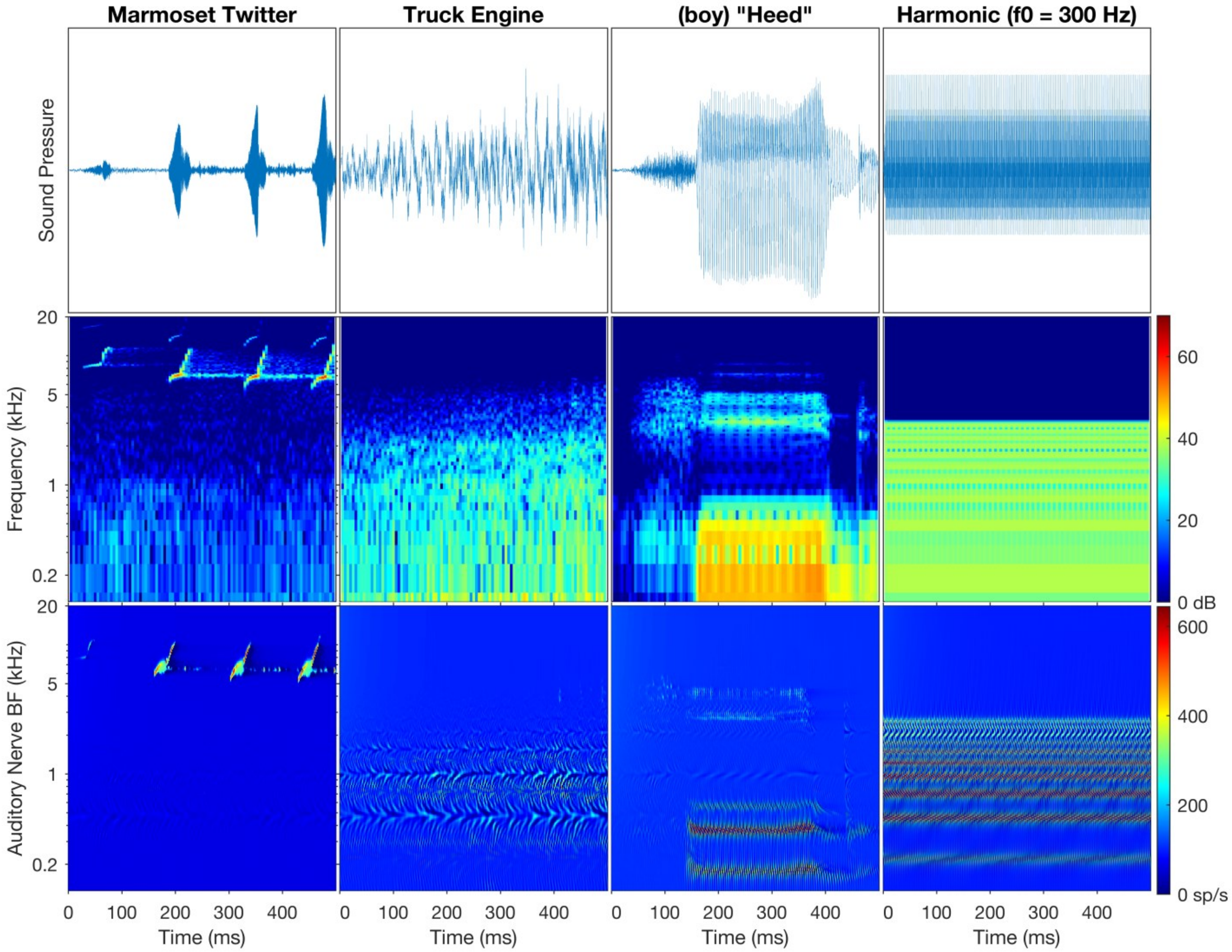


Figure 2.1: Examples of sounds used in the training dataset. This shows the raw sound waves (top), spectrograms (middle) and the cochleagram of the model auditory nerve's responses (bottom). The frequencies of the spectrograms are shown on log scale for easy comparison with the AN model bank. The spectrograms are at a time resolution of 10 ms and all sounds are clipped to 500 ms long in this figure. The AN model bank has good spectral resolution at low frequencies to clearly resolve the harmonics and also retain good time resolution at high frequencies to track the twitter's sweeps. The barking and engine sound are from the generic bank and the phee, trill, and twitter are from the marmoset dataset. The frequency scale is logarithmic and the colorbar is in dB for the spectrogram and spikes/s for the cochleagram.

Time windows

For both the raw waveform model and the AN model, we used a 10ms window since that was found to be near optimal. Longer windows demand more training examples because there are more degrees of freedom to train; under-training creates weight vectors with more noise added, longer windows require more computation power. Windows shorter than around 10 ms introduce window-dependent artifacts for the raw sound waves, and at shorter than around 5 ms the window size completely controls the weight vector structure (Figure 2.2). This suggests that there is strong temporal coherence around 5 ms that sets the scale of the weight vectors as long as the window size is significantly longer than 5 ms. This could be investigated further by examining autocorrelations. For the AN model we also used a 10ms window (not including the 30 ms of padding) but since time was averaged to emphasize spectral information, the window size was not nearly as important in determining the shape of the weight vectors.

At or above 10ms, the structure of the weight vectors no longer depends on window-size all that much (Figure 2.2), and the 20 ms case looks much like two 10 ms cases attached together and the 50 ms case looks like five 10 ms cases set side-by-side (not shown). Although this is only a visual analysis, there are other reasons to believe that 10ms is reasonable. The window used in (Lewicki, 2002) was a very similar 8 ms window on speech "because it covers the relevant time scale for a broad range of AN fibers" and "captures most of the short-range temporal correlations in the sound ensembles as revealed by the auto correlation functions". Thus 10 ms, which was also used in (Lewicki, 2002) and elsewhere in the literature, is likely a good window choice.

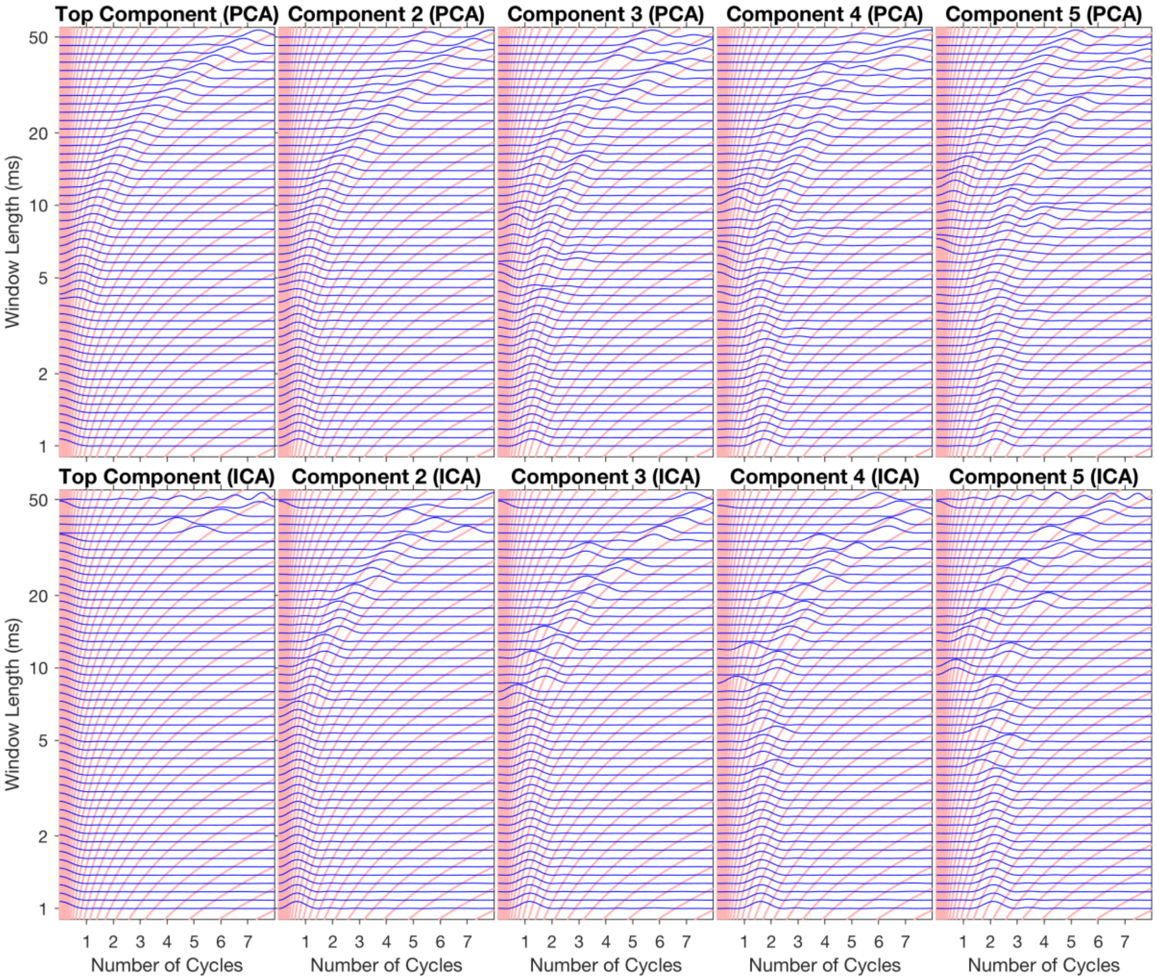


Figure 2.2: The effect of different window sizes on the trained weight vectors. They are trained on the raw waveforms in the generic bank. Shown is the zero-padded discrete Fourier transform of the 5 most important weight vectors in descending order of importance for various window sizes. The x-axis is the number of cycles and y-axis is spectral energy. On this plot a sinusoidal weight vector that completes a single 2π cycle within its window will show a "lump" of energy at x=1. Each pink curve represents where the spectral energy would be for a constant time-period sinusoidal column; a longer window will fit more cycles of a given time-period. Above 5-10 ms the curves tend to follow the pink lines (though weight vectors of similar importance may swap order), indicating nearly constant widths in the time domain; if a 20 ms weight vector were shown in the time domain it would look like two of the 10 ms time vectors side-by-side from Figure 3, indicating that little new structure is gained by moving from 10 to 20 ms.

Weight Training

The steps outlined above gave us a matrix of data X where each of the 10,000 columns of X is a single training instance. Here each column is a sound sample of 10 ms at 50 kHz for the direct wave case; for the AN model case, each column consists of 720 AN fiber responses described above, each averaged over 10 ms. The final outcome of the training is a weight matrix W of which each column is a *weight vector,* a single weight vector determines the input to a hypothetical cortical unit given a certain stimulus. We used PCA and a slightly-modified *fastinfomax* for our ICA. We apply the algorithm for both the ICA and PCA on the raw sound waves and a model AN bank (Zhang et al., 2001; Zilany et al., 2013).

For the PCA we simply used MATLABs PCA function on X^T. This gave us a series of vectors of descending fraction of explained variance.

For the ICA, many ICA algorithms have been proposed in the literature (Comon & Jutten, 2010; Himberg et al., 2004). We used *fastinfomax* (Huang et al., 2017) which tends to produce final results very similar to the infomax ICA (Bell & Sejnowski, 1995) but with faster convergence. We used *fastinfomax* on the square setting (equal numbers of inputs and outputs) with the default parameters except for a few modifications. The original algorithm enforces orthogonality of the iterated matrix C using Ghram-Schmidt. Here we used the formula $C_{ortho} = C/\sqrt{C^T C}$ (Horn et al., 1988). This avoided the preferential treatment that Ghram-Schmidt gives to the first few vectors and was found to be faster in MATLAB. After the orthogonalization is relaxed we used an even looser normalization criteria than that in the original algorithm: we simply constrained the Frobenius norm of C matrix to be n for an n by n matrix (i.e. what an orthogonal matrix's norm would be). This constrained only a single degree of freedom.

ICA generates a list of weight vectors that can be ranked much like PCA components. In the PCA case you rank vectors by the fraction of variance explained. The total fraction of variance explained by any subset of PCA vectors is the sum of the fractions of variances explained by each vector. This isn't the case with the ICA as the vectors are no longer orthogonal. However, we can still calculate an "importance" of each vector which is the amount of variance explained by that vector *alone*. This means *larger* weight vectors are *less important* because they need to amplify a smaller signal.

Our method also calculates the number of weight vectors (i.e the effective rank of X). The rank is computed by the minimum number k_0 of singular values needed such that the sum of squares of the largest k_0 values is 99% of the sum of squares of all singular values (Huang et al., 2017). Furthermore, the resulting penalty functions favor sparsity in activity by rewarding outliers (Huang et al., 2017).

Harmonicity Detection

We considered a hypothetical unit receiving input directly from the AN model bank with a weight vector from W. The output or firing rate of that cortical unit is

$$y = \max(0, w^T s), \tag{1}$$

where vector s is the responses of the auditory nerves to either tones, harmonics, or mistuned harmonics and w is the weight vector obtained by unsupervised learning. A rectification-linear model is realistic because cortical neurons are rarely driven near saturation (Douglas et al., 1995; Gutnick & Mody, 1995). Furthermore, they have faster learning than sigmoidal models so long as a global control on the firing rate is included (Glorot et al., 2011).

The maximum response to tones and harmonics was found by testing a grid of frequencies and then locally optimizing the frequency around the most responsive location to polish up the result. From this, the facilitation index and periodicity index were calculated as:

$$FI = \frac{y_{hm} - y_{tone}}{y_{hm} + y_{tone}}, \quad (2)$$

$$PI = \frac{y_{hm} - y_{mistune}}{y_{hm} + y_{mistune}} \quad (3)$$

where y_{tone} and y_{hm} are the best tone response and harmonic response, respectively, and $y_{mistune}$is the response to the best harmonic stimuli but mistuned by half of f_0. The FI is a measure of how much the response favors harmonics over tones and the PI is how much it favors harmonics over mistuned harmonics. Harmonic template units were defined in (Feng & Wang, 2017) as units with both FI>0.33 and PI>0.5. We applied these indices to our simulation results to allow direct comparison with the experimental data.

The stimuli are different from (Feng & Wang, 2017) due to our lack of a one-octave radius bound on the harmonics, but a comparison is still reasonable. In (Feng & Wang, 2017) the best-frequency of a neuron was found and then a one-octave radius bound around said frequency was used to restrict the components of the soundwave. However, the concept of BF is unnatural for multipeaked units: many receptive fields in both our study and (Feng & Wang, 2017) have two or more comparable peaks. Choosing a BF entails picking a peak just because it is slightly taller. Furthermore, many of the units in (Feng & Wang, 2017) had extremely little responses to tones, making a BF hard to find in the first place. Many of our results are fairly narrow anyway, so adding a one-octave bound wouldn't change in responses all that much. Finally, in an *in vivo* model there are animal comfort and broadband adaptation issues with presenting loud sounds, thus the desire to minimize the energy as much as possible. Although not perfect, removing the one-octave bound probably doesn't make a fundamental difference.

Population statistics

Each weight vector defines a cortical unit so the weight matrix as a whole can be considered a population. Thus, we can use the binomial and ranksum statistical tests to make inequality inferences about the different populations and how likely it is to find each unit type given different training conditions and use the two-sample Kolmogorov-Smirnov test to make inferences about differences in distributions.

2.3 Results

Training directly with raw sound waves in our datasets tended to produce filters that look like sinusoids; in other words, the resultant filters often resemble Fourier components (Figure 2.3).

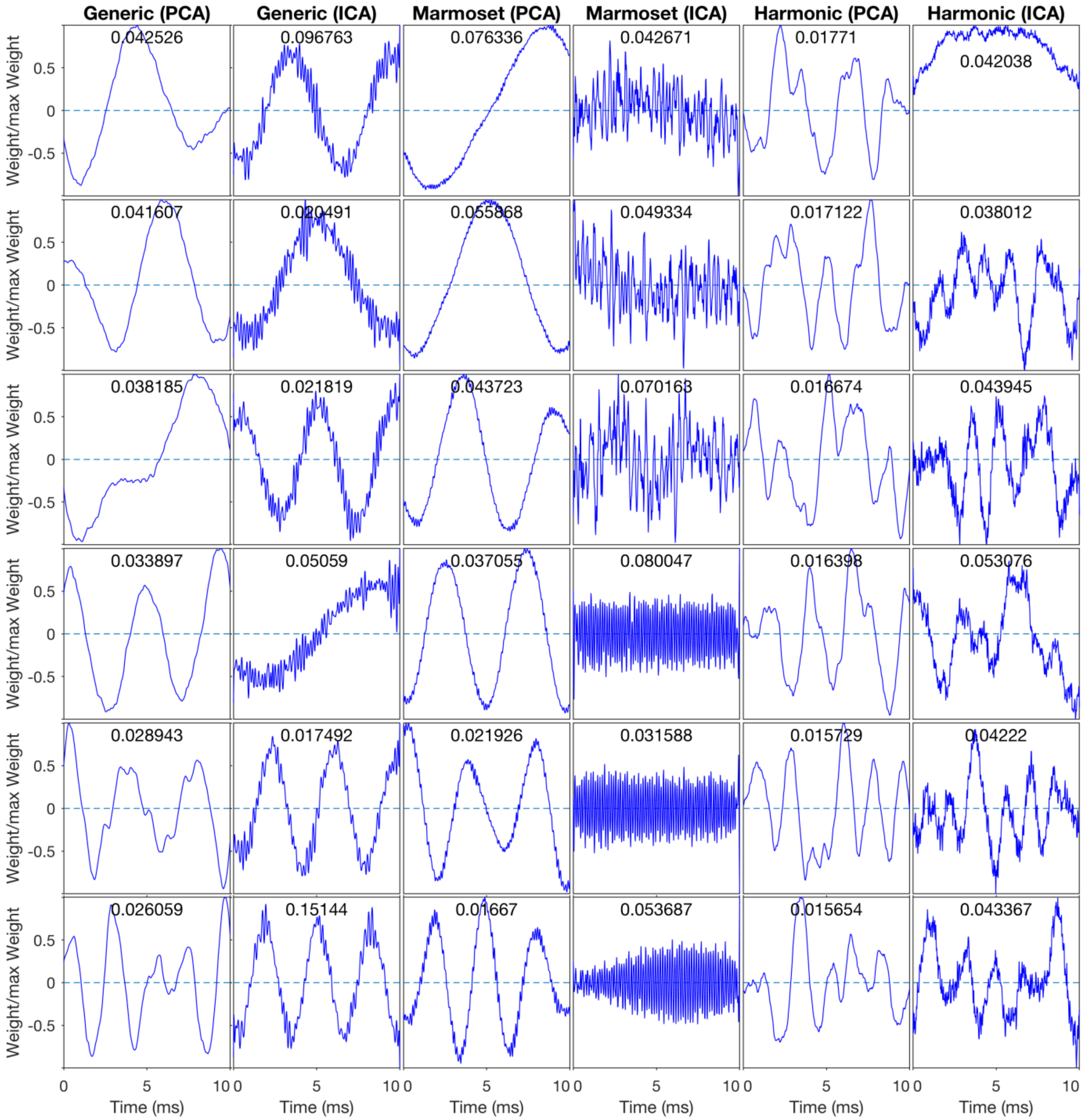


Figure 2.3: Top raw-sound weight vectors for three soundbanks. Shown is the top six PCA and ICA weights for the generic, marmoset, and harmonic soundbanks. As indicated by the numbers, the ranking was based on the fraction of variance explained by PCA or ICA. The y-scale is normalized. Sometimes the PCA weight vectors come in pairs of stimuli that are shifted by a phase, such as the first two marmoset examples. This is much like the SIN vs COS phase Fourier components.

From a mathematical point of view, this result is expected. The random sound sampling will mean that the expected value of the covariance between two points in time will only depend on the separation, i.e. for a random sound snippet x, $E(x_i - x_j) = f(|i - j|)$ where f is a one-dimensional function that encodes the auto-correlation. This means that the covariance matrix is a symmetric Toeplitz matrix because each diagonal is constant. A symmetric Toeplitz matrix's eigenvalues form a Discrete Fourier series (Makhoul, 1981). A perfect Fourier series for the PCA weight vectors would only occur given an infinite number of samples, but our sample size was still sufficient to yield nearly a Fourier series in most cases (Figure 2.3).

There are exceptions to this, however. The components from artificial harmonics or mistuned harmonics tended to look like each one has multiple frequencies rather than a single sine wave and the ICA components from marmoset calls have high frequency carrier that falls in the ~6-9 kHz range of the typical marmoset call dominant frequencies. Slightly further down the marmoset PCA weight vectors there is also a high-frequency carrier (not shown). Vowels are similar to the generic bank, and white noise produces almost-white vectors for both the PCA and ICA (not shown). None of our training produced wave shapes that were compact in time. The ICA and PCA methods tend to be similar, although the ICA has more noise. For the other training sets, the more important weight vectors tend to oscillate at lower frequencies. The generic bank components and first few marmoset ICA components are very much like a slightly-out-of-order Fourier series (see Figure 2.3). The rule-of-thumb that ICA tends to be sparser than PCA does not apply here.

Although the direct training sets never produced sparse weight vectors, this is not due to the details of *fastinfomax*. We were able to reproduce the gammatone-like filters reported by (Lewicki, 2002) when we applied *fastinfomax* to the same sound bank he used (See figure 2.4).

Thus the lack of sparseness should be due to the datasets we used which had a more diverse range of sounds.

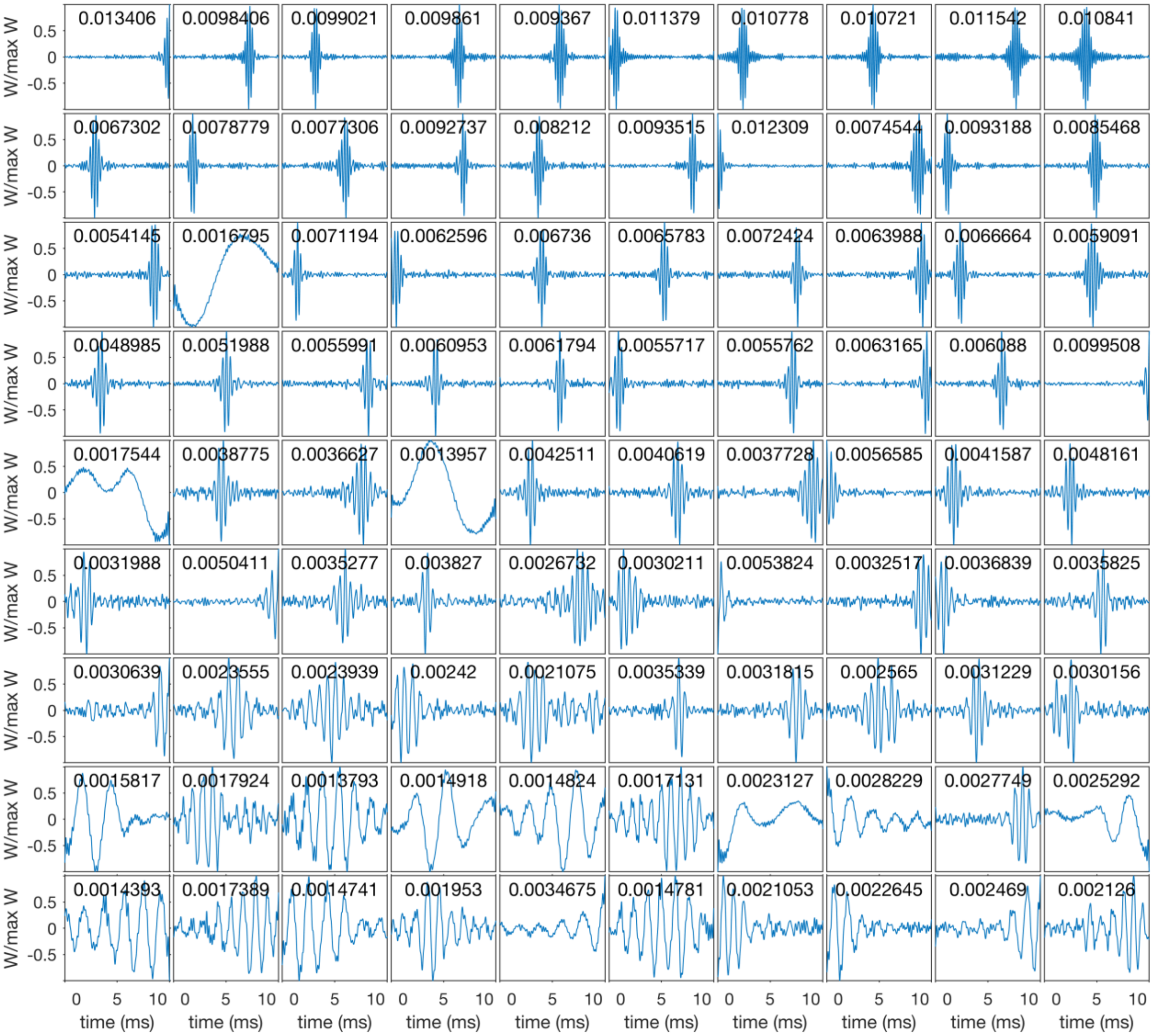


Figure 2.4: Weight vectors of hand-selected “non-harmonic environmental sounds”. This dataset was hand-selected in (Lewicki, 2002) and the weights are shown in their Figure 1a. Both our figure and his figure exhibit wavelet-like weight vector. Note that their Figure 1a is sorted by frequency while ours is sorted by our importance metric, yielding a different order. Other than that technicality the results are qualitatively similar.

To measure how compact a weight vector w_i is, we computed its width σ as follows:

$$\omega_{\mathrm{i}} = \frac{w_i^2}{\sum_{i=1}^{n} w_{\mathrm{i}}^2},$$

$$\mu = \sum_{i=1}^{n} \omega_i t_i \,,$$

$$\sigma = \left(\sum_{i=1}^{n} \omega_i (t_i - \mu)^2 \right)^{1/2}, \tag{4}$$

where *n* is the number of elements in the weight vector and t_i is the time that corresponds to the weight element w_i . This is essentially treating the square of the weights at each time t_i as a probability density function and computing its standard deviation. Weight vectors that are more spread out over a wider range of times will have a larger σ. These distributions are usually different from each other. All of the direct training sets ($1.59\times10^{-23} < p < 0.0334$, ranksum) except for white noise ($p=0.5$) were smaller for the ICA than the PCA, but the difference was not great. The direct training widths for all PCA and ICA cases averaged between 2.16 ms and 2.89 ms. The greatest difference was the marmoset calls, at a ratio of widths ICA/PCA = 0.78 ($p = 1.59\times10^{-23}$), but all other cases the ratio was between 0.9 and 1 even in cases where the statistics were stronger due to having a larger number of weight vectors. Despite the high significance, these actual differences were relatively small.

The training based AN data produced results qualitatively different from above. We find that the ICA-based model generates much more compact receptive fields which are not always more selective to harmonic sounds than the receptive fields generated by the PCA-based model. We can still ask to what extend the ICA is more or less compact than PCA, we simply use the log of frequency in place of time. The ICA weight vectors are much more narrower than the PCA, averaging across all banks only σ = 0.48 octaves for the ICA vs 0.88 octaves for the PCA

(Figure 2.5). The only exception to this rule is for marmoset calls where the ICA was slightly wider ($p = 0.0044$), but on everything else the ICA was narrower ($5.4\times10^{-25} < p < 1.5\times10^{-5}$). For mistuned harmonics and vowels the ICA/PCA ratio of compactness was 2.7 and 2.5, respectively. For the generic bank the ratio was 4.08 (0.27 vs 1.12 octaves). These large differences stand in stark contrast to the direct training sets (see Figure 2.5).

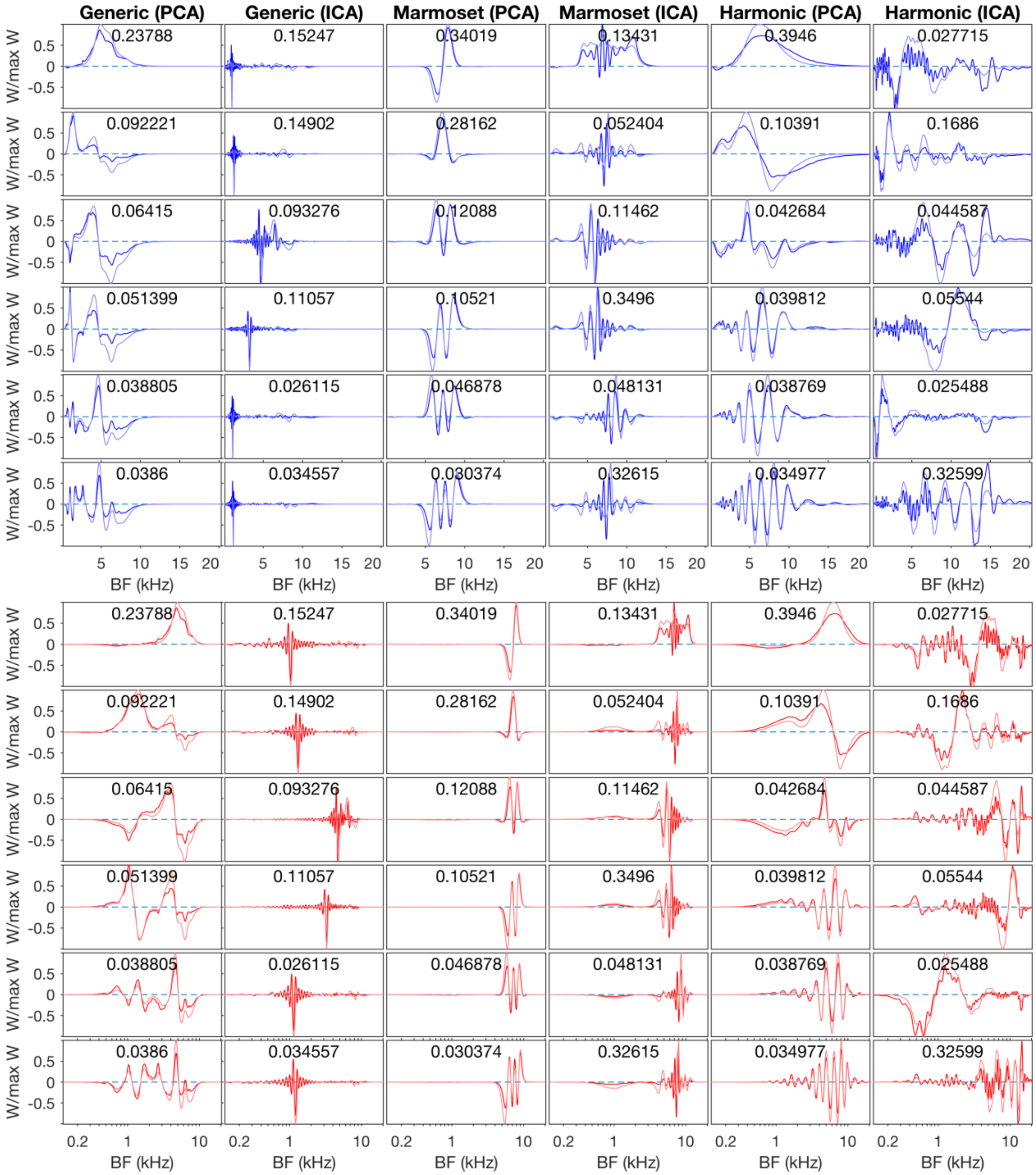


Figure 2.5: The top six weight vectors for PCA and ICA on the AN model. This is shown with both a linear scale (top) and log scale (bottom). The faded lines show the weighted response to a “sliding tone” such that at each frequency the curve is the dot product of the weight vector and the AN responses to said frequency. Below 500 Hz and above 10 kHz the AN responses begin to die down.

Sparsity

By regarding the filters obtained from training as a population of neurons, we examined their responses to natural or artificial stimuli. We expect that the ICA filters should produce sparser population activity because of the earlier results in the literature (Bell & Sejnowski, 1997; Olshausen & Field, 1996; Van Hateren & van der Schaaf, 1998). To examine whether this is the case, we performed the following tests. For each weight vector we have a distribution of responses to the sound bank that the vector was trained on; that is, the input for a single cortical unit is a random variable and each element of $w^T X$ is a sample of the input, giving us a distribution to work with; sparser responses should indicate heavier tails. Indeed, the ICA had greater kurtosis for almost all cases. This was highly significant by the rank-sum test ($p=0.00090$ for vowels, $1.06\times10^{-85} < p < 2.9\times10^{-5}$ otherwise) with only two exceptions. For the direct white noise case the response was forced to be a normal distribution of zero mean so getting any sparsity whatsoever was impossible. The AN marmoset case also failed to achieve significance ($p=0.052$). Often the ICA kurtosis was *much* greater than the PCA; the median kurtosis ratio was about 4 times in the AN generic bank and direct wave harmonic case and it was over 7 times in the AN harmonic case. When we added a rectification step, i.e. $\max(w^T X, 0)$, the statistical differences got a little weaker but remained significant.

The activity ratio in (Olshausen & Field, 1996) is an *inverse* metric of sparsity and makes sense only when we rectify to prevent negative responses. We see results that do not disagree with the PCA having a higher activity ratio (being less sparse). Compared with the kurtosis measure, the activity ratio for the direct sound wave model yielded compatible results (PCA>ICA, $10^{-300} < p < 0.00043$) but gave somewhat different results for the AN model cases: although the results were also consistent for the generic bank ($p=0.037$), for the marmoset calls, vowels and harmonics, the ratio was not significantly different from 1 (p=0.08, p=0.5, and p=0.121) and the white noise results even indicated that ICA was less sparse (p=0.026) although white noise is very unnatural and only serves as a control. However, lack of significance does not necessarily mean disagreement with the kurtosis results. The overall evidence still strongly indicates that the ICA filter's response distributions are sparser, and higher statistical significance was seen in the kurtosis measure than the activity ratio.

Harmonic selectivity

All training cases yielded harmonic-template weight vectors using the criteria of Feng & Wang (2017) as shown in equations (2) and (3). An example of a couple template-vectors is shown in Figure 5.

Both PCA and ICA can produce harmonic selectivity but the harmonically selective weight vectors show different patterns of excitation and inhibition. The Fourier-transformed direct training sets can be selective, but they did not yield large selectivity differences, except for white noise (which was *more* likely to have template units, calling into question the validity of this metric for the direct training sets). Furthermore, the direct sets did not represent harmonic waveforms in the time-domain.

The AN training cases, on the other hand, have template-vectors with sieve-like patterns, but they are modified as shown in Figure 2.7 (the pattern for the top 5 most selective units is similar for the top two shown in this figure). The sieves for the PCA cases are skewed to be approximately linearly spaced on a log scale, i.e. the lower frequency components are spaced more closely. The ICA sieves tend to be restricted, consisting of a relatively narrow range of harmonic numbers. The ICA harmonics case, on the other hand, produces many more nonchirped, broad sieves. This leads to far greater harmonic selectivity: the FI distribution of the ICA harmonic case, as well as the fraction of template units, was much higher than any other case ($1.2\times10^{-17} < p < 9.1\times10^{-6}$, see figure 2.8).

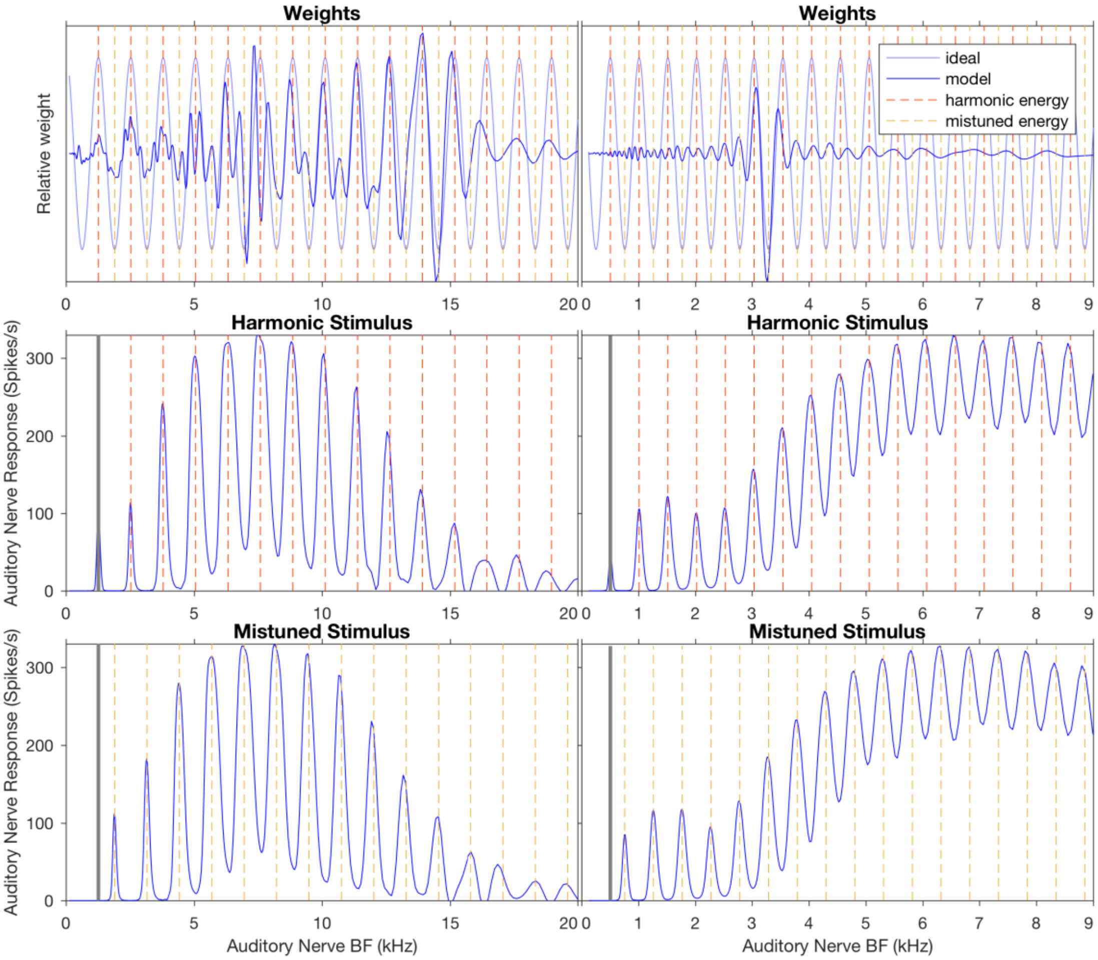


Figure 2.6: Examples of a harmonic template unit weight vectors. We have a strongly harmonically selective unit (left) and the center region of a slightly selective unit (right) obtained by ICA learning. The response of each unit is taken as the dot product of its weight vector (top panels) and the AN response pattern (middle and bottom panels) to a given sound, followed by rectification so that the responses cannot be negative. The thick vertical grey lines indicate each unit's Bf_0. The top panels show the weight vectors along with idealized sinusoidal "sieves" that are aligned to Bf_0. The middle panels show the responses of AN fiber array to an idealized harmonic stimuli at Bf_0, with the dashed lines showing the stimulus frequencies. The sieve's excitation regions match perfectly with the locations of peak AN responses so that each unit responds much more strongly than it does to tones. The bottom panels show that when a half-Bf_0 mistuning is added to the stimuli, the AN responses land on the inhibitory regions of the sieve so that each unit is inhibited. The unit in the left panels matches the sieve over several cycles and thus responds 8.6 times as strongly to harmonics at its Bf_0 than its maximal response to tones. The match of the unit in the right panels is only 2.4 times as much, barely exceeding the 2.0 criteria for a harmonic template unit, because the oscillations have approximate uniform spacing only on a log scale. Neither of the units responded at all to the mistuned stimuli because strong inhibition kept them below threshold.

These differences also make the ICA-harmonic template-vectors more selective to a single *Bf*$_0$ than the template-vectors in the other cases. The chirping or range-restriction means that each weight vector is selective to a wider range of *f*$_0$ values because the fit is not precise; this produces a multipeaked response to harmonics (not shown). However, in the ICA harmonic case we find selectivity to a single *f*$_0$ value (not shown).

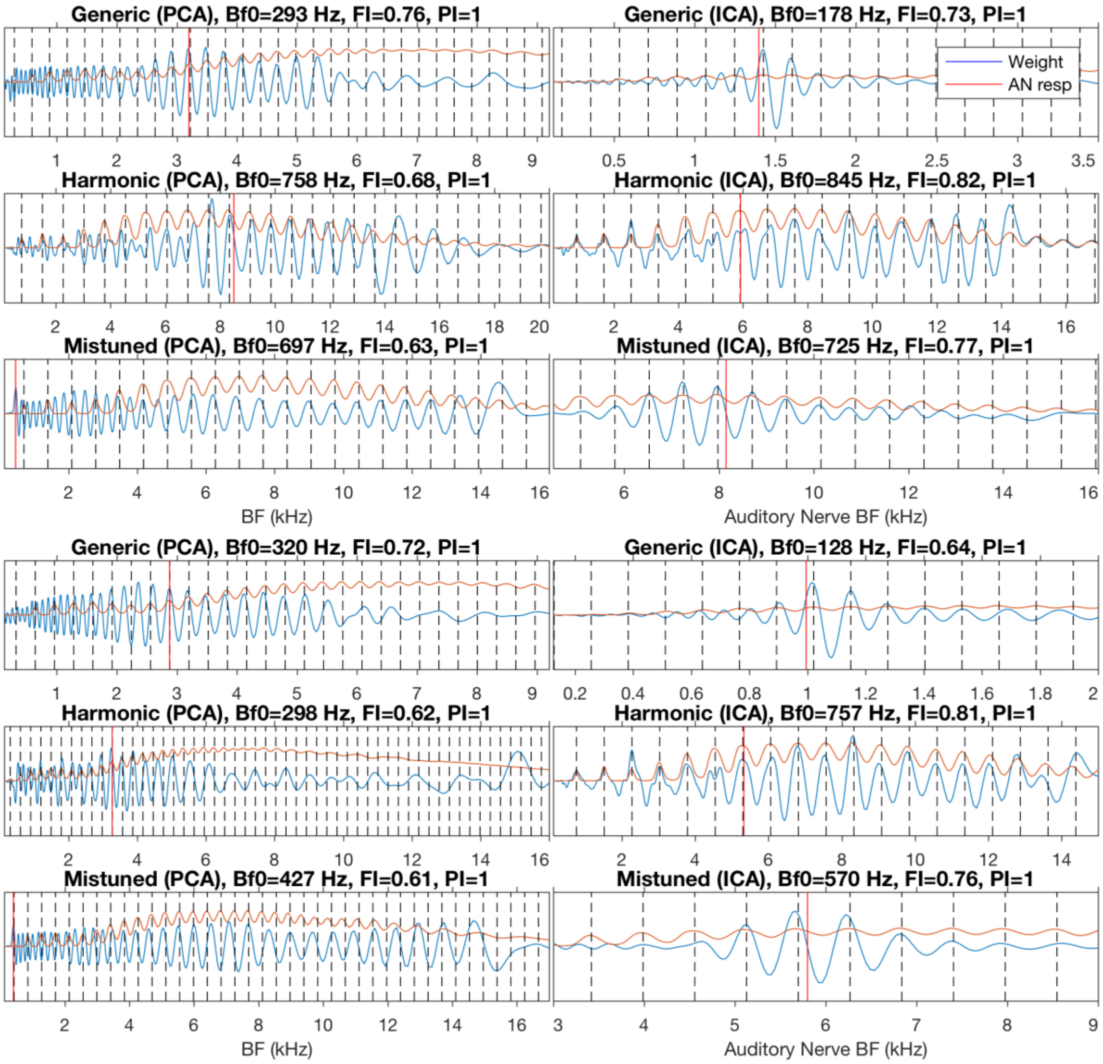


Figure 2.7: The most harmonically selective PCA and ICA weight vectors. Top half: The most selective vectors. Bottom half: The second most selective vectors. The PCA and ICA weight vectors are obtained with the auditory nerve model, where selectivity was evaluated by the minimum of the FI and PI indexes. Only the center regions of the weight vectors are shown. The vertical dashed lines mark multiples of *Bf*$_0$ and the vertical red line is the BF for tones. The blue curve is the weight vector and the red curve is the AN responses for a harmonic complex for a *fixed stimuli* at *Bf*$_0$ (not sliding as in Figure 4); the input to the cortex was calculated by taking the dot product of the two curves.

All training combinations having at least 10% of harmonic template units, but there were differences between banks (even beyond the case of ICA harmonics). Since we have 7 different banks we can make 21 pair-wise between-bank comparisons, which corresponds to a Bonferroni corrected p value threshold of 0.0024 for significance. Using this cutoff, the generic bank, harmonics, and mistuned harmonics were both more likely to have template units than white noise or the *in vivo* data ($1.5\times10^{-5} < p < 1.8\times10^{-3}$); the generic bank had 50% template units, vs 27% for in vivo. White noise was also less likely to get harmonic template units than vowels ($p=0.00085$). All other bank vs bank differences were insignificant ($0.007 < p < 0.5$).

We can also detect differences in the FI and PI distributions using the Kolmogorov-Smirnov test, comparing bank vs bank for all 4 combinations of PCA and ICA and FI and PI, also with $p<0.0024$:

- All *in vivo* cases were different from any bank ($1.8\times10^{-18} < p < 0.00059$), except for the marmoset FI (due to lower statistical power).
- White noise was different in 14 of the 24 bank vs bank comparisons across the 4 combinations ($8.5\times10^{-18} < p < 2.2\times10^{-4}$), but these differences were sporadic with no simple pattern.
- Harmonics were different from everything else in the ICA FI case ($1.2\times10^{-15} < p < 4.9\times10^{-5}$). It was different from everything except the marmoset (again, due to statistical power) and mistuned harmonic cases for PCA PI case ($3.1\times10^{-13} < 8.3\times10^{-8}$). Finally, it was different from white noise and mistuned harmonics in the ICA PI case ($p=3.74\times10^{-8}$, and 1.44×10^{-5}), as well as *in vivo* ($p=1.83\times10^{-18}$).
- Mistuned harmonics were also different from the bank and vowels for the PCA and ICA PI cases ($8.4\times10^{-17} < p < 8.4\times10^{-5}$).

- The only other bank vs generic bank difference was the vowels vs generic bank in the ICA PI case (p=0.0016).

Thus, the main result of all of this is that harmonic selectivity is very strong for filters trained by ICA with harmonics, but the in vivo data were better (though far from perfectly) matched by the other categories. These distributions are summarized in Figure 2.8 and Figure 2.9.

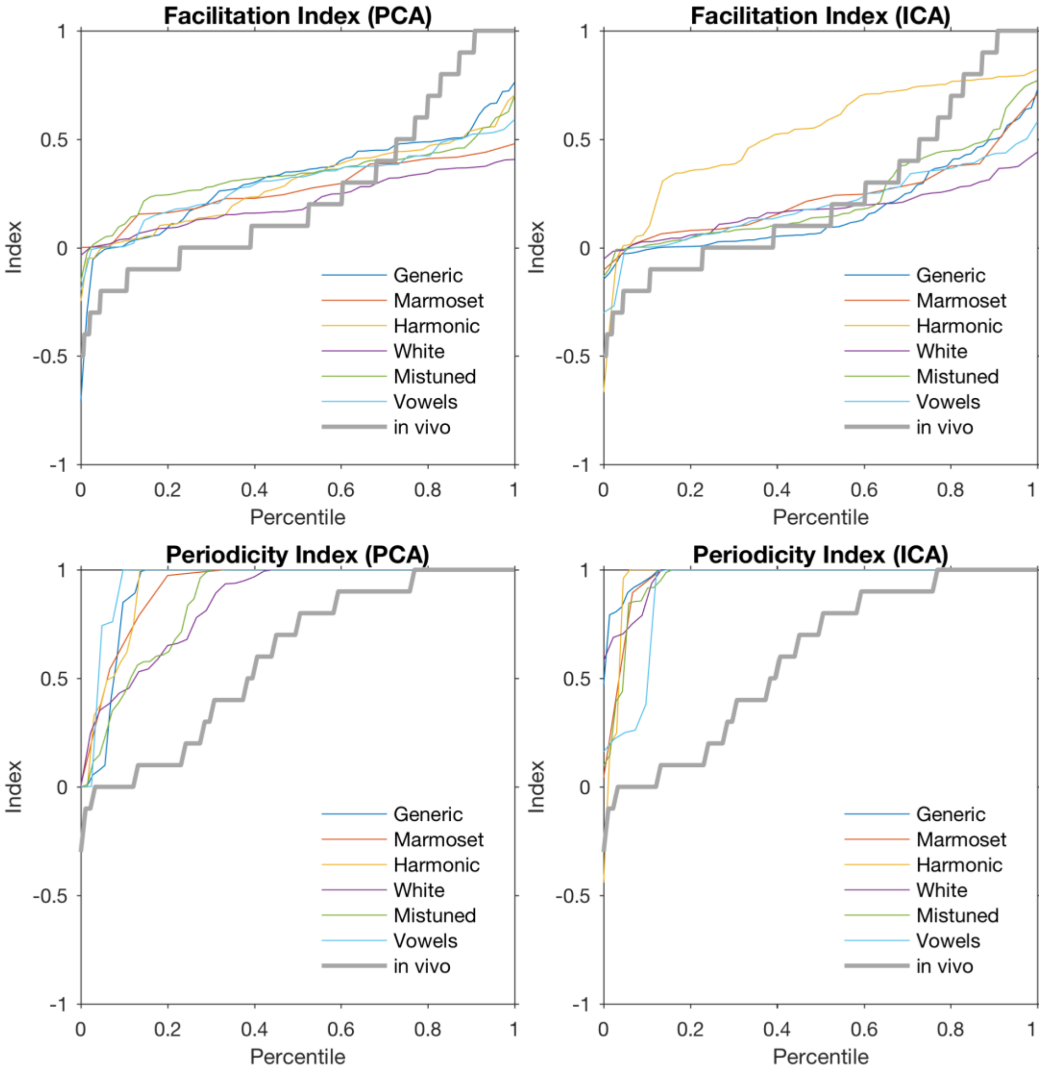


Figure 2.8: Cumulative distributions of the FI and PI. These are for AN models of various training sets, the x-axis is the cumulative probability. The ICA harmonics stands out as highly significant at increasing the FI distribution. The gray curves represent the *in vivo* experimental data (Feng & Wang, 2017). Compared with all the models, the in vivo FI has a broader range, and the *in vivo* PI lies below all the theoretical curves.

Since the results seem to lack the diversity or heterogeneity of the experimental data, we wonder whether adding variable threshold would produce more diverse units. We measured harmonic selectivity for a random non-zero threshold, as would be found in the cortex. For each weight vector w we calculated the standard deviation d of the cortical input given w and the training data (the standard deviation of $w^T X$). We then added a random normally distributed threshold with mean μd and standard deviation σd, where μ and σ represent a normalized bias and spread of thresholds. We tested a grid of μ and σ to find what parameters create the closest match to experimental distribution of PI and FI, based on average Kolmogorov-Smirnov test statistics. The greater kurtosis of ICA vs PCA in almost all cases was still the case with this non-zero threshold. We found that μ of -2/3 and σ of 10/3 gave the best match of all values tested, however the exact value didn't affect the results much and a μ of 0 and σ of 4 gave slightly worse test statistics but more realistic FI distribution results as shown in Figure 2.9. We conclude that just adding a variable threshold was sufficient to produce units with PI and FI distributions that were roughly comparable with the neurophysiological data, with it understood that we could also introduce variations in other factors to increase heterogeneity in the neural population.

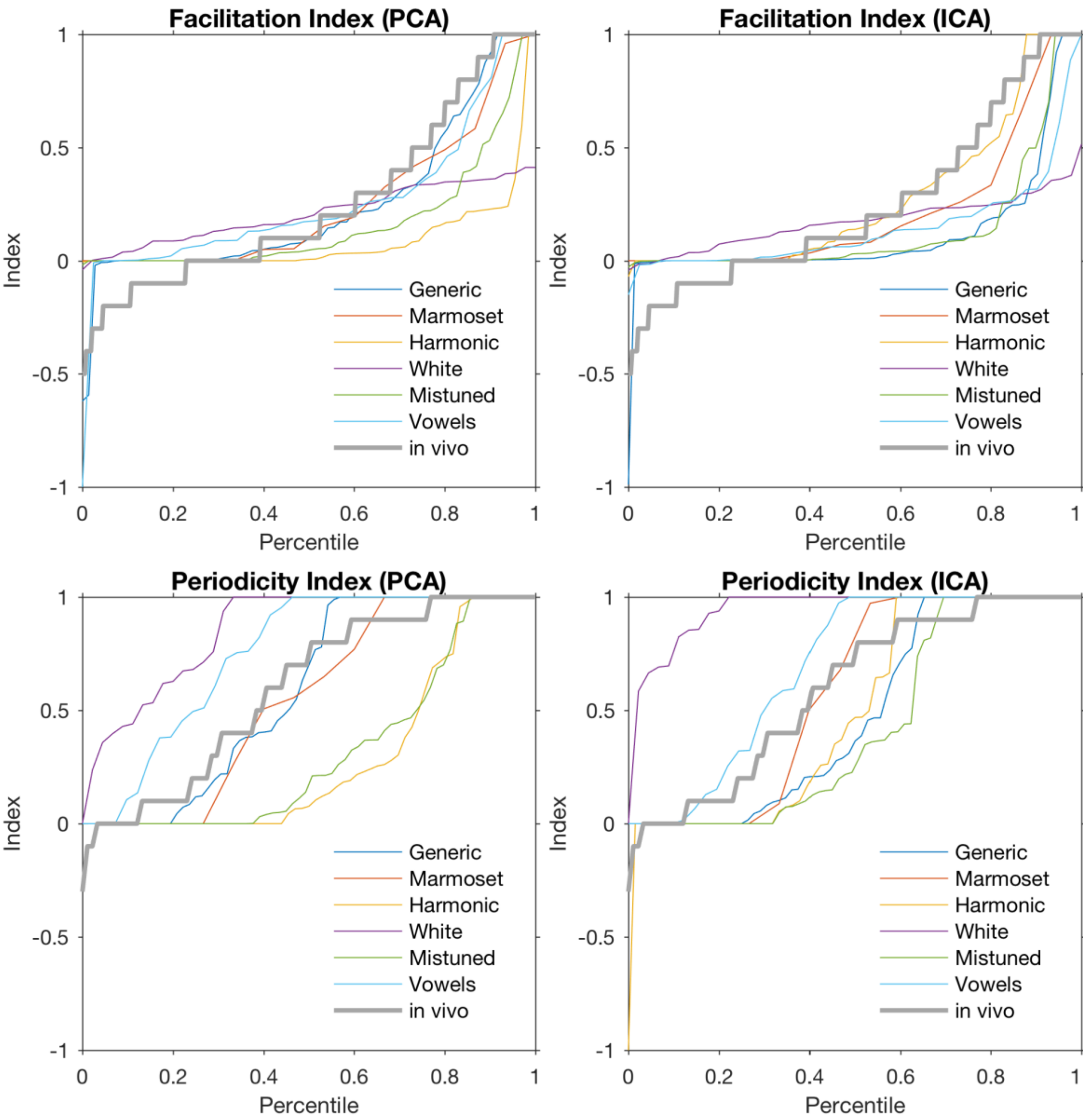


Figure 2.9: Cumulative distributions of the FI and PI under a randomized threshold. In this figure we use a relative $\sigma = 4$. The PI distributions are a much better match to the *in vivo* case than the zero threshold model.

2.4 Discussion

The goal of this study is to examine whether harmonic selectivity found in the marmoset auditory cortex could be accounted for by efficient coding of natural sounds that contain harmonic structures. As a control, we first performed PCA and ICA on the raw sound waves from multiple natural and artificial sound sources and found that the weight vectors tended to be broad and periodic, much like Fourier series (Figure 2.3). The exception to this was the case of marmoset vocalizations where the two methods picked up on the high frequency 6-9 kHz carrier. With a smaller dataset restricted to “ambient sounds” as used in (Lewicki, 2002), we reproduced compact wavelets from ICA learning. The sounds in that dataset have far less perceived pitch than most sounds in other banks, even less than most background-noise sounds. Also, they tend to be “crackly” such as fire, and in the time domain these sounds are similar to very compact random clicks with little other obvious higher-order statistical structure. This property likely enables ICA models to produce compact wavelets. However, with our much larger and more diverse datasets we found that both PCA and ICA training on raw sounds tended to produce unrealistic units for harmonic selectivity. Neither method produced sparse units, and the less important weight vectors often had more sinusoidal-like oscillations, although the ICA results sometimes were slightly narrower.

Adding a proper AN model allows for greater biological realism, incorporating effects such as Q-factors, two-tone suppression and power-law nonlinearities. Since our purpose here was to learn harmonic units based on spectral weight pattern, the overall role of the AN model was not fundamentally different from the weight vectors on the raw sound waves because they both served to approximately transform the sounds into the spectral domain. The PCA weight vectors for the AN model were similar to a windowed version of the direct wave, in that the components were sinusoid-like and less important components showed more oscillations (Figure 4). The windowing was determined by the location of energy in data X and acted to confine the weight vectors. However, the ICA model tended to form more compact wavelets that occurred in different places within the window, thus they only responded to a sub-region of X. For mistuned harmonics and vowels, ICA weight vectors averaged over twice as compact as the ICA results, and for the generic bank, they were four times more compact (as quantified by the width defined in equation 4).

Both ICA and PCA can yield realistic levels of harmonic selectivity

Despite the large difference in the AN-model-trained receptive fields, ICA and PCA generated similar ranges of the harmonic selectivity metrics FI and PI (equations 2 and 3), which were not too different from the ranges of selectivity observed in the marmoset when a randomized threshold was added to the model. The ICA results based on artificial harmonic sounds were an exception: we saw a very strong harmonic selectivity with FI values well above what was found in the cortex. However, vowels and marmoset calls did not yield high selectivity despite containing harmonic components. Marmoset calls rarely have more than 2-3 harmonic components, which is far fewer than most stimuli in our artificial harmonic training set. Also, the vowels have harmonics at a narrower range of f_0 values than our harmonic training set, and at much lower f_0 values than the best fundamental frequency (Bf_0) values of units trained on that training set. At these low frequencies, Q-factors are so low that picking out individual components becomes more difficult. Thus it appears that ICA can be very sensitive to harmonics under contrived conditions that rarely occur *in vivo*.

ICA with AN model shows more realistic tuning curve width and BF range

The efficient coding hypothesis states that neurons optimize their firing distributions to maximize information about the stimulus (Barlow, 2001; Barlow & others, 1961), but how does it apply in the marmoset auditory cortex? The *fastinfomax* ICA algorithm we use is based on maximizing Shannon mutual information, and does it do a better job at predicting cortical response than simply decorrelating the stimuli as does PCA? However, the FI and PI harmonic selectivity metrics do not show much difference between ICA and PCA, except for harmonic stimuli where the ICA's high FI and template unit proportion is way above *in vivo* levels. Thus these metrics are not useful at differentiating PCA and the efficient-coding-based ICA.

Cortical tuning widths and distribution of BFs seem to mirror the generic bank ICA behavior; the generic bank with its wide range of sounds is the most realistic training case. Like our results, the cortical units tend to be relatively compact with usually 1-3 peaks (Kadia & Wang, 2003). This is in contrast to the broader and greater number of peaks found in the PCA case. However, a more careful scrutiny with a variety of quantitatively shape metrics needs to be done in the future to conclusively determine whether harmonic selectivity in the auditory cortex is PCA or ICA-based, or neither.

Biological relevance of PCA and ICA learning

We have compared PCA, which decorrelates the data but does not remove statistical dependency beyond the second order, with an ICA algorithm based on (Huang et al., 2017), which produces results with comparable quality to the infomax ICA (Bell & Sejnowski, 1995), but it is numerically more robust and does not require equal input and output dimensions while still being as fast as Fast ICA (Hyvarinen, 1999). Although we did not explicitly model how the learning takes place in the auditory system, both PCA and ICA can potentially be implemented by biologically plausible learning rules. PCA is essentially a Hebb rule whereas the infomax ICA is essentially an anti-Hebb rule (Bell & Sejnowski, 1995). Hebbian learning produces PCA-like behavior (Oja, 1992). Adding a convex non-linearity to Hebb's rule means unusually strong stimuli will dominate the plasticity and cause sparseness (Brito & Gerstner, 2016). Biological mechanisms can potentially implement approximations to ICA. Volume transmission feedback (Zoli et al., 1999) can regulate and stabilize the global activity.

Relation with harmonic selectivity in auditory cortex

As mentioned in this Chapter's introduction, both ICA and PCA can potentially be realized with a local learning rule combined with a global activity regulator that stabilizes the overall firing rate. In the brain, the wiring and synapses determine local connections while diffusion of neurotransmitters outside of the synapses affects neuronal activity on regional or global scales (Agnati et al., 1995). In an artificial neural network, the local Hebbian learning rule finds the components and the global term keeps the overall activity from exploding to infinity (Oja, 1992). Hebbian learning will pull weight vectors toward the largest eigenvector, extracting the largest principle component (Oja, 1992), or recursively extracting smaller and smaller components (Sanger, 1989). For ICA learning one possible local rule adds a cubic term that amplifies the Hebbian learning at large firing rates, rewarding a heavy-tail (Hyvärinen & Oja, 1997).

The sparsity inherent to brain activity is more consistent with ICA-like behavior rather than PCA which finds the linear transformation that minimizes the square error (Linsker, 1988). ICA but not PCA can reproduce Gabor filters in vision (Olshausen & Field, 1996) and gammatone filters in audition (Lewicki, 2002). Nonlinear multistage compression algorithms in the form of a deep neural networks have proven very useful in image and sound recognition in the machine learning field (LeCun et al., 2015), but these do not have to resemble the cortex with sparse activity. For more complex tasks that the brain accomplishes, it may be the case that sparse responses are a much better representations for performing complex tasks beyond V1 (Olshausen & Field, 1996), such as for responding to stimuli as specific as an individual's face or voice.

Our model predicts that the harmonically selective neurons could be made of excitatory and inhibitory weights arranged as a sieve with even spacing in the frequency domain on a linear scale; such sieves are selective to harmonics over tones and mistuned harmonics (Figure 2.6). This model is consistent with the finding by the simulations in (Terashima & Hosoya, 2009) and the experiments in (Feng & Wang, 2017) who found similar weight vector patterns by fitting a linear model to the responses elicited by random spectral stimuli. This interleaving excitatory-inhibitory pattern of connections could potentially be tested in future anatomical studies. Our model is also able to make testable predictions at the population level. The statistics of harmonic neurons such as the sparsity of population activity in response to a given stimulus set as predicted by the model could potentially be tested in real auditory neurons.

Although our model neglects all the intermediate steps before the cortex, those steps rarely integrate multiple spectral bands, they instead build up a single band from a narrow range of frequencies and integrate other properties such as spatial location (Pickles, 2013). As our model focuses on spectral integration, it probably corresponds to the step from the thalamus to primary cortex most closely. In the future our work could be extended by incorporating realistic thalamic response properties.

Other models of harmonic selectivity, such as the temporal periodicity detection in (Bendor & Wang, 2005), could complement the spectral sieving. This could be tested by training a time-domain model on click trains with various periods and amounts of jitter, as the pitch-sensitive neurons in (Bendor & Wang, 2005) responded less strongly to jittered click trains. Even with complex spectral-temporal training paradigms, ICA would likely end up being more compact in whatever space the weight vectors are in than PCA.

Finally, we mention several factors that could explain the remaining discrepancy between model and experimental data in the FI and PI distributions. If the PI distributions are matched by a randomized threshold the FI distribution *in vivo* remains broader than the distributions on any of the training sets (Figure 8). The actual distribution of the *in vivo* thresholds (i.e. the effective thresholds that result from the very complex underlying neural network) is unknown. We choose a normal distribution for its simplicity and symmetry and then find the best σ; it also was convenient to express σ as a dimensionless ratio with respect to the range of inputs.

Technical difficulties with experimentation such as electrode-induced property changes, baseline drift, and adaptation could add noise and broaden the distribution. Also, the resolution limits of BF collection could make it hard to find as precisely the BF and Bf_0. The stochastic nature of Poisson-like spike trains would introduce additional randomness which was not accounted for in the simulation. The *in vivo* responses used a one-octave radius cutoff around BF, which was deemed artificial for the neurons that were multipeaked. The use of this cutoff likely does not make a large change, but it conceivably could have softened the *in vivo* PI distribution since there would be less components acting together to build up a strong inhibition when mistuned.

CHAPTER 3:

Training harmonicity selectivity in feedforward neural networks

3.1 Introduction

In Chapter 2 we showed that summary statistics can produce harmonically selective units in a similar proportion to what was found in the marmoset cortex in (Feng & Wang, 2017). Here we reverse this question: if harmonic selectivity is desired, what do the weights look like for feedforward cortical networks?

The subcortical levels are monospectral

The peripheral auditory system begins in the cochlea, where sounds are broken down into a frequency-time representation (Pickles, 2013). A harmonic sound that is resolved will be separated in the frequency domain, while an unresolved sound will produce temporal repetition in the firing pattern of the auditory nerve (AN) fibers (Bernstein & Oxenham, 2003; Bidelman & Khaja, 2014; Carlyon & Shackleton, 1994). Throughout the auditory system is a tonotopy, which is mechanically defined in the cochlea and propagates all the way up to the cortex. The system has interaction between nearby isofrequency laminae at all levels.

The inferior colliculus (IC), the main processing center below the cortex, integrates nearby spectral bands with center-surround patterns; with a strong non-linearity that complicated predicting complex spectra responses from pure tunes (Ehret & Merzenich, 1988). The IC also has neurons sensitive to modulation frequency, with bandpass neurons as well as other response types; it is likely that these neurons form an axis perpendicular to the tonotopy (Langner et al., 2002). Modulation is potentially a mechanism to detect pitch that can complement raw frequency mechanisms. However, the ascending IC neurons in the Marmoset have not been found to integrate widely-separated sounds (Kostlan, 2015).

Psychophysical evidence bolsters this “monospectral subcortical performance” picture. There is degraded harmonic discrimination performance in hemispherectomy patients (Zatorre, 1988) or other disabilities *without* a loss of tone discrimination ability (Whitfield, 1980). Given the relatively monospectral nature of the IC and below, we will assume that the cortex is where harmonicity is constructed.

At the cortex, it is unclear the mechanisms behind developing such selectivity in the first place, i.e. if it can be explained through a learning model. As shown in Chapter 2, it is possible to train a degree of harmonicity using summary statistics, namely independent component analysis, that produces roughly similar harmonic selectivity population histograms to the marmoset cortex as found in (Feng & Wang, 2017). It is also reasonable that a supervised learning model will produce sieve-like selectivity. However, given the wide range of possible network topologies, nonlinearities, local minima in training bio-algorithms, the choice of the training set, Dale’s principle, and the performance metric(s) it is unclear whether a sieve is a universal property of harmonically-selective networks.

Can temporal cortical models be used?

The cortex also responds to temporal periodicity: some neurons are modulation sensitive with various synchronization and best modulation frequency patterns. As shown later in this chapter, these can be explained with integrate and fire models modified to account for synaptic depletion. More relevant to producing a fine-tuned pitch perception are the so-called pitch sensitive neurons at low frequencies that are very sensitive to regularity in click trains (Bendor & Wang, 2005, 2010). A simple model of these is elusive, however. Long short term memory (LSTM) networks can learn precise timing (Gers et al., 2002) but it is questionable as to how relevant LSTM is to biology as well whether it could act quickly enough. Indeed, better micro circuitry mapping is needed before we have locked down enough degrees of freedom to build models that are reasonably likely to be qualitatively correct, thus we will focus in the spectral domain.

Is a sieve trivial?

It is likely that training a feed-forward network selective to harmonics makes it a spectral sieve, but given the nonlinearities, considerations with deep neural networks, Dale's principle, and different training options as well as constraints a sieve isn't the only possibility and thus should be verified.

3.2 Methods

Spectro-temporal subcortical models

We explored trying to build a broad spectral-temporal model of the subcortical parts of the auditory system. The timestep at all levels is 0.01ms.

One recurring theme is lowpass filtering during subcortical processing. We make heavy use of a *first order lowpass filter* which is equivalent to an RC circuit and has an impulse response $K \sim \exp(-2\pi t F_c)$ where F_c is the corner frequency and the filter is discretized and normalized to sum to one (unit gain at zero frequency).

For each unit above the AN, it is possible to specify the threshold and best frequency in addition to the type of the unit.

From air to eardum:

The first step of converting sounds into percepts is the acoustics of getting the sound to the eardrum. We used the experimentally measured head-related-transfer function for the Marmoset from (Slee & Young, 2010). We used the HRTF in for the superiorly olivary complex trials, otherwise we didn't apply any HRTF model because sound localization wasn't crucial in those cases. The system is linear, so the time-domain HRTF of the nearest azimuth and elevation to the desired stimulus is convolved with the input stimulus to calculate the eardrum waveform.

The auditory nerve model:

We developed a simple auditory nerve model which provides an alternative to (Zilany et al., 2013). Our model isn't "better", it is comparable in computational speed and accuracy. However, our model is more easily controllable being written in pure MATLAB and includes most of the same features. Also, it allows us to see which features are needed to best simulate the actual auditory nerve.

We define a spectral gammatone filter which is then convolved with the sound waveform, which is commonly used in the AN, see (Katsiamis et al., 2007):

$$K = t^3 \exp(-2\pi E_{RB} t) \cos\left(2\pi \beta B_f t\right),\ E_{RB} = 0.0247\alpha \left(\left(4.37 \frac{B_f}{Hz}\right) + 1000\right)$$

This kernel is then normalized to have unit gain for the ideal-tuned sinusoid. We calculate this for $\beta, \alpha = \{1,1\}$ and $\beta, \alpha = \{0.7,2\}$. The former is used for quieter sounds than the latter, namely we use a linear combination of the two convolutions based on the envelope's amplitude (with "full loudness" defined at 100 dB). The envelope is calculated by a first-order band-pass filter of the sound around BF. This gives us a nearly-linearly filtered signal.

We then calculate the envelope of said nearly-convolution (Hilbert transform). We need to compress said envelope, so we define a sigmoid mechanical compression function that far from zero it behaves like the p-th power and near zero is linear:

$$g_p(x,p) = \sinh\left(asinh\left(\frac{x}{p}\right), p\right)$$

We then map the envelope to mechanical motion with:

$$g_m(x) = x_{on} g_p(x/x_{on}, p)(1 - g_L) + x g_L, x_{on} = \sqrt{10}^{3}, x_{off} = \sqrt{10}^{10}, g_L = x_{on}/x_{off}, p = 0.32$$

This function represents outer hair cell sound amplification in the cochlea, which was first discovered by (Gold, 1948). It is the identity at small levels (full linear amplification regime) but gets a compressive nonlinearity around x_{on}. Finally, it becomes linear again after x_{off}, which represents very loud sounds overwhelming the amplifier and the cochlear again becomes a linear system.

Call x_{filt} the result of the "almost linear" convolution. Call H_m the envelope that is compressed through the g_m function. The *total mechanical motion* is the fine-structure of x_{filt} combined with the envelope of H_m. The formula for this is $r = x_{filt} H_m / envelope(x_{filt})$. For a tone at Bf at zero dB, the motion would have an root-mean-square of 1.0 and the amplifier would be at maximum gain.

We calculate the *normalized hair-cell current* with a sum of two sigmoid functions:

$$I_{hair} = \frac{1}{2}\left(g_s(k_0(r - b_0)) + g_s(k_1(r - b_1))\right), \qquad g_s(x) = \frac{1}{1 + \exp{(-x)}}$$

The gains and thresholds on the sigmoid depend on how sensitive the unit is:

$$b_0 = lgt(14.25, 4.75), b_1 = lgt(61.75, 4.75)$$

$$k_0 = lgt(0.278, 0.9), k_1 = lgt(0.06, 0.9)$$

The log-interpolate function is:

$$lgt(x, y) = \exp\left(\log(x)(1-\alpha) + \log(y)\,\alpha\right)$$

With α continuously varying from zero (for the low spontaneous fibers) to one (for the high spontaneous fibers). The normalized hair current ranges from zero to one.

We apply a low pass filter to calculate the *normalized depolarizing current*:

$$I_{exc} = K_{low1} * K_{low2} * I_{hair}$$

Each low-pass filter is a first order filter with a corner frequency of 1400 Hz and 2400Hz, respectably.

The *un-adapted firing rate* is an approximate leak, integrate, and fire function (the actual implementation uses additional code to prevent numerical overflow):

$$R_{unadapted} = \ln\left(1 + \exp\left(\frac{I_{exc}}{I_{leak}} - 4\right)\right)$$

We use an adaptation model which calculates a "fatigue factor", again with first-order low-pass filters:

$$F_F = \sum \mathrm{w_i} \mathrm{K}_{\mathrm{low},\tau_i^{-1}} * R_{unadapted};\ R_{adapted} = \frac{R_{unadapted}}{1+F_F};\ w_i = \{1, 1.4, 2, 2.8\},\quad \tau = \{1, 4, 16, 64\}ms$$

Finally we multiply by a global rate of into spikes per second: $R = R_{adapted} 2300 \frac{sp}{s}$

No neuron can fire anywhere near that quickly, but transients in the peri-stimulus-time histogram *can* be higher for extremely brief periods of time.

The auditory nerve model can reproduce the fundamental properties of the auditory nerve's response to tones. In terms of temporal response, there is strong phase-locking up until just over 2kHz, which is shown in figure 3.1. By 5kHz, the phase-locking is very weak as shown in figure 3.2. These match the auditory nerve phase-frequency sensitivities in (Zilany et al., 2013).

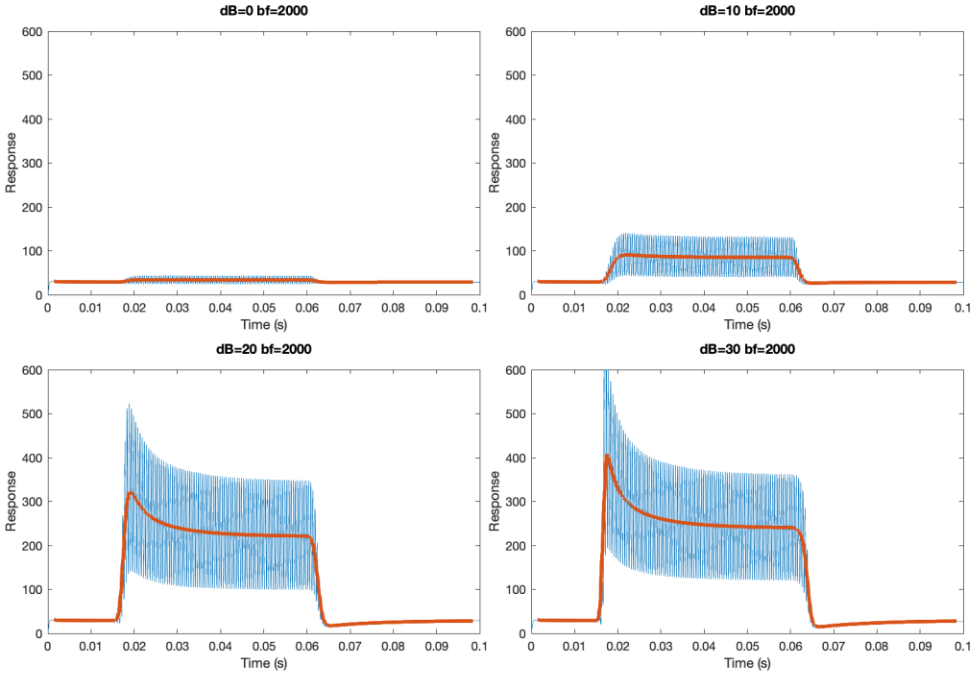


Figure 3.1: PSTH's of an AN fiber unit at 2000 Hz. Phase-locking is strong up to and including this frequency.

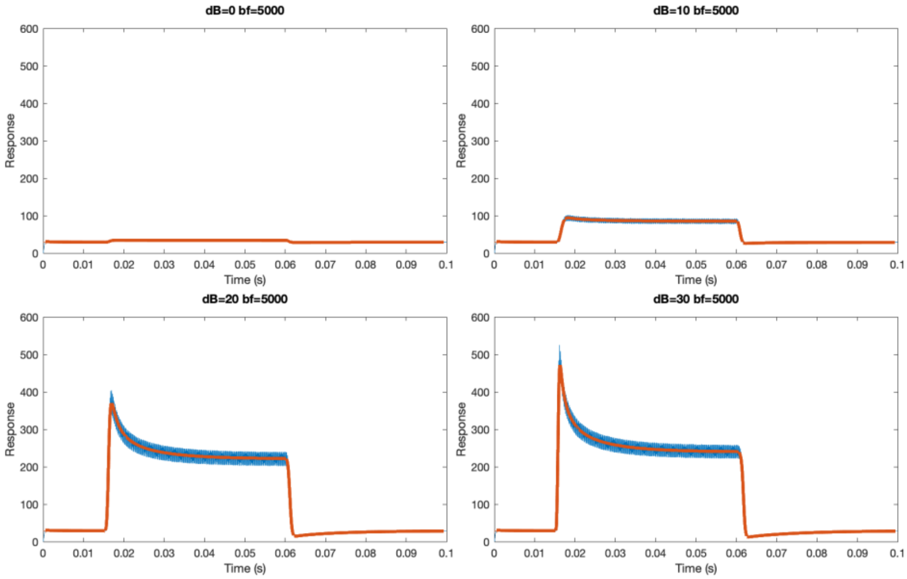


Figure 3.2: PSTH's of an AN fiber unit at 5000 Hz. Phase-locking is much weaker than the 2000Hz case.

The auditory nerve model reproduces the total average response in (Zilany et al., 2013) to tones of various frequencies. At BF there is a sigmoid curve of the response to tones of various dB shown in figure 3.3. The response map of a single unit to tones with different frequency and dB shows the classic skewed-V response, shown in figure 3.4.

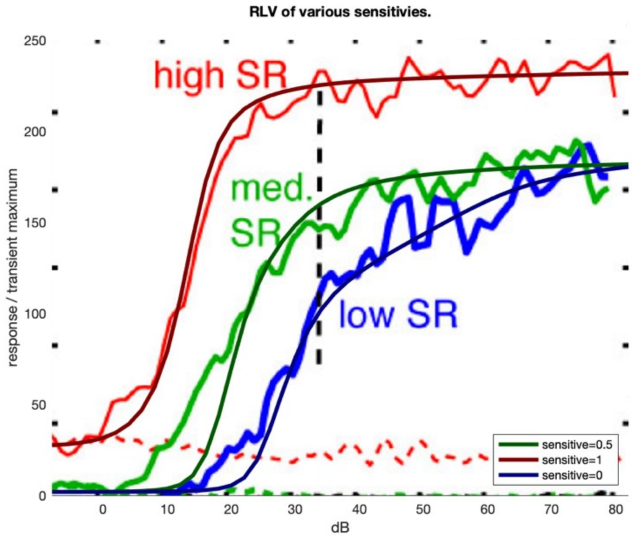


Figure 3.3: Rate-level functions in comparison to experimental data refenced in (Zilany et al., 2013). The BF and tones are 2kHz.

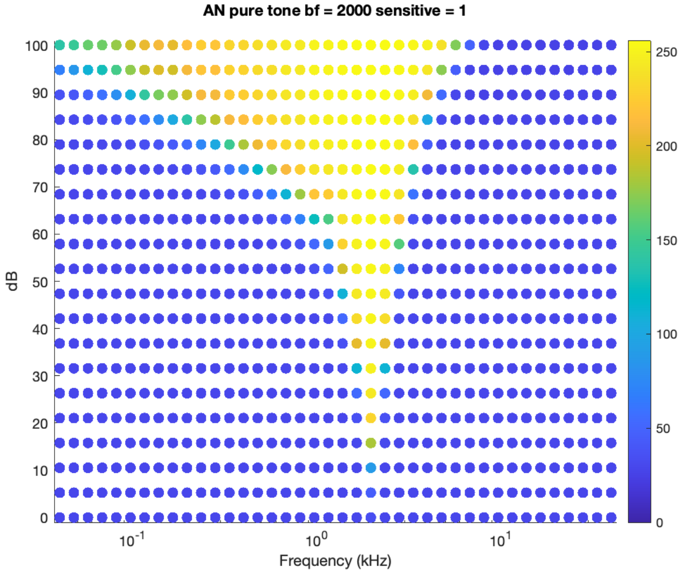


Figure 3.4: Steady-state response map of a model high spontaneous AN unit. This is to pure tones, in spikes/second. Its BF is 2kHz. The fiber saturates at high sound pressure levels.

The cochlear nucleus model:

The next part of the ascending pathway from the auditory nerve is the cochlear nucleus. We use corner frequencies of F_{cn} ranging from 500-700Hz to match the capabilities of the cochlear nucleus neurons as seen in (Rhode et al., 2010).

Primary-like units: These units behave like the AN fibers so have similar properties (Rhode et al., 2010). These simply are AN fibers but with an added first-order low-pass filter at corner frequency $F_c = 700Hz$. For these simulated units, their temporal response to a tone at BF is shown in figure 3.6.

Type 2 units: These units are sharpened by surround inhibition. 7 inhibitory AN units are used spaced linearly on a log scale ± 3 octaves from Bf weighted by a gaussian with a sigma of three octaves. The total inhibitory weight is scaled to sum to 4 and is passed through a first-order lowpass filter of $F_c = 700Hz$. This filtered inhibition is subtracted from the activity central excitatory AN unit and first-order-low-passed to $700Hz$. Finally it is passed through a sigmoid gain (50% at 250 spikes/s, 1000 spikes/second maximum rate, gain of 0.025 s/spike). The effect of inhibition is apparent in figure 3.5. This pattern of applying a filter and gain at the end is used for most types of units. Note that the actual sustained rate doesn't get anywhere near 1000 spikes/s. Their temporal response (not shown) is very similar to the primary-like units although it is a little slower.

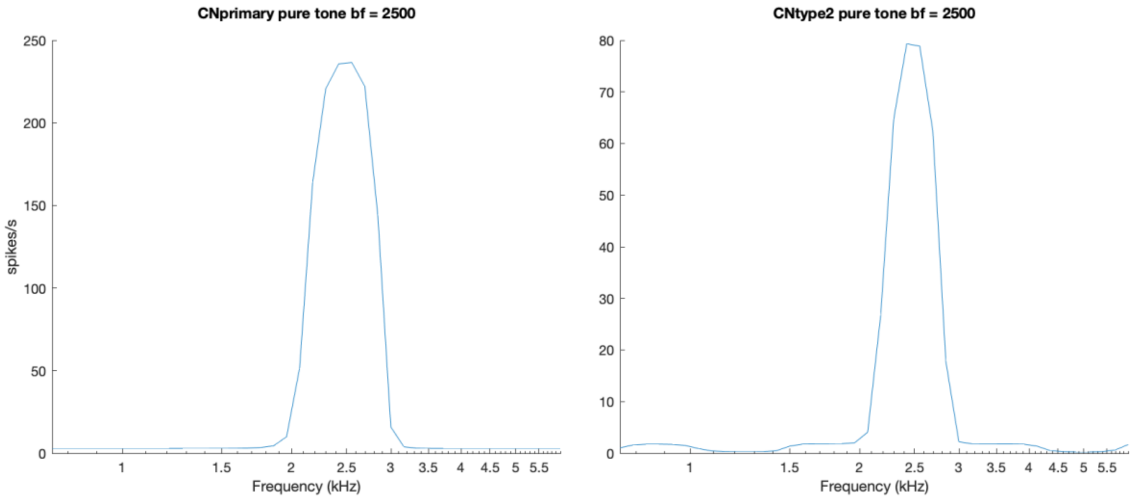


Figure 3.5: Primary-like and type 2 CN unit response maps. Surround inhibition sharpens the type 2 unit slightly (right).

Type 4 units: Type 4 units that are excited by a wide range of frequencies but inhibited by type 2 units (Rhode et al., 2010). Our model uses a gaussian bank as explained previously that is added together and filtered but is used as excitatory rather than inhibitory stimuli. A type two unit is subtracted out from this excitation (with a weight of 2). This is then first-order-lowpassed at 700Hz and a gain is applied (50% at 100 spikes/s, 1000 spikes/second maximum rate, gain of 0.025 s/spike). Their temporal response to a tone at BF can be inhibitory, if we define “BF” as the center of the receptive field. This is shown in figure 3.6.

Onset units: These have a strong but short-lived response to a wide range of frequencies (Rhode et al., 2010). We create a gaussian bank of AN fibers as in the Type 2 units. This bank is used for both the excitatory and inhibitory parts of the network: The inhibitory part is delayed 3ms and twise first-order-lowpassed to 500Hz. The excitation minus 2.3 times the inhibition is then first-order lowpassed (700 Hz) and sigmoid-gained (50% at 260 spikes/s, maximum 1000 spikes/s, gain of 0.0175 s/spike). Their temporal response to a tone at BF is shown in figure 3.6.

Primary notch units: These units have a gap, or “notch” in their response but otherwise are primary-like (Rhode et al., 2010). An onset unit is delayed 3 ms, first-order-lowpassed to500Hz and multiplied by 0.8. This is subtracted from a primary-like unit. The result is first-order-lowpassed (700 Hz) and sigmoid-gained is applied (50% at 450 spikes/s, maximum 1800 spikes/s, gain of 0.01 s/spike). Their temporal response to a tone at BF is shown in figure 3.6.

Pauser units: These units have a short transient response and then take time to build back up to a steady state (Rhode et al., 2010). The excitatory neuron is an AN unit. The inhibitory neuron is an onset unit which is 5dB more sensitive. It is weighted by a factor of three, delayed 3ms, and first-order-lowpassed to 25 Hz (which is much slower than other timeconstants). The excitation minus inhibition is first-order-low-passed (700 Hz) And sigmoid-gained (50% at 450 spikes/s, maximum 1700 spikes/s, gain of 0.01 s/spike). Their temporal response to a tone at BF is shown in figure 3.6.

Chopper units: These units respond to the total amount of sound, and have some oscillations when a sound is first given which decay over time (Rhode et al., 2010). *The model here is very different from previous models*. These units are modelled as a continuous population of noisy integrate-and-fire units. Let $y(v)$ be the number of neurons at dimensionless voltage v (zero is the equilibrium with no input, and one is when the neurons fire). When neurons fire we assume that v is reset to zero.

We compute the forward and reverse fluxes which represent neurons changing their voltage. The equation governing the evolution of the population distribution is given by:

$$\frac{\partial y}{\partial t} = \frac{\partial y}{\partial t}_{Leak} + \frac{\partial y}{\partial t}_{Noise} + \frac{\partial y}{\partial t}_{Input} + \frac{\partial y}{\partial t}_{Fire}$$

Leakage acts to concentrate the population near zero: $\tau \frac{\partial y}{\partial t}_{Leak} = v\frac{\partial y}{\partial t} + y$

Noise spreads the population out: $\frac{\partial y}{\partial t}_{Noise} = D\frac{\partial^2 y}{\partial v^2}$

The input shifts the population to higher voltages: $\frac{\partial y}{\partial t}_{Input} = -j_{ext}\frac{\partial y}{\partial t}$

We enforce a Dirichlet boundary condition of zero at $v = \{-1,1\}$. The boundary at -1 is artificial but it is so very little mass even reaches it. We compute the firing rate by looking at the fraction of neurons per unit time that cross the boundary at 1:

$$R = -D\frac{\partial y}{\partial v}_{v=1}$$

The firing rate is "injected" back at v zero (with the [] indicating if-statements):

$$\frac{\partial y}{\partial t}_{Fire} = \lim_{\epsilon\to 0}\frac{R}{\epsilon}[2v > -\epsilon][2v < \epsilon]$$

The initial condition approximately solves $\frac{\partial y}{\partial t} = 0$ for $j_{ext} = 0$, i.e. it is an equilibrium condition.

Parameters: $\tau = 10ms$, $D = 1.75s^{-1}$, $\frac{j_{ext}}{AN_{rate}} = 1.3sp^{-1}$,$\frac{output}{R} = 0.53\ sp$

There are several numerical considerations. 61 grid points are used ranging from -1 to 1. Numerical diffusivity is estimated and accounted for by reducing the amount of diffusion added (however, for stability we do not let it go below zero). The fraction of mass that would cross the boundary is calculated each timestep, used R, and shifted down by 1, avoiding the need to actually estimate the derivative at that point.

The total firing rate is 0.53 spikes divided by R (because R has units of s^{-1}). No filtering is used, but the noise in the model will act like a lowpass filter. Their temporal response to a tone at BF is shown in figure 3.6.

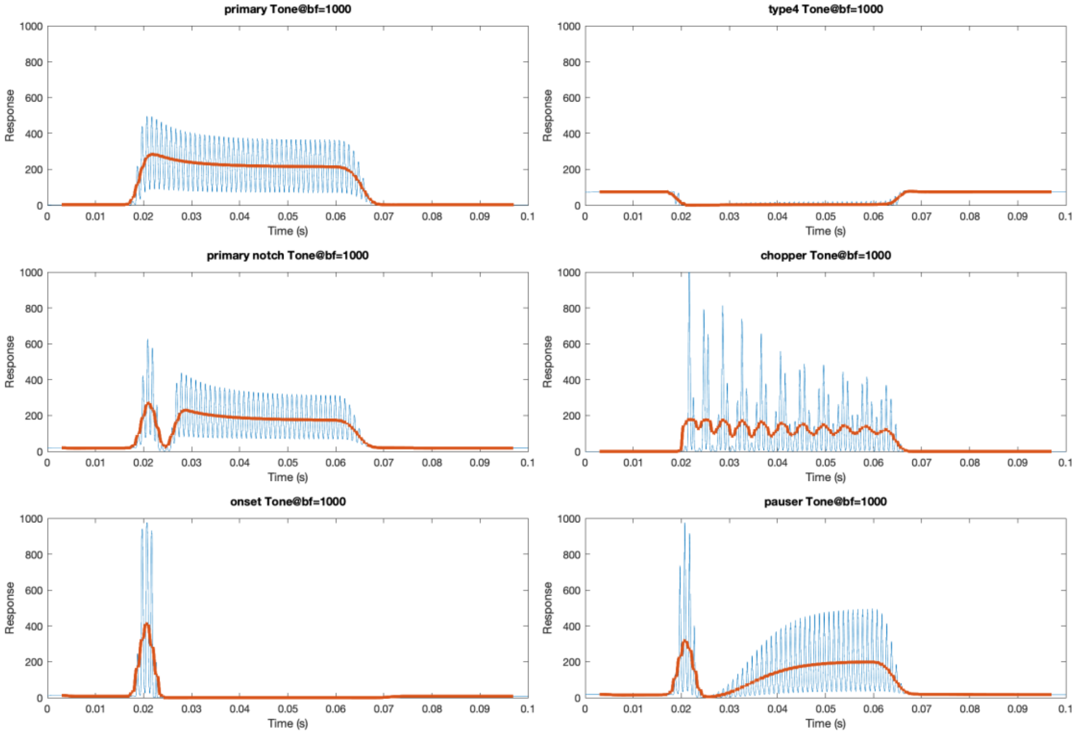


Figure 3.6: Instantaneous firing rates of model cochlear nucleus neurons. A**:** Primary-like. B: Type 4, C**:** Primary-notch. D**:** Chopper. E: Onset. F: Pauser. Note that the type 4 units can be inhibited by tones at their BF.

The inferior colliculus model:

The inferior colliculus combines a collection of primary-like cochlear nuclear units. The collection are weighted-summed together, first-order low-passed at 200Hz, and sigmoid-gained with a gain of 1/1000 s/spike (adjusting the weights instead) and variable maximum rate and thresholds. We only consider the spectral response maps to tones in building the model, matching the unit types as closely as possible to (Schreiner & Winer, 2005). The weights and locations are shown in table 3.1.

Unit type	Unit description	Bf_{CN}/Bf_{IC}	dB_{CN}-dB_{IC}	Weights	Sigmoid parameters*
V	Broad	{1.0,0.7,1.5,0.5,2.0}	{-10, 10, 10, 10, 20, 20}	All 1.67	300, 825, 1/1000
I	Narrow	{1.0,0.5, 2, 1, 0.25}	{-15, 30, 30, 30, 20}	{10, -6, -6, 40, -50}	, 250, 1/1000
O	Non-monotonic	{1.0, 0.5, 2, 1, 0.25}	{-15, 30, 30, 30, 20}	{30, -30, -30, -30, -30}	, 325, 1/1000

Table 3.1: Model IC unit parameters. Each IC unit has multiple CN primary-like units feeding in, the relative value of which are listed here. *Sigmoid parameters is listed {50% level in sp/s, maximum in sp/s, gain in s/sp}.

The type V units use a collection of excitatory units to broaden the receptive field. The type I units use center-surround patterns to keep themselves as sharp as possible (real-like type I unit can maintain sharpness at higher frequencies than our model. The type O units use high threshold inhibitory units to prevent response at high amplitudes, but this is only one of *many* type O patterns which would each require a separate row to be added to table 3.1. The responses are shown in figure 3.7.

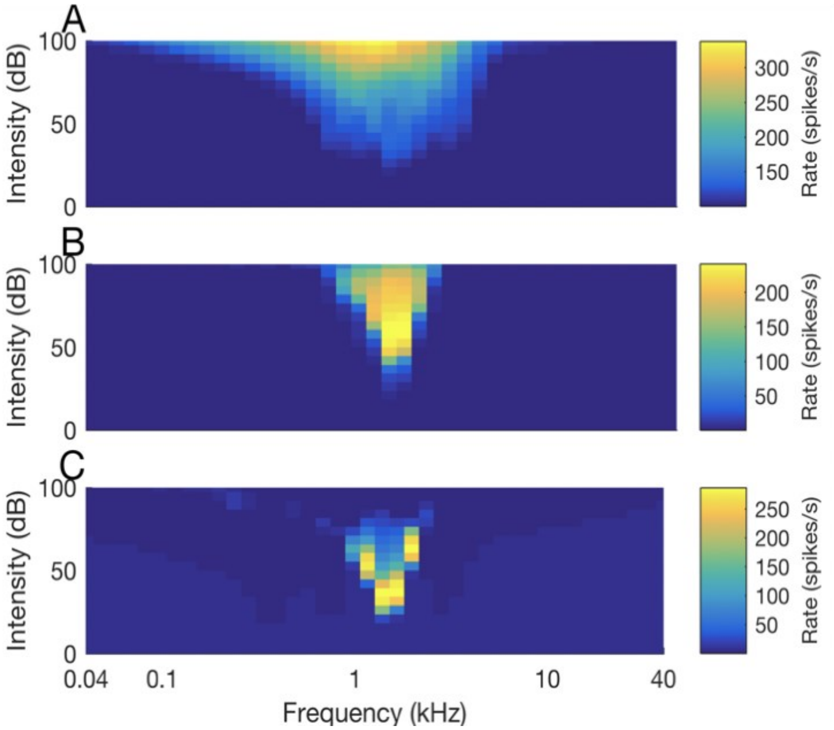


Figure 3.7: Model inferior colliculus neurons. These showing a loose resemblance to the response maps of type V (A) I (B) and O (C) neurons. The type I neuron fails to be as narrow as it should be despite our best efforts at using surrounding inhibition to sharpen the receptive field. What is shown is the steady-state response to pure-tones of the specified frequency and intensity.

However the model hit difficulty at the IC in terms of performance issues but more importantly a combinatorial explosion of the degrees of freedom. Thus we settled on modelling the IC as a gaussian convolution, where we worked with both *fixed-bandwidth* and *fixed-Q-factor* kernels. This only considers type I neurons, which is reasonable given we are looking for the maximally spectrally selective neurons. Note that the ascending IC neurons in marmosets were not found to integrate across disparate frequency bands, the (narrow) range of frequencies essentially make the sharpest neurons respond as a center-surround EI pattern rather than being strongly multipeaked neurons (Kostlan, 2015).

Temporal cortical modelling

We briefly explored temporal models of cortical neurons in which sounds are click trains with no definition of stimulus frequency. Neurons have been found with a variety of click-train-response properties. Some are *synchronized* to the waveform while others aren't, and there are *lowpass*, *highpass*, and *bandpass* responses; this is shown in figure 3.8 (Gao et al., 2016). Computational models of these neurons would need to look at synaptic time constants.

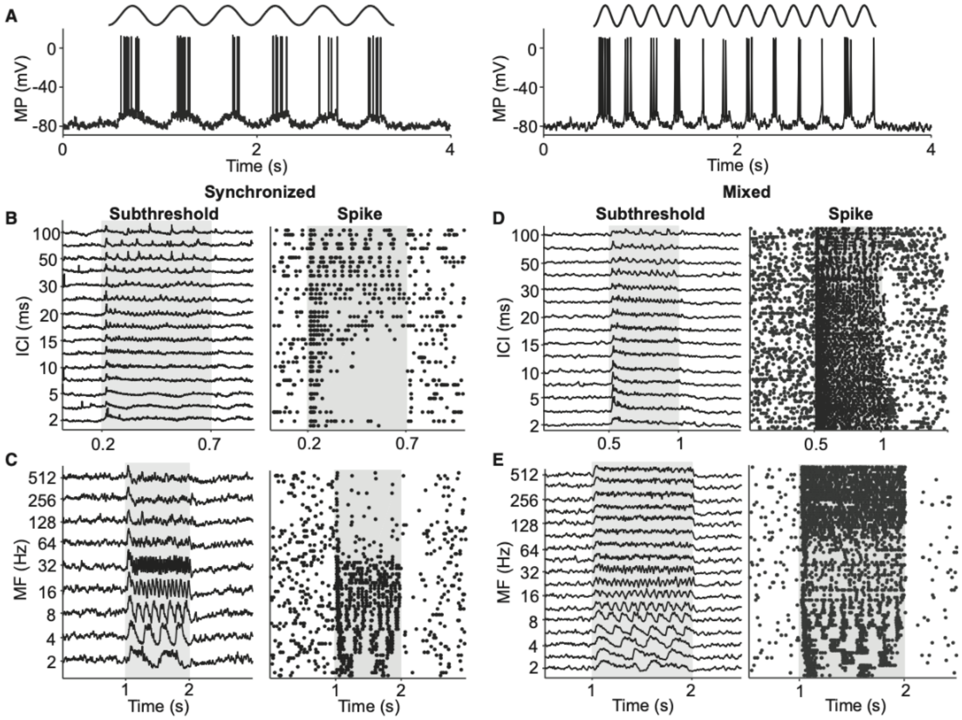


Figure 3.8: Modulation-sensitive neurons in the marmoset A1. A: The envelopes and subthreshold response, with the modulation frequency 2 Hz (left) and 4 Hz (right). A: The envelopes. B, C: A synchronized units that phase-locks to a click train, with the inter-click interval raging from 2 to 100 ms. The sub-threshold is the average of five trials. *Figure from (Gao et al., 2016).*

We extend the model from (Bendor, 2015). Briefly summarizing their model: it was an integrate-and-fire model in which there is a bank of jittered excitatory synapses and inhibitory synapses. Each synapse, when "triggered" by a click, generates a rise-fall conductance $\sim \frac{t}{\tau} \exp\left(-\frac{t}{\tau}\right)$. The timing of the triggering is jittered randomly with a gaussian noise. There are leakage currents and noise currents as well. When the neuron exceeds it's threshold it fires and it's voltage resets.

The model was extended by including a "vitality" parameter. Synapses deplete when they are used and take time to recover. To compute the vitality at the next timestep, we have: $v_{j,t+1} = \min\left(1, v_{j,t}\left(1-\frac{\Delta_t}{\tau_{sat}}\right)+\frac{\Delta_t}{\tau_{sat}}\right)[S_{t+1}=0]$, where τ is the time-constant for recovery, Δ_t is the timestep, and $[S_{t+1}=0]$ tests whether no stimulus has been played at that timestep. This the vitality gets "reset" to zero after the synapse fires and recovers exponentially, as summarized in figure 3.9.

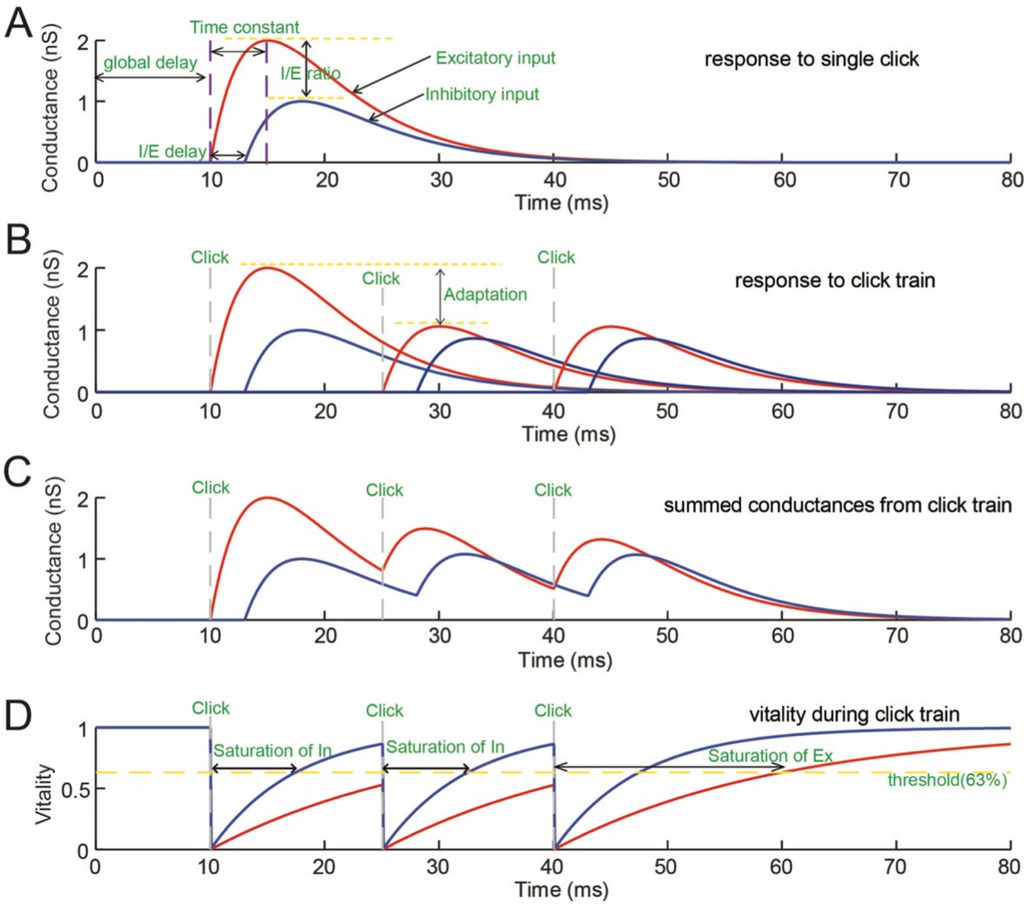


Figure 3.9: Demonstration of the "synaptic vitality" at work. A: The EPSC and IPSC's created from a synapse firing. There is random jitter to the location of the curves (not shown). B, C: Depleted synapses don't fire as strongly, so the summed conductances isn't as much as a linear addition would be. D: The underlying vitality over time that depletes and recovers. *Figure from* (Gao et al., 2016).

Our extended model sometimes had extremely high synaptic conductances that would create numerical instability. To prevent this, we calculated the equilibrium voltage:

$$V_{eq} = \frac{V_r g_{leak} + \sum_j g_{jt} E_j + I_{noise}}{g_{leak} + \sum_j g_{jt}}$$

If the ratio between the total conductance and the capacitance fell below the timestep, the voltage was set to V_{eq} instead of integrating the time-dynamics. This represents the membrane time constant falling below the timestep.

We used the range of parameters shown in table 3.2.

Parameter	Meaning	Value(s)
Δt	Timestep	0.0001 s
t_d	Global delay (visual appearance only)	0.01 s
τ_{sat}	Synaptic depletion time constant	0-0.025 s for excitatory, 0.0075 s for inhibitory synapse
E_j	Synapse zero-current voltage	0 mV for excitatory, -85 mV for inhibitory
W_e	Excitatory synaptic weight	Set per-neuron to make the maximum firing rate about 50-55 spikes/s over a 10Hz-500Hz range of click train frequencies.
I_{noise}	White noise current	800 pA. Note: I_{noise} scales ~ $\Delta t^{1/2}$ for an equivalent model at a different Δt.
Σ_{syn}	Synaptic jitter standard deviation	0.001 s
g_{leak}	Leakage conductance	25 nS
C	Membrane capacitance	0.25 nF
V_r	Resting voltage	-60 mV
V_t	Threshold voltage for spiking	-53 mV
I/E	Inhibitory/Excitatory conductance ratio	0-2
τ_{syn}	Synaptic time constant of rise-fall	0.005-0.025 s
d_I	How much more delayed is inhibition than excitation, on average	-0.002s to 0.007s (negative values indicate inhibition arrives earlier on average).

Table 3.2: Parameter values and ranges used for the integrate-and-fire model. *Parameters are named as used in (Bendor, 2015).*

We also developed a way to test for statistical significance of Fourier components. Consider a discrete noisy signal $x = k \sin(2\pi f - \Theta) + \varepsilon$, sample rate r. The noise ε is a time-series with some autocorrelation but with no sinusoidal structure. The null-hypothesis is $k = 0$. Let X be the square magnitude of the Fourier transform of x^* where x^* is resampled and/or padded *x* such that the index on X that corresponds to f, if= 1+N f/r, is as close to an integer as possible. We want to see if the signal at if significantly rises above random chance.

We assume that the noise contribution to X is IID (independent and identically distributed) in the vicinity of if. Here, "vicinity" means any index between max(0.8if, if – 150) and min(0.8if, if + 150) (inclusive), but not including if *-1*, if, or if *+1*. Let *m* be the sample median of the noise in the vicinity of X (the median is more robust to outliers). Let *n* be the number of degrees of freedom (number of bins "in the vicinity").

The heuristic is a balance of two competing factors. If it is too wide, it may misestimate the background noise level because the background is not white-noise and thus varies with frequency. This could either inflate or deflate the p-value. If it is too narrow, there are less degrees of freedom which makes rejecting the null hypothesis harder.

For large *n* the test statistic $1.3863X_{i_f,null}/m{\sim}\chi_2^2$ (a χ^2 distribution with 2 degrees of freedom). With a corresponding p-value of $p = 1 - X_2^2$ (1.3863 X_{if}/m), where X_2^2 is cumulative distribution of χ_2^2. This is because 1.3863 is the median of X_2^2. For small n-values, we have to use the equivalent of the T-test to account for uncertainty in the true median. We compute p-values numerically by Monte Carlo sampling the cumulative test-statistic distribution, $T_n(z)$ under the null hypothesis. Then we run our test on the data to get z and calculate p (z) = 1 – T^{-1}(z), where T^{-1}(z) is the inverse function of T (z). This ensures that null-hypothesis p-values will be uniformly distributed on [0-1) as it should be in any significance test. This method is very general so will work as long as care is taken for numerical stability at the tails. For small n the p-values will be higher on average, which makes rejection more difficult (Bloomfield, 2004).

We also adapted a power-law fitting model. For each model neuron, we calculated a time-dependent spike rate and mean membrane potentials (excluding the spikes) by placing a 10 ms window every 1 ms (with overlap). We fit the data with a least-squared power-law model from :

$$Fr = k(V - V_{threshold})^{p}[V > V_{threshold}]$$

where Fr is firing rate, V is the membrane potential, p is the exponent and k is a gain factor. $V_{threshold}$is not the biophysical threshold, but is instead a free parameter determined by the regression and represents the point above which it becomes feasible for EPSP's to trigger the neuron (Priebe et al., 2004).

For parameter stability, we found it necessary to add a weighting factor, weighting bins with more spikes more heavily in the regression but with the weights being constant for bins below 1% of the maximum number of spikes in any one bin.

Our extension of the model was able to reproduce "negative monotonic neurons" that respond preferentially to low (10Hz) click train frequencies. It also was able to reproduce both the presence and absence of synchronization, see figure 3.10.

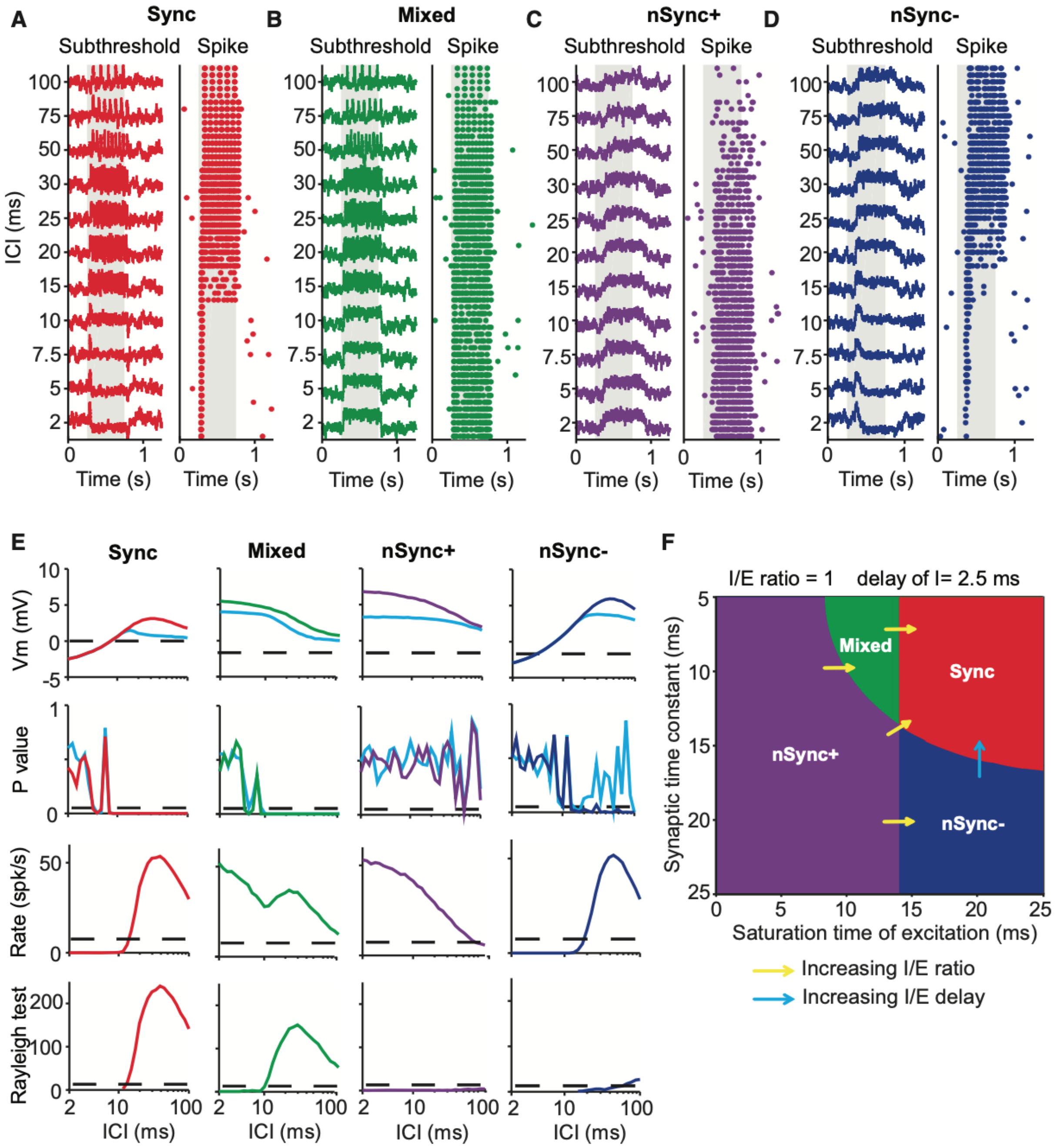


Figure 3.10: The integrate-and-fire model's results. A: Subthreshold and dot-plot of a "synchronized" model neuron that phase-locks to the click train down to an inter-click interval of about 20 ms, below which it ceases to have a sustained response. B: a "mixed" neuron that phase-locks at low frequencies but fails to do so at higher frequencies. C: A non-synchronized "+" unit that doesn't phase-lock and any of the frequencies supplied but prefers smaller ICI's. D: A non-synchronized "-" unit that doesn't phase-lock and any of the frequencies supplied but prefers ICI's around 100 ms. E: the average membrane potential, phase-locking p-value, spike rate, and Rayleigh test statistic of the four neurons. In the first and second row, the light blue curve is the membrane potential when spikes are allowed and the other colors is the potential when spikes are disabled (allowing it to rise higher). F: One plane in parameter space that contains all four kinds of units, with the effects of changing two other parameters shown by arrows (Gao et al., 2016). *Figure from (Gao et al., 2016)*.

However, our model failed to produce the pitch sensitivity to temporal regularity as found in (Bendor & Wang, 2010). Lacking a model of pitch and periodicity, we decided to focus on spectral models for harmonic selectivity.

Training stimuli

Stimuli were modified harmonic complexes with various ranges of fundamental frequency, jitter, intensity range, and mistuning (exact values need to be summarized in a table). In all cases the background stimuli were "degraded" versions of the harmonic stimuli, i.e. higher jitter or mistuning ranges (combinatorics need to be summarized in a table). Random seeds were used for repeatability.

Spectral subcortical models

Stimuli are represented as lists of spectral location and energy intensity (see below for descriptions of what stimuli are used). For training, 256 stimuli were used.

Subcortical input was computed at 80 channels linearly spaced from zero to 4.5 times the training fundamental frequency. It was generated by convolving the stimuli with a gaussian kernel that integrates to one with σ that was 3 bins (channels) for most test cases but 1.2 bins for the "narrow" case and 7.5 bins for the "broad" case:

$Subcortical = \frac{1}{Z}\sum p_i exp\left(-\frac{Bf - f_i}{2\sigma^2}\right)$, where Bf is the vector of best-frequencies, f is the stimulus frequency, p is the stimulus intensity, and Z normalizes the kernel. The effect of applying this convolution to is shown in figure 3.11.

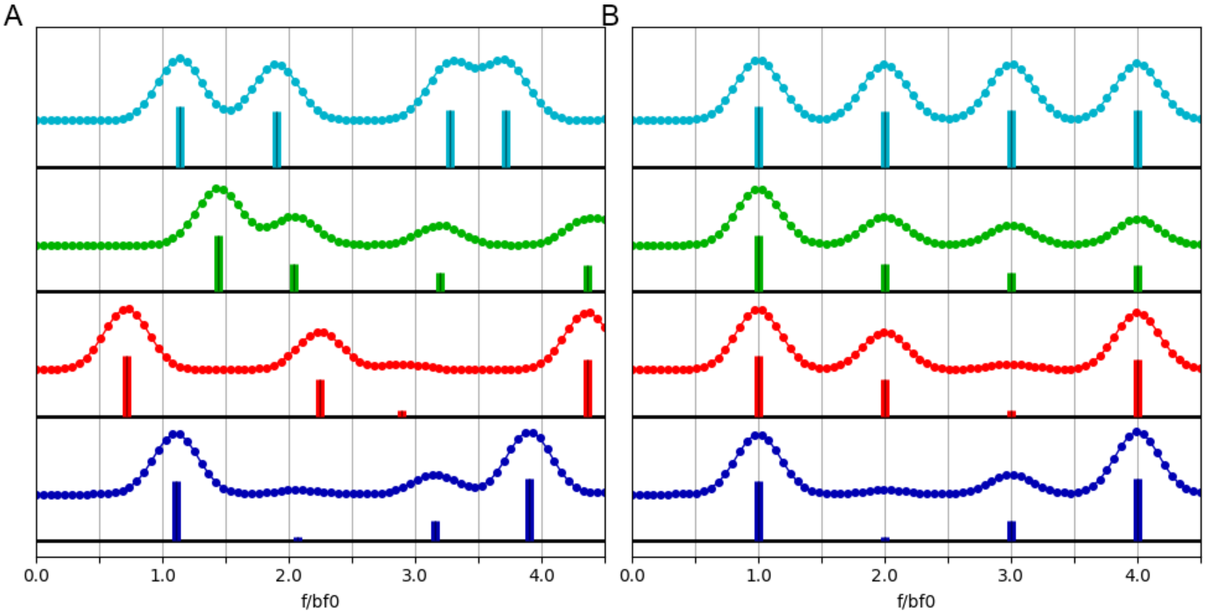


Figure 3.11: Example training stimuli and subcortical input. The stimuli are represented as vertical bars, while the input is the curve. A: Background stimuli (stimuli for which the network must not respond to). B: Target stimuli, for which the network is trained to respond to. The zero frequency would most realistically correspond to an extremely low frequency neuron that is close to zero on this linear scale.

Noise model

The network output was calculated without any sort of randomization. However, we estimated how noise would propagate. There are three potential sources of noise, with different simulation runs using different sources. Input noise is the amount of noise in the stimuli vector. Neural noise is noise added after the gain function for each batch of neurons. Finally, output noise is added at the very end.

Noise is propagated through the network. The variance is tracked at each step, and assumed to be infinitesimally small: for $x \rightarrow g(x)$ the variance would propagate as

$$v_x \rightarrow v_x \left(\frac{dg}{dx}(x)\right)^2$$

When several variances are added (when a neuron gets a weighted input from other neurons) we assume they add linearly (statistically uncorrelated). We also assume lack of correlation between a neuron's added output variance and the noise that it adds.

The noise variance can be either constant or proportional to the response. The latter case models a Poisson distribution. However, the output is modelled as a normal distribution; in the interest of simplicity we don't capture higher order moments. The default noise used was "output" Poisson noise with variance 0.1 when the maximum firing activity of 1.0 is reached, but the weights were insensitive to different noise levels and models.

Neural network topology

Neural networks were set up with various feed-forward topologies, both obeying Dale's principle or not. Feed-forward networks had between 2-4 layers (0-2 hidden layers), feeding to a single output unit. The input layer had 80 channels (160 neurons if Dale's principle was invoked). Higher layers had progressively fewer neurons. A sigmoid gain was used. The simplest network we used is shown in figure 3.12, and the simplest Dale's-principle-respecting network is shown in figure 3.13 and shown in a different way in figure 3.14.

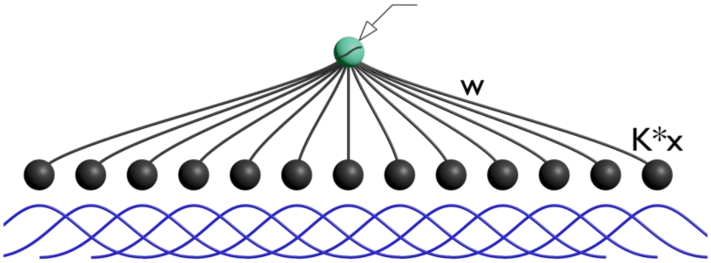


Figure 3.12: Diagram of the simplest feed-forward training network. The input layer (black) convolves the stimulus with a gaussian kernel which represents it's receptive field (each blue curve). For most experiments the kernel has a fixed width, but we occasionally tried fixed Q-factor instead (fixed ratio of width to center frequency). The neurons have zero threshold and no saturation (i.e. identity gain function). The output neuron (cyan) has a sigmoidal gain function and non-zero threshold.

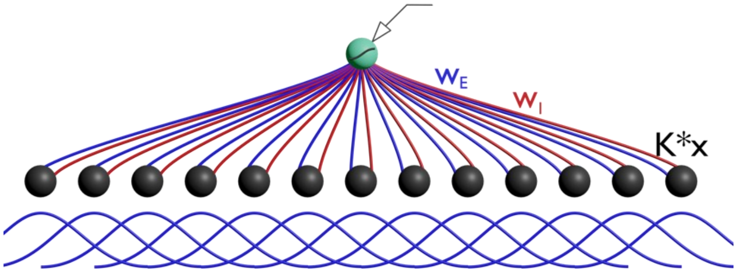


Figure 3.13: Diagram of the sign-constrained network that respects Dale's principle. Red curves are excitatory, blue curves are inhibitory. The weights are independently adjustable, giving us twice the number of weights to train.

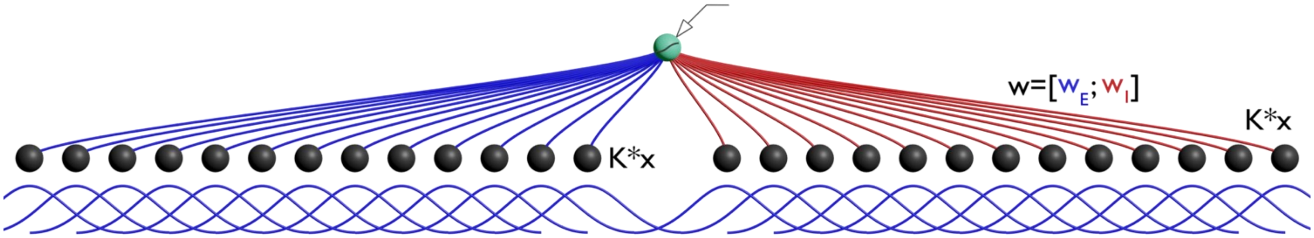


Figure 3.14: The sign-constrained network represented differently. It is sometimes easier to visualize the system if we concatenate the excitatory weights and inhibitory weights into one vector. This is mathematically equivalent to the [previous] figure.

Neural network model

We developed a method to associate both a discreet time model and a continuous time model with the underlying dynamics of the system. In the discreet model, each neuron's state is given by $s_{i,t+1} = \sum_{layer} \sum W_{ji} y_{j,t}$, where the outer sum is over all layers that connect to the target neuron's layer, W is the weight matrix, b is the threshold, and g is the gain function. The neuron's firing rate is given by: $y = g(s - b)$ for a sigmoidal gain g. There is a special *input layer* which is set to the stimuli.

For the continuous model, we have a similar equation with an additional leakage term, and it is solved by the adaptive Runga-Kutta 4-5 method, which was found to be robust.

$$\frac{ds_i}{dt}\tau^{-1} = \sum_{layer} \sum W_{ji} y_j \ - s_i$$

In a steady-state situation the continuous model's activity is equal to the discrete model's activity. The discrete model is used for supervised learning in this Chapter because it is much easier to calculate gradients. The continuous model is used for the Hebbian recurrent cases in Chapter 4 because it better tracks temporal dynamics.

Supervised training and testing

Networks were trained by applying 128 iterations of the conjugate gradient method to an objective function with penalties for weights that have the wrong sign or get too big in magnitude. The penalty function, which is piecewise defined but differentiable, is a combination of the weight magnitude penalty function and the sign penalty function:

$$2X \sim \max\left(0, \left(\left(\sum w_{ij}^{p} \right)^{\frac{1}{p}} - W_{max} \right)^{2} \right) + \sum \max\left(0, \left(-s_{ij} w_{ij}\right)^{2}\right)$$

Where s_{ij} is the sign restriction function, 1 forces a positive sign, -1 forces a negative sign, and 0 allow either sign. The p-norm is usually 1 or 2. The objective function uses one of two criteria. The first is the "D'" of the output: $D' = \frac{(y-0.5)(2L-0.5)}{\sigma}, L \in \{0,1\}$, where L is the label. The second is based on expected value of the accuracy:

$$a = \phi\left(\frac{y - 0.5}{\sigma}(2L - 0.5)\right), L \in \{0,1\}$$

The strength of the penalty is set so that the relative violation is around 1-5% (assuming that the optimization was hitting the penalty is applying in the first place). After gradient descent, the weights are projected onto the nearest set of allowed weights to remove this minor violation. The effectiveness of a trained network on a harmonic stimulus is shown in figure 3.15.

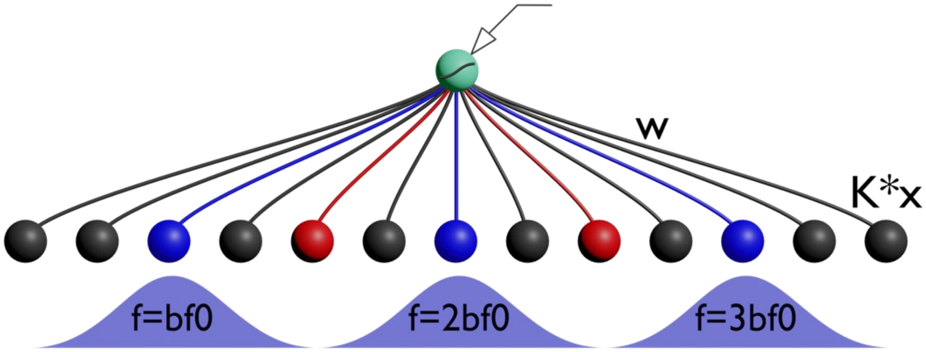


Figure 3.15: How a sieve network detects harmonics. The blue curves represent the subcortical input of a harmonic sound. They activate the excitatory weights (blue neurons) but avoid the inhibitory weights (red neurons). This maximizes how strongly the output neuron is driven.

Effective linear weights

If all gain functions are replaced with the identity function, a feed-forward network produces an output scalar that is a linear function of the input. We call this weight vector the *effective linear weights*. These are defined for feed-forward networks but not recurrent networks.

Linearized weights around an input

Neural networks given a stimulus can be evaluated by linearized weights. This is accomplished numerically with a difference quotient, by computing $(\frac{dy}{dx})_i \approx \frac{\Delta y}{\Delta x_i} = \frac{1}{2\epsilon}(y(x + \epsilon\delta_i) - y(x - \epsilon\delta_i))$ where x is the input stimuli vector and y is the scalar output. Care was taken to ensure that the relative numerical precision and discretization errors are less than 1%.

RSS weights

We calculate RSS weights as in (Slee & Young, 2013). For a matrix of stimuli columns X we fit the response as a linear regression pattern: $y = ax + b\ + \epsilon, \epsilon \sim N(0, \sigma)$. We generate 512 background stimuli as our X. In practice the RSS weights were very similar to the linearized weights so we show the latter; RSS weights are essentially linearized weights but with a large finite difference quotient.

Facilitation and periodicity indexes

We calculate these indexes as in (Feng & Wang, 2017), except that we only consider a single harmonic fundamental frequency (the one we use for training).

If necessary, we add a small ridge on the matrix that gets inverted such that $\frac{\lambda_{\max M^T M}}{\lambda\_\min M^T M} \leq 4n$, where n is the number of stimuli. We don't compute any quadratic regression terms because the number of degrees of freedom is very high (we do explore quadratic terms in Chapter 4, however).

Sieveness and linclassness

Two vectors can be measured as to how similar their directions are:

$C(x, y) = \frac{x \cdot y}{|x||y|}$

This index ranges from -1 (vectors pointing in opposite directions) to 1 (vectors pointing in the same direction).

We calculate the *sieveness* by comparing the effective weights to a sieve:

$sieveness = C(w_{eff},\ \cos{(2\pi Bf/f0)})$

We calculate the *linclassness* by comparing the effective weights to a linear classifier:

$$linearness = C(w_{eff}, \langle \mathrm{x}_{\mathrm{L}=1}\rangle - \langle \mathrm{x}_{\mathrm{L}=0}\rangle)$$

Where $\mathrm{x}_{\mathrm{L}=0}$ and $\mathrm{x}_{\mathrm{L}=1}$ are the average subcortical input for the background and foreground stimuli.

3.3 Results

Sieves are robust

Harmonic sieves emerge naturally and appear to be a universal feature of harmonicity detection at a particular bf0, not just an effect of a linear classifier. Changing the source of noise (or switching between constant-variance and Poisson noise) had little effect on the weights. Also, the evaluation criterion had little effect on the weights. The basic sieve result is shown in figure 3.16.

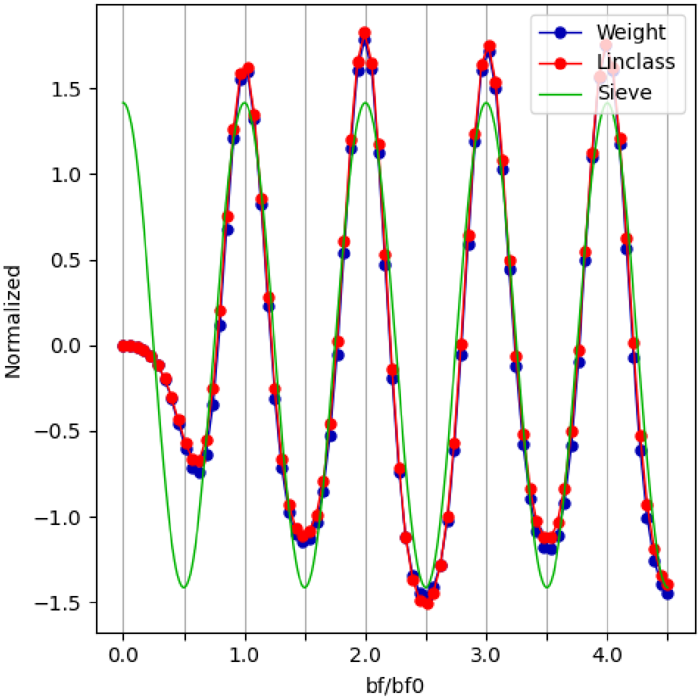


Figure 3.16: The weight vector for the simplest harmonic training case. The weights are very similar to a sieve (green curve). They are even closer to a linear classifier (red curve). The zero-frequency component actually would correspond to a very low frequency unit.

Some changes did have an effect on the sieve, but didn't degrade it's vector-strength significantly. The closer to linearity that the neural network is operating under the closer the weights come to an optimal linear classifier, so when strong weights were allowed (which brings the neurons toward their rails, enhancing nonlinearity), the sieve was further from a linear classifier. When the network was trained with a hidden layer, the input to the hidden layer acted as a sieve in the dimension corresponding to the frequency while being more or less random in the dimension corresponding to the input to the output neuron; training was slower with two levels and unsuccessful from random initial conditions with three levels, but if it had been successful it is likely the results would still be a sieve.

When the input is broken down into an excitatory layer (only positive weights allowed) and an inhibitory layer (only negative weights allowed), the positive parts of sieve are assigned to the excitatory layer and visa-versa.

The stimuli had some effect on the sieve, but again preserved the sieve unless the problem became too hard for the network to do much of anything. When the convolution width became less than about 0.2 f0 the network and the linear classifier begin to get noticeably sharper than a sine-wave. In the limit of no convolution the weights and classifier became a comb of Dirac-deltas concentrated on the integer values (with no Dirac deltas on the half-integer values). When the convolution width gets above about 0.3 f0 the weights (again both for the network and the linear classifier) show strong edge effects and weakened oscillations; the oscillations become negligible above around 0.5 f0.

The background stimuli represent a degraded version of the target stimuli, where "degraded" means further from harmonicity in terms of having components that are not harmonic, i.e. mistuning. When the components in the background stimuli were mistuned globally, with the same mistuning across all trials, the sieve did become. However, neither per-component or per-trial mistuning changed the linear response or sieve significantly nor did the amount of per-component energy "height" randomization in the stimuli affect the sieve noticeably. Thus neural-network sieves are very robust and quite similar to linear classifiers over a wide range of parameters.

Sieves respect sparsity constraints

When the weight constraint becomes sparser (1-norm instead of 2-norm) the sieve has sharper peaks. A similar phenomenon occurs for Poisson noise, however it only happens in cases when weights to intermediate neurons were forced to be all positive or all negative.

One may consider a modified sieve of the form:

$$w \sim sgn_{pow}\left(\sin\left(\frac{2\pi f}{bf_0}\right), q\right), sgn_{pow}(x, p) = sgn(x)|x|^p$$

In this form $q = 1$ represents a sieve and $q > 1$ represents a spiky sieve, which was the case in the one-norm. This modified sieve with $q > 1$ is qualitatively similar to the results shown in 3.17.

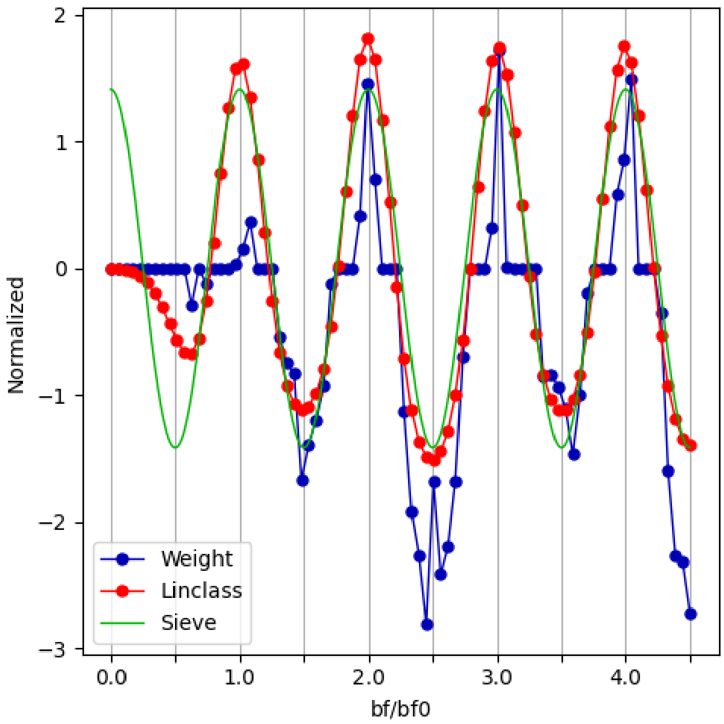


Figure 3.17: The weight vector for 1-norm-constrained training case. The weights form a sieve pattern, but they are sparser, with more values near zero and a few large values.

Dale’s principle splits the sieve

When Dale’s principle is used the network behaves almost identically to the sign-agnostic version. There are two banks of neurons, one is excitatory and one is inhibitory. If the sign-agnostic case would produce a positive (negative) weight at a given Bf the excitatory weight is positive (zero) and the inhibitory weight is zero (negative) for the corresponding Bf values. Training takes more iterations to converge as the problem is more nonlinear but it converges to almost the same behavior as a sieve anyways. This division of labor is shown in figure 3.18.

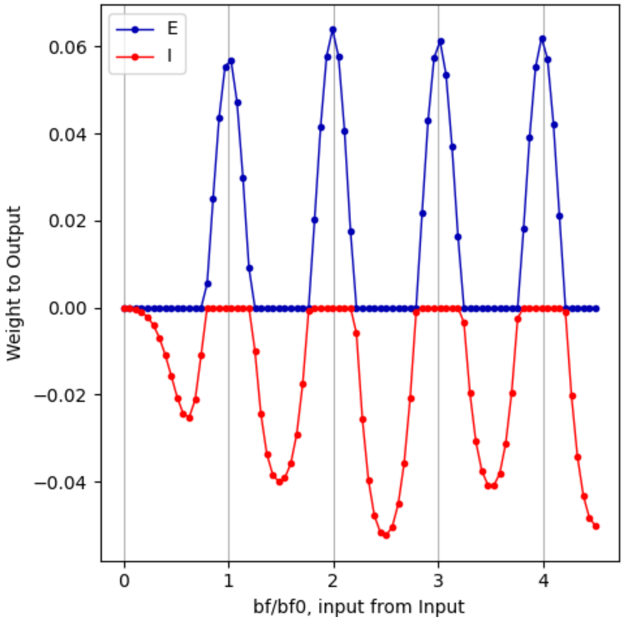


Figure 3.18: The weight vector for the Dale’s principle training case. The sieve is split between the excitatory and inhibitory parts.

<u>RSS weights are also sieve-like</u>

RSS weights resemble sieves as well, in line with the sieves. They are much nosier than the actual weights because they are essentially deconvolving the blurred subcortical input back into the stimuli. They still are a sieve, as shown in figure 3.19.

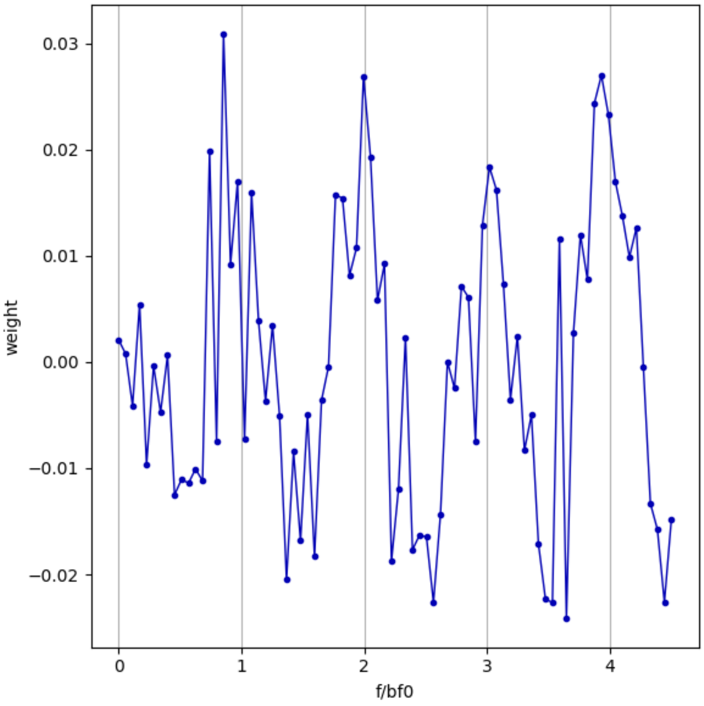


Figure 3.19: RSS linear weights showing a sieve-like structure. 512 RSS stimuli were used.

If it can be solved, it is a sieve

The sieve was the only solution that was found. We tried networks with hidden layers, as well as many other combinations. Figure 3.20 shows a network with one hidden layer, and figure 3.21 shows a network with two hidden layers. There never was a case where the network failed to have a high sieveness score while having a reasonably high accuracy. The combinations used are summarized in table 3.3.

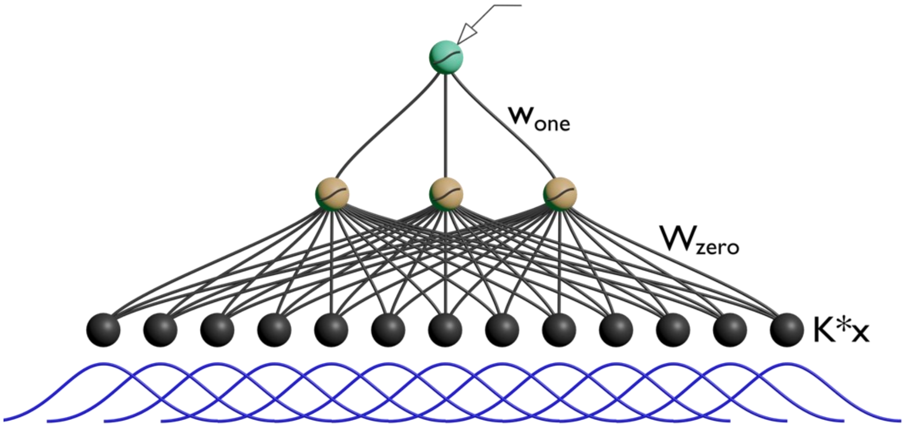


Figure 3.20: A one-hidden-layer network. There is an intermediate layer of neurons with a sigmoid gain (tan color).

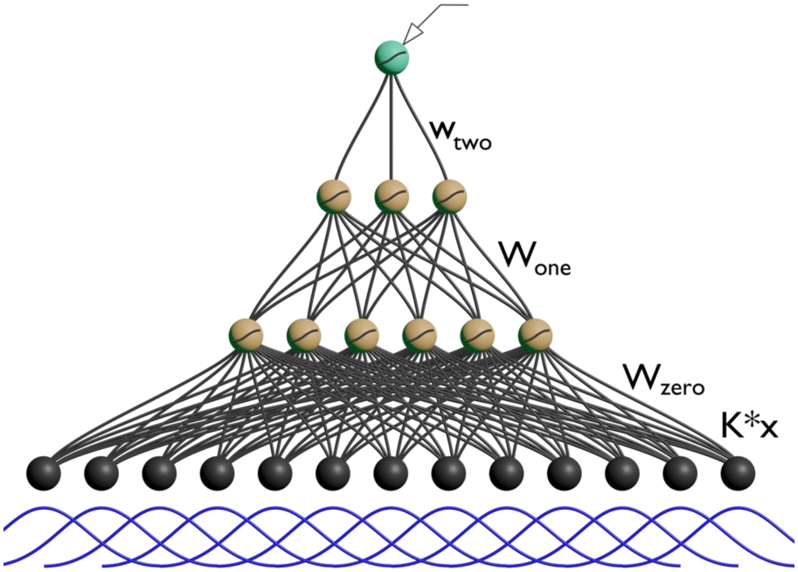
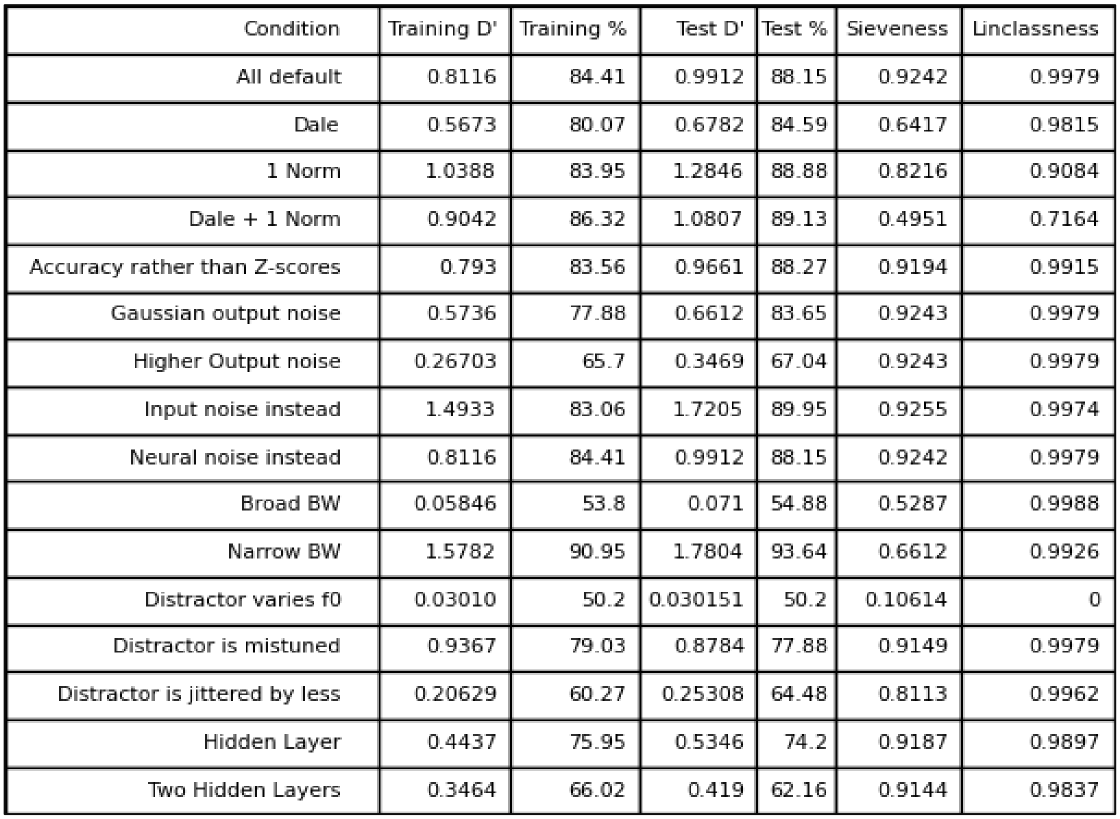


Figure 3.21: A two-hidden-layer network. There are two intermediate layers of networks (tan color).



Condition	Training D'	Training %	Test D'	Test %	Sieveness	Linclassness
All default	0.8116	84.41	0.9912	88.15	0.9242	0.9979
Dale	0.5673	80.07	0.6782	84.59	0.6417	0.9815
1 Norm	1.0388	83.95	1.2846	88.88	0.8216	0.9084
Dale + 1 Norm	0.9042	86.32	1.0807	89.13	0.4951	0.7164
Accuracy rather than Z-scores	0.793	83.56	0.9661	88.27	0.9194	0.9915
Gaussian output noise	0.5736	77.88	0.6612	83.65	0.9243	0.9979
Higher Output noise	0.26703	65.7	0.3469	67.04	0.9243	0.9979
Input noise instead	1.4933	83.06	1.7205	89.95	0.9255	0.9974
Neural noise instead	0.8116	84.41	0.9912	88.15	0.9242	0.9979
Broad BW	0.05846	53.8	0.071	54.88	0.5287	0.9988
Narrow BW	1.5782	90.95	1.7804	93.64	0.6612	0.9926
Distractor varies f0	0.03010	50.2	0.030151	50.2	0.10614	0
Distractor is mistuned	0.9367	79.03	0.8784	77.88	0.9149	0.9979
Distractor is jittered by less	0.20629	60.27	0.25308	64.48	0.8113	0.9962
Hidden Layer	0.4437	75.95	0.5346	74.2	0.9187	0.9897
Two Hidden Layers	0.3464	66.02	0.419	62.16	0.9144	0.9837

Table 3.3: Summary of training conditions, accuracy, and performance.

3.4 Discussion

Optimal linear classifiers are sieves

The optimal linear classifier weight vector with a 2-norm constraint for the z-score metric is proportional to the difference of the average of the harmonic and the background stimuli. This is a sieve in most cases, which explains the robustness of the sieve. Sparsity inducing norms work as expected: producing sparser weights (kurtosis of the weights-as-probability-distribution). Poisson noise models tends to act similar as it disfavors strong responses *in some training cases*.

Optimal classifiers and log likelihood

We can consider the background and foreground stimuli as coming from two distributions, and make an optimal classifier that selects which distribution has more probability density at a given point. If we have two normal distributions the log likelihood ratio is quadratic. With equal covariances the loglikelihood ratio is linear and a linear classifier is optimal.

Our two distributions (the background and foreground subcortical input) aren't precisely equal-variance or even gaussian. However, the problem seems to be close enough to linearity that a linear classifier works well for all the problems tested.

Locality constraints as biological budgeting

Axoplasm takes space and energy, so the brain tries to minimize distances between connected neurons by wiring neurons locally when possible. If a sieve can be built with only local connections it is advantageous. A multi-level network with locality constraints proved difficult to train, but if there is a way to train it still may provide an alternative to a single-summary neuron sieve.

Breaking down a sieve into excitation and inhibition

Splitting up the excitation and inhibition separately yields very similar results to a linear sieve. Dale's principle demands that, for example, glutamatergic synapses can't become GABAergic synapses as easily as a weight can switch from positive to negative *in silico*. However, a similar effect can be achieved by having an independent set of excitatory and inhibitory neurons feeding into the same neuron.

Future experiments

For a harmonic template unit there may be “excitatory” and “inhibitory” units that feed into it and could be found electrophysiologically. The “inhibitory” units would respond preferentially respond to *half-integer* multiples of some Bf_0 (or *odd-integer* multiples of ½ bf_0). Due to the tonotopy, it may be possible to trace a template neuron retrograde to see where it gets input from and use a second electrode to probe the best-frequencies of the region, similar to the trans-synaptic modified GFP two-photon calcium tracers used in (Maskos et al., 2002).

Harmonic template units only respond to a *single* best frequency. However, it may be the case that broadly tuned harmonic template units exist and can be found by experiments. We discuss this highly nonlinear problem more in Chapter 5.

CHAPTER 4:

Training harmonicity selectivity in recurrent neural networks

In Chapters 2 and 3 we showed that two different training paradigms can produce harmonic selectivity in feedforward networks. However, recurrent networks are paramount in the cortex. They have also seen numerous successes in machine learning; convolutional recurrent networks have seen success in including for removal of noise in speech (Hu et al., 2020; Strake et al., 2020). Here we address recurrent networks, which are pervasive in the cortex.

4.1 Introduction

The abundance of auditory cortical recurrent networks

The model in (Feng & Wang, 2017) has no recurrent connections, but the cortex is *highly* recurrent. In addition to the extensive physiological evidence, recurrent networks have several hallmarks. Hysteresis may be one of the most obvious examples, as Hopfield networks admit a non-increasing Lyapunov function which can have multiple local minima “memories” (Hopfield, 1982, 1984). Another advantage is pattern completion, in other words content addressable memory (Hopfield, 1982, 1984). They also make stimulus specific adaptation more robust to changes in the neural network’s parameters, which is useful for directing attention to novel stimuli (Yarden & Nelken, 2017).

Dye-injection experiments have found recurrent connections in the cat auditory cortex which are disproportionately at ratios that have relatively large greatest common factors (Wang, 2013). A summary of one such experiment is shown in figure 4.1. This indicates likely harmonicity in recurrent connections between cortical neurons.

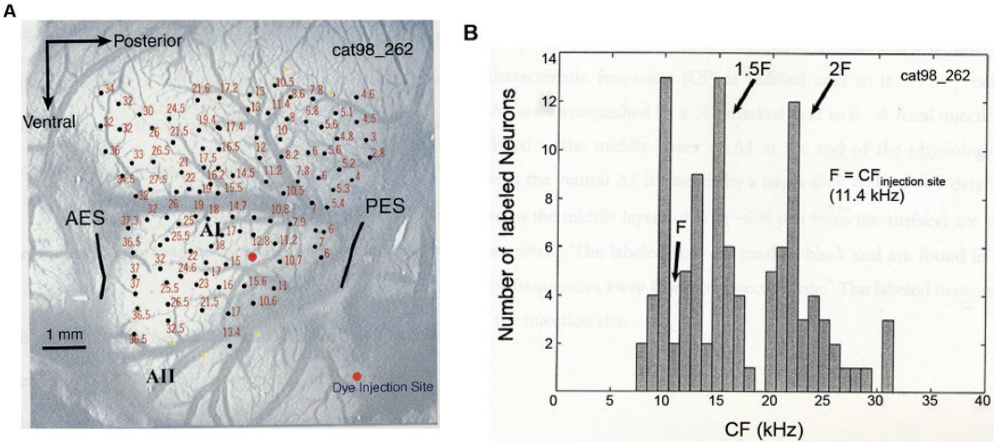


Figure 4.1: A cat auditory cortical dye-injection experiment. The experiment is (shown in A) in which the injection is in a neuron at best frequency 11.4 kHz. The best frequencies in are mapped throughout the cortex, and the location of each labelled neuron is used as a proxy of it's best frequency. These locations are compiled into a histogram. The peaks of neurons at 1.5*11.4kHz and 2*1.4kHz suggests a recurrent connections between harmonically-related units. The fundamental frequency would be at 0.5*11.4kHz with the injection site at twice the fundamental frequency and the connections at three and four times the fundamental. *Figure from (Wang, 2013).*

Training of recurrent networks

Learning processes in the cortex must maintain global stability while extracting statistical trends in the data. The balance between the positive feedback needed to "lock on" to a pattern and maintaining global stability is a challenge for local learning methods. The classic Hopfield network, which only finds second-order correlations, relies on the gain function to saturate for stability (Hopfield, 1982, 1984; Little, 1974). However, the highest-performing machine learning activation functions tend to be "ReLU" or "softplus" functions that don't saturate (Agarap, 2018; Agostinelli et al., 2014). Furthermore, the cortex itself under most conditions operates in the regime way below saturation, instead the rate is maintained at a much lower set-point (Hengen et al., 2013). Thus saturation is not the likely mechanism for global stability.

Learning is spike-timing dependent. It was found that Hebbian reinforcement was much weaker at higher input frequencies (Markram & Tsodyks, 1996), which has a stabilizing effect to the total activity. Volume transmission may also act as a regional or semi-global stabilization method (Zoli et al., 1999).

A wide variety of Hebbian timing functions have been measured. For excitatory to excitatory connections, we have (most commonly) the "classic" patter of strengthening the weight when the preceding synapse fires first and weakening the weight when the order is backwards, with a sharp change near zero timing difference (Caporale & Dan, 2008). Connections from excitatory cells to inhibitory cells followed a reversed timing rule (Caporale & Dan, 2008). Connections from inhibitory cells to excitatory cells followed a variety of rules (Caporale & Dan, 2008). This wide range of rules is likely due to the extreme variety of inhibitory cell types (Bota & Swanson, 2007).

We are most interested in how weights change over a long period of time in a neural network. To do so requires averaging the Hebbian learning over time. For the "classic" case of excitatory neurons, suppose two EPSP's come in at the same time. It is more likely for the neuron to fire and it will fire after said EPSP's and reinforce the connection. Thus this rule, averaged over a bank of stimuli, approximates the Hopfield training of $\langle xx^{\mathrm{T}} \rangle$ from (Hopfield, 1982, 1984).

The story of "time-averaging" is more complex for inhibitory connections. Suppose inhibitory-to-excitatory connections have fixed weight with no learning rule. A concordant EPSP pair hitting the inhibitory neuron will tend to reduce the connection to it, and in turn reduce the strength of inhibition. This is also similar to the Hopfield learning case. However, the inhibitory neurons have a menagerie of learning rules, of which the function is unclear.

The neural network itself must also be simplified into a mathematical model. We assume that the stabilization processes that prevent excessive activity can be approximated as sigmoidal neurons that saturate. Each neuron could in fact represent a microcircuit that stabilizes itself. Also, we assume that the weights have a fixed norm as there are also mechanisms.

We will be considering a simple Hebbian learning with global normalization as well as a Dale's principle inspired learning rule with separate excitatory and inhibitory inputs. Most of the results will focus on the Dale case. This represents an averaging over time of excitatory and inhibitory models.

4.2 Methods

Neural network topology

Recurrent networks had one (hidden) recurrent layer which feed into itself and a single output layer, with 80 frequency channels. A unit sigmoid gain function was used.

Two recurrent network models were used. The sign-agnostic network (80 recurrent neurons) is shown in figure 4.2 and the network with separate excitatory and inhibitory regions (160 recurrent neurons) is shown in figure 4.3.

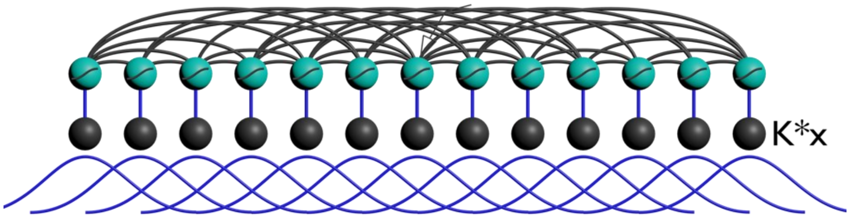


Figure 4.2: A sign-agnostic recurrent network. The input layer convolves the stimuli and is linear, and gives weights of +1 to the recurrent network. The recurrent layer has sigmoid gain, and weights between it can be positive or negative. The best selectivity happens with a fixed K, but variable K (fixed Q-factor) was also considered.

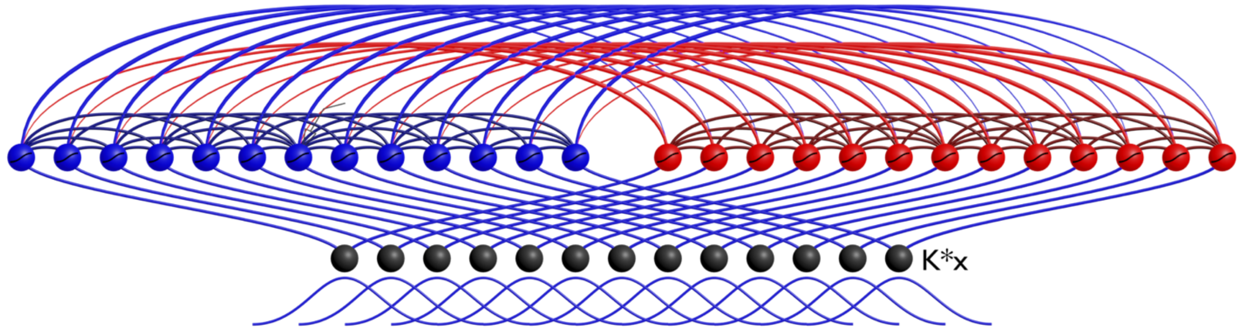


Figure 4.3: A recurrent network respecting Dale's principle. The best selectivity happens with a fixed K, but variable K (fixed Q-factor) was also considered. The input layer convolves the stimuli and is linear, and gives weights of +1 to the recurrent network. The recurrent layer has sigmoid gain. The left half of the layer emit excitatory inputs to itself and the right half. The right half emits inhibitory inputs to itself and the left half. Only identity weights are shown across E vs I for brevity.

Neural network and noiseless model

We use the same model as the feed-forward chapter, but only in the continuous time domain and without noise. However, the expected firing rate isn't any different with or without our (simple) model, so removing the noise doesn't change the mean response.

Subcortical input

Stimuli are represented the same way as Chapter 3 (as a list of locations and energies). However, we considered both fixed with and fixed q-factor subcortical models:

$$Subcortical_{BW} = \frac{1}{Z}\sum p_i exp\left(-\frac{Bf-f_i}{2\sigma^2}\right), \quad Subcortical_Q = \frac{1}{Z}\sum p_i exp\left(-\frac{Bf-f_i}{2(\sigma f_i)^2}\right)$$

The variety of stimuli location-energies we used for training and evaluation are described in other sections of this chapter. Example stimuli are shown in figure 4.4.

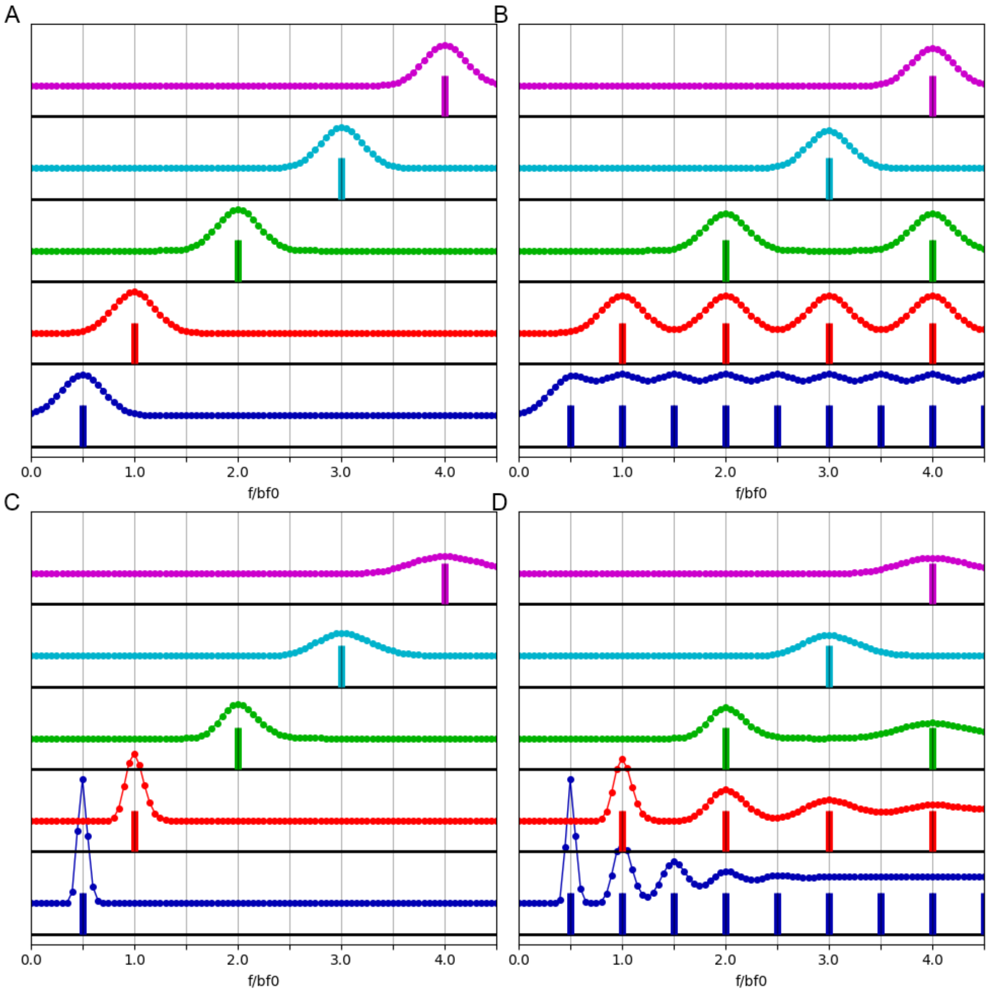


Figure 4.4: Subcortical inputs for stimuli for fixed BW and for fixed Q. A: Tone stimuli, fixed BW. B: Harmonic stimuli (fixed BW). C: Tone stimuli, fixed Q. D: Harmonic stimuli, fixed Q. Only the red stimulus is used for training.

We also define a subcortical temporal adaption model. This model is used when the stimulus over time is a step function. Before the step the input is zero. During a step input, the excitatory and inhibitory input are given by an sum of exponential rises with different time constants:

$$I_{ext,E} = \sum k_{E,i}\left(1 - exp\left(-\frac{t - t_{on}}{\tau_{E,i}}\right)\right), \qquad I_{ext,I} = \sum k_{I,i}\left(1 - exp\left(-\frac{t - t_{on}}{\tau_{I,i}}\right)\right), \quad t_{on} \leq t \leq t_{off}$$

After the input is over, both of these decay:

$$I_{ext,E} = \sum k_{E,i} \left(1 - \, exp\left(-\frac{t_{off} - t_{on}}{\tau_{E,i}}\right)\right) \exp \frac{\left(t_{off} - t\right)}{\tau_{E,i}} , \qquad t > t_{off}$$

$$I_{ext,I} = \sum k_{I,i} \left(1 - \, exp\left(-\frac{t_{off} - t_{on}}{\tau_{I,i}}\right)\right) \exp \frac{\left(t_{off} - t\right)}{\tau_{I,i}} , \qquad t > t_{off}$$

The total *temporal input strength multiplier* is the excitation subtracted from the inhibition, but normalized so that the steady-state input is unity:

$$I_{ext} = \frac{I_{ext,E} - I_{ext,I}}{\sum k_{E,i} - \sum k_{I,i}}$$

We use $\tau_E = \{6.375, 12.75\}, k_E = \{1,1\}, \tau_I = \{15\}, k_I = \{1.75\}$. This generates an onset-sustained-offset curve which is shown in figure 4.10. The subcortical input at each point in time is multiplied by this value.

Hebbian training and testing

We considered two training methods. The first is a Hebbian sign-agonistic method based on the activity of a neural network. Stimuli for this are harmonics with randomized heights. It allows neurons to simultaneously give excitation to and inhibition to other neurons. The recurrent weights are updated in a batch: $W_{ij,t+1} = normalize(\eta\langle y_i y_j\rangle + (1-\eta)W_{ij,t})$, and the output weights are learned using a similar formula: $w_{i,t+1} = normalize(\eta\langle y_{output} y_{recur,i}\rangle + (1-\eta)w_{i,t})$, where $y = g(x-b)$ is the response *x-b* is the state and *g* is a sigmoid gain function, with $normalize(W) = kW/\left|\left|W\right|\right|_{frobenous}$ for a user-prescribed *k*. With $\eta = 0.8$ and 8 batched steps we get the network to converge to well within 5% of it's final value. The effects of this training are summarized in figure 4.5.

We also have a sign-constrained learning rule that is based on the stimulus statistics. Neurons are divided into two equal groups, excitatory and inhibitory, given the recurrent weights this rule (in this case x represents the subcortical input). We first define a ridge matrix that penalizes connections that are far away from each-other:

$$W_{Ridge,\ i,j} = \exp\left(-\frac{(i-j)^2}{\sigma_{BW,ridge}^2 + \left(0.5(i+j)\sigma_{Q,ridge}\right)^2}\right)$$

We then define the four weight quadrants, of which three use Hebbian or anti-Hebbian learning:

$$x_+ = x - \min(x)$$
$$W_{EE} = k_{EE} normalize(\langle x_+^T x_+ \rangle)$$
$$W_{EI} = k_{EI} normalize(\langle x_+^T x_+ \rangle)$$
$$W_{IE} = -k_{IE} normalize(W_{Ridge})$$
$$W_{II} = -k_{II} normalize(\langle x_+^T x_+ \rangle)$$

The use of the stimulus statistics produces similar results as using the neural activity. However, it is much easier to control. A set of parameters that make highly selective harmonic template units are listed in the results section of this chapter. The training stimulus for this is a single harmonic stimuli with constant heights; randomized heights would produce a significant ridge. Many assays required computing a steady state response to various stimuli. This was estimated by fixing the subcortical input and measuring the network's activity after 20 time constants.

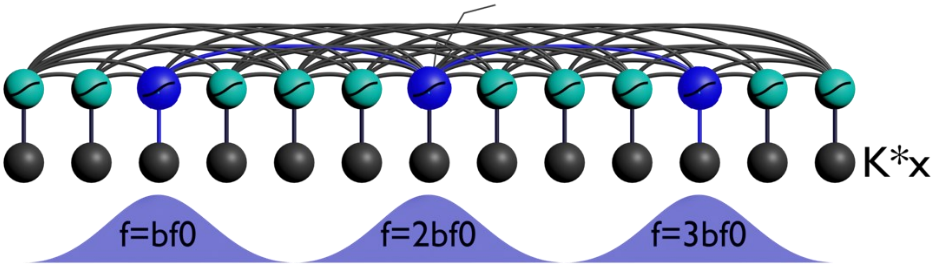


Figure 4.5: How training produces an egg-crate network. When presented with a harmonic stimulus, the subcortical inputs excite neurons at integer multiples of the fundamental frequency. That cause Hebbian reinforcement of the weights between them.

Eigen-analysis of linearized dynamics over time

Recurrent networks admit a non-trivial Jacobian in between the state of the network and the rate of change of the state. The Jacobian is: $J = \frac{\partial\left(\frac{\partial y}{\partial t}\right)}{\partial y}$, where y is the network state.

We are most interested in the Jacobian at the middle of a hysteresis cycle loop; see figure 4.6. This gives us at least one unstable eigenvector.

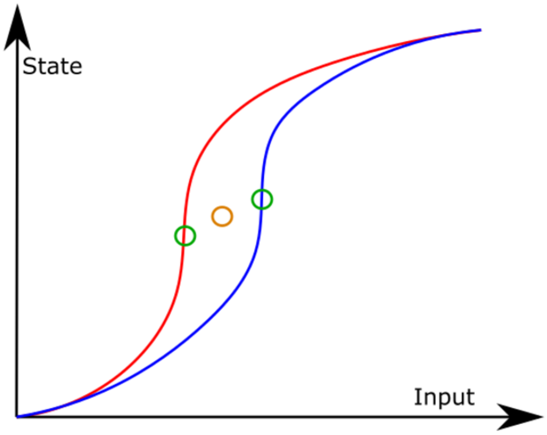


Figure 4.6: Finding the point in the approximate center of the hysteresis loop. Shown is a hysteresis loop (the rising curve is blue, the falling curve is red). The green circles are the points of fastest change. The orange circle is the average of the green circles and is where the Jacobian gets calculated.

4.3 Results

We are able to produce harmonically selective units that under the experiments in (Feng & Wang, 2017) would resemble some of the harmonic template units measured. The Hebbian learning rule produces egg-crate recurrent weights. We also have hysteresis, pattern completion, and/or oscillation depending on the parameters used; these are hallmarks of recurrent networks.

Weights are an eggcrate

The weights that are trained under Hebbian rule resembles egg-crates, see figure 4.7 and figure 4.8. These egg-crates have excitatory connections between integer multiples at inhibitory connections between half-integer multiples of the fundamental frequency. The half-integer-integer connections are saddles in the egg-crate and have nearly zero weight. The sign-agnostic case with a fixed bandwidth generate the "classic" egg-crate shown in figure 4.7. The sign-agnostic case with a variable bandwidth also generates an egg-crate but the peaks get broader (and the "valleys" get narrower) at higher frequencies as seen in figure 4.8.

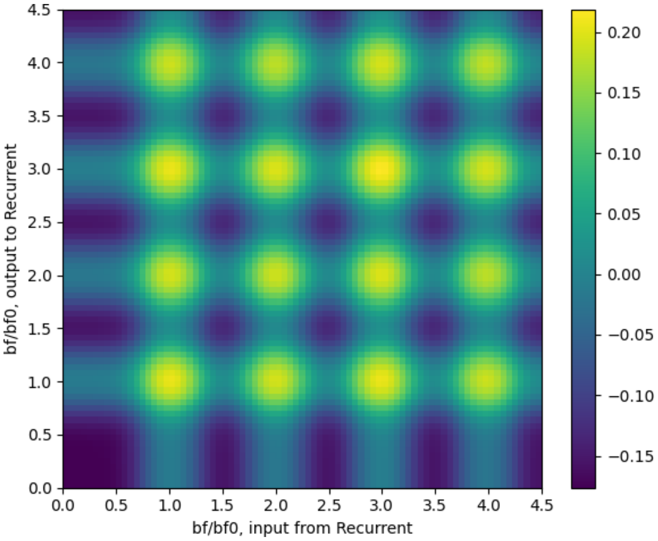


Figure 4.7: The recurrent weights for a fixed bandwidth sign-agnostic case. These weights are an egg-crate.

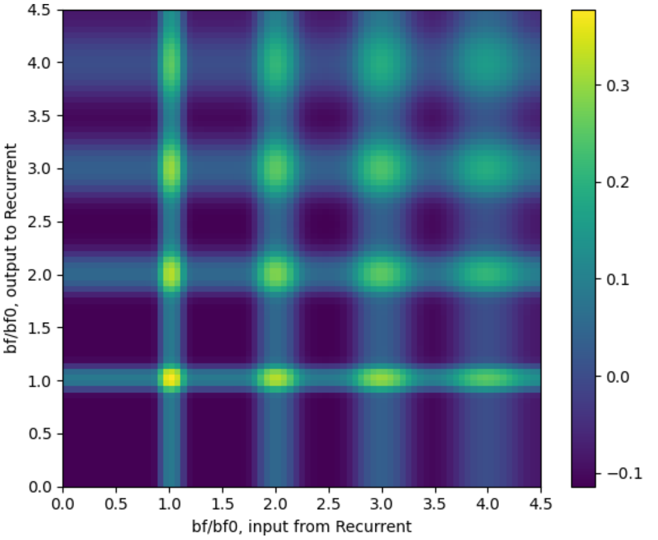


Figure 4.8: The recurrent weights for a fixed Q-factor sign-agnostic case. These weights are eggcrate-like but with a "variable width".

The statistical-based Dale's principle learning case, under our training rule, produces a weight matrix with three egg-crate sections (one of which is inverted) and an upside-down ridge shown in figure 4.9.

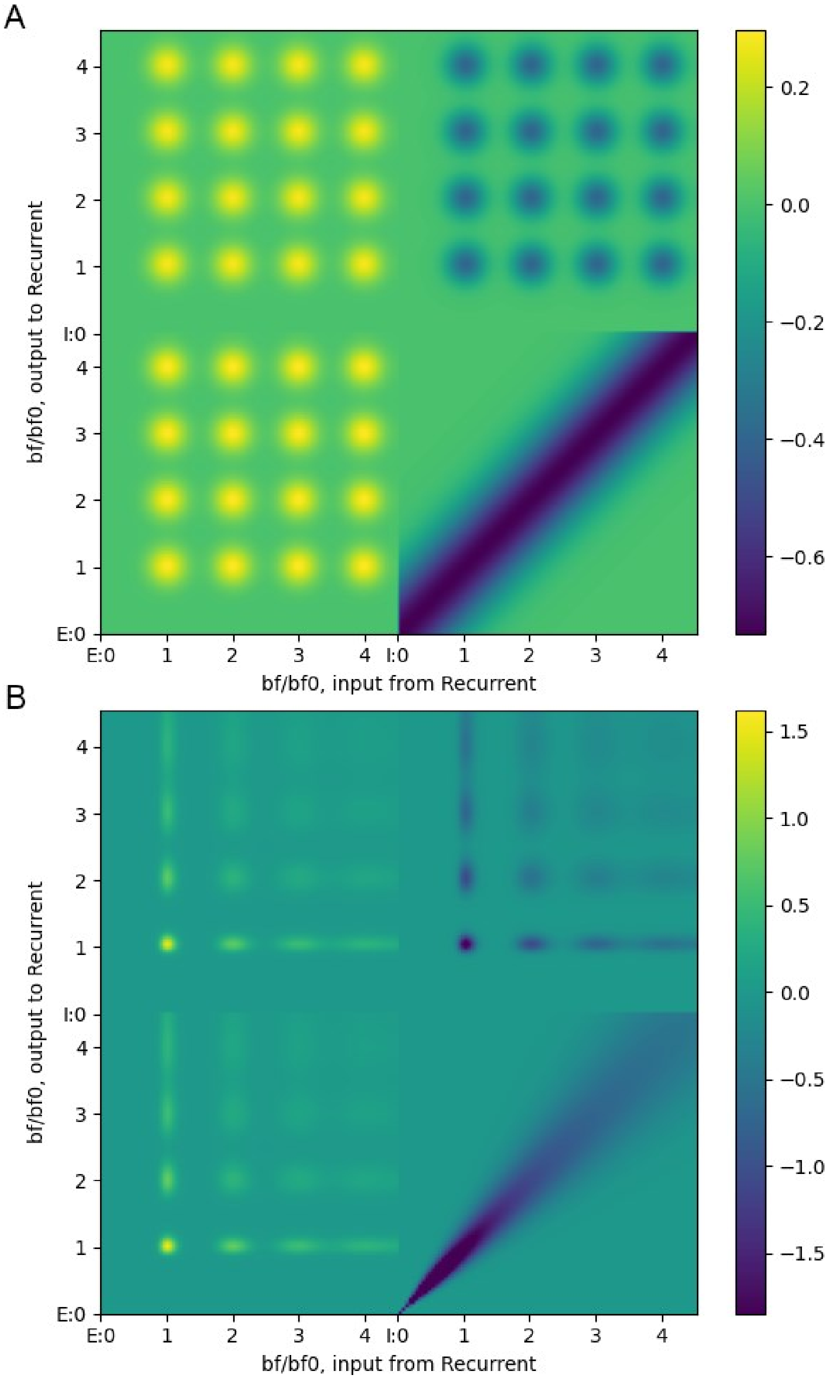


Figure 4.9: The recurrent weights for the Dale's principle case. A: for a fixed bandwidth. B: For a fixed Q-factor (with a restricted color range to show the shorter, broader peaks more clearly).

The following results are all for the fixed-bandwidth Dale's principle case. Although all cases can produce a degree of harmonic selectivity, the aforementioned case was found to give the best results with the correct choice of parameters.

Response to a fixed stimulus over time

We tested the network on a fixed harmonic stimulus at the training f0 = Bf0. The stimuli before and after subcortical processing are shown in figure 4.10.

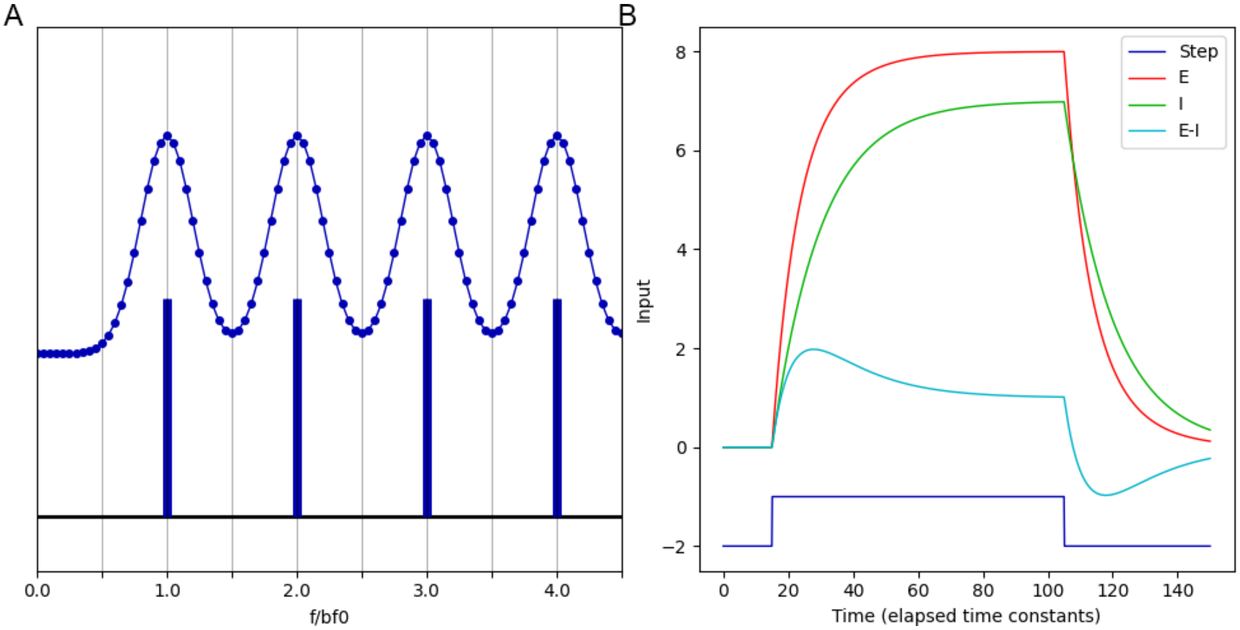


Figure 4.10: Presenting a fixed stimulus to the network. A: The stimulus and subcortical input for a harmonic network. B: The subcortical onset/transient/offset model. The blue curve in A is multiplied by a time-dependent scalar, which is The cyan curve in B.

The network activated very strongly to this stimuli near integer multiples of Bf0. The state of the network shows strong depolarization for both the excitatory and inhibitory neurons near the integer multiples of Bf0 as shown in figure 4.11. The activity shows a similar, if more confined, pattern as seen in figure 4.12.

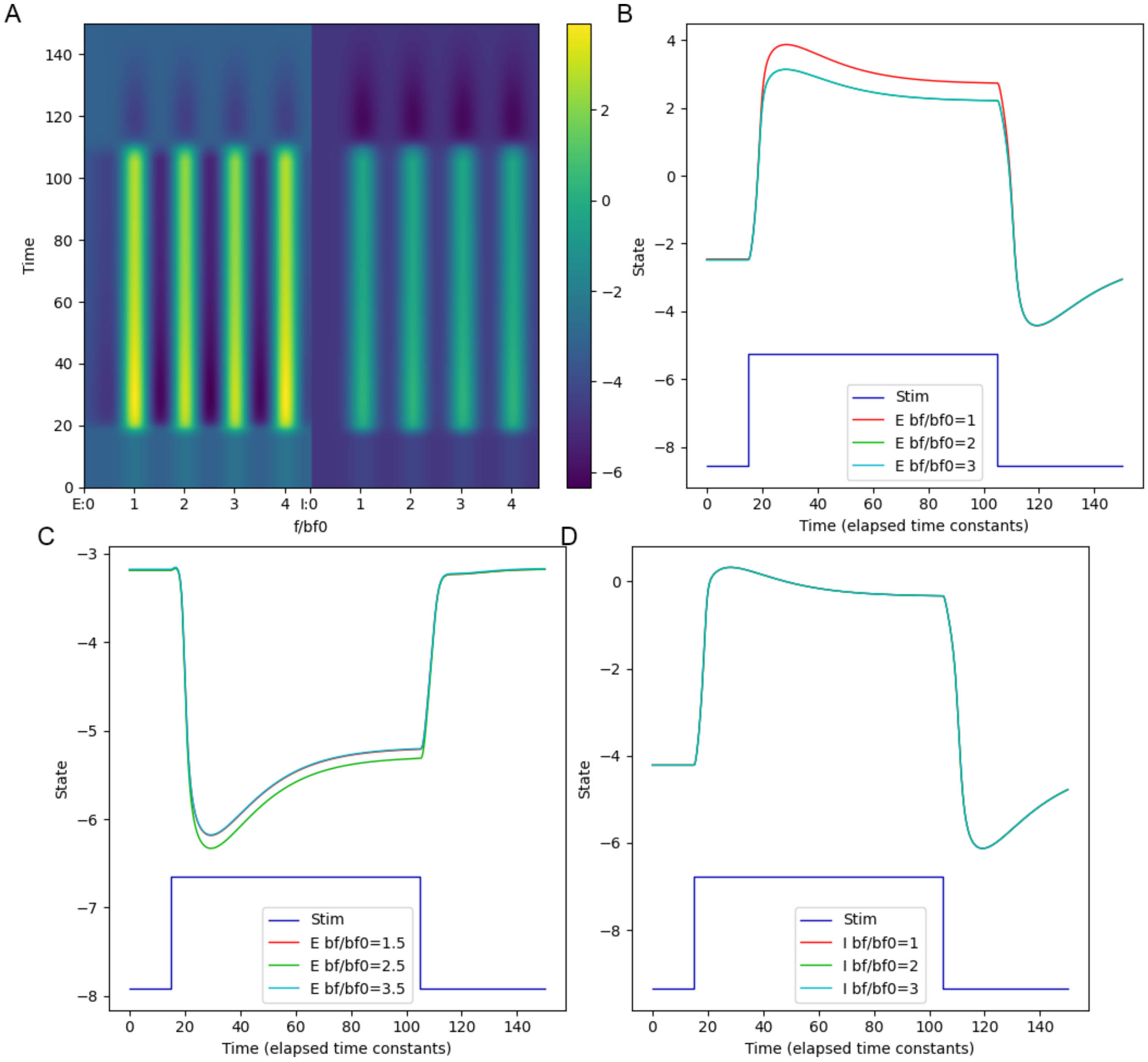


Figure 4.11: The network state over time to a fixed harmonic stimulus. We use Dale's principle under a fixed bandwidth. A: The state over time of the entire network, values at zero indicate the neuron is firing at 50% it's maximum rate. B: The state over time of three integer-multiple excitatory units. C: The state over time of three half-integer multiple excitatory units. D: The state overt time of three integer-multiple inhibitory units. It is similar to the excitatory units in shape but a weaker effect, as is the half-integer multiple inhibitory units vs the half-integer excitatory units (not shown).

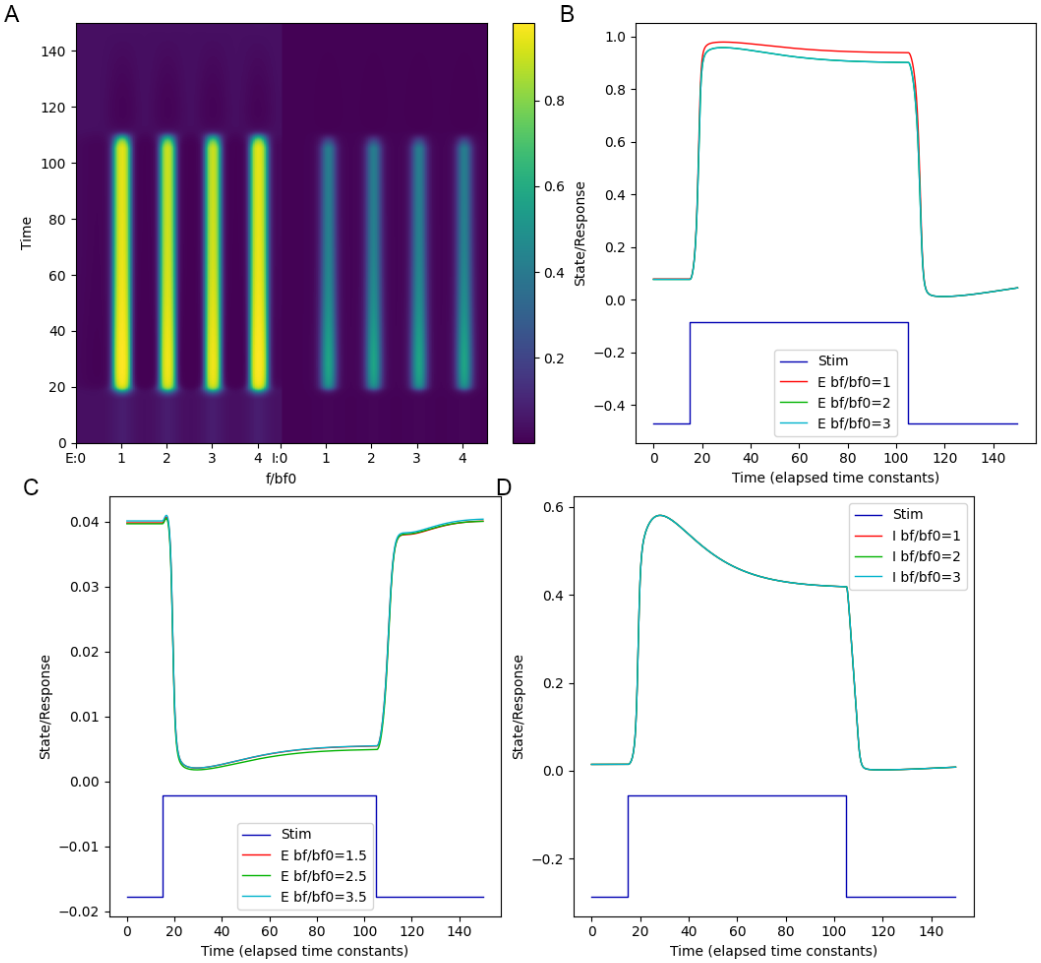


Figure 4.12: The network response over time to a fixed harmonic stimulus. We use Dale's principle under a fixed bandwidth. A: The response over time of the entire network. B: The response over time of three integer-multiple excitatory units. C: The response over time of three half-integer multiple excitatory units. D: The response over time of three integer-multiple inhibitory units. It is similar to the excitatory units in shape but a much weaker effect, as is the half-integer multiple inhibitory units vs the half-integer excitatory units (not shown).

Response to tones, harmonics, and mistuned harmonics

Tones, harmonic, and mistuned harmonic stimuli, shown in figure 4.13, are used to assess the networks selectivity to the presence and the tuning of harmonics. The network's integer-multiple units prefer harmonics over tones *or* over mistuned stimuli, see figure 4.14.

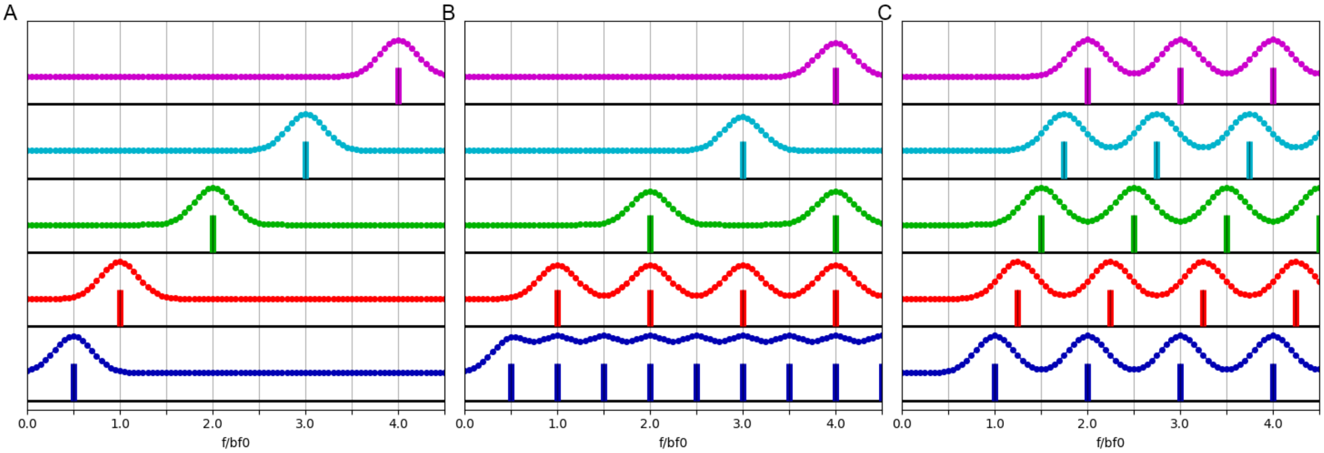


Figure 4.13: Stimuli and subcortical input to assess harmonic selectivity. We use tones (A), harmonics (B), and mistuned stimuli (C).

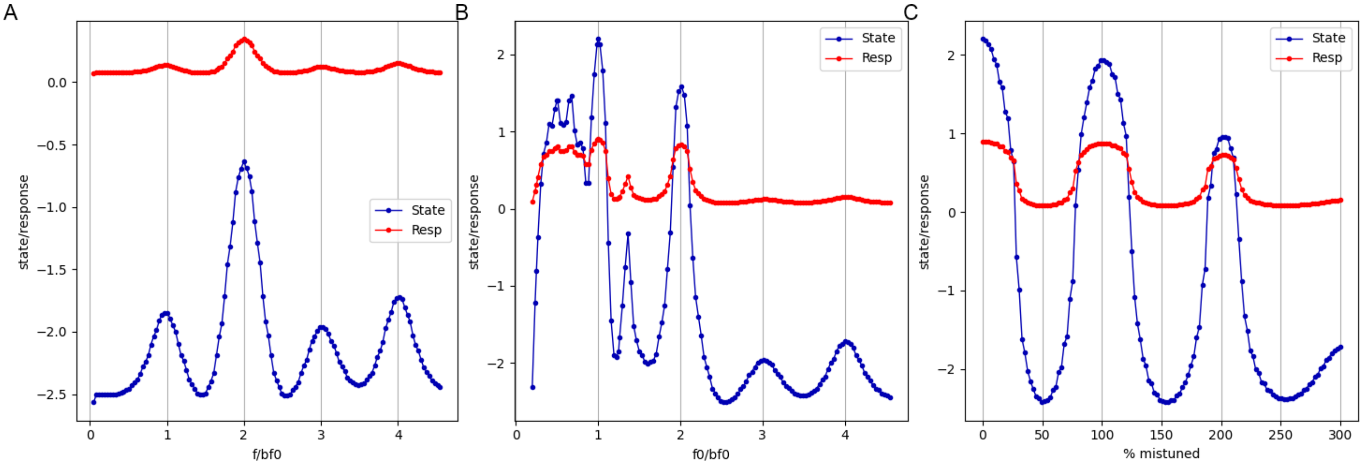


Figure 4.14: Response maps to tones, harmonics, and mistuned stimuli. The neuron is excitatory with bf=2bf0. The neuron responds to harmonics(B) at Bf0 much more than to tones (A) and mistuned stimuli (C).

Facilitation index (preference for harmonics over tones) and periodicity index (preference for harmonics over mistuned complexes) are the metrics of harmonic selectivity used in (Feng & Wang, 2017). These are shown in figure 4.15 with their histogram in figure 4.16. Our integer-multiple units meet both criteria, as do units near integers. The inhibitory units actually are somewhat *more* harmonically selective. The half-integer multiple units and their nearby neighbors aren't selective at all to harmonics.

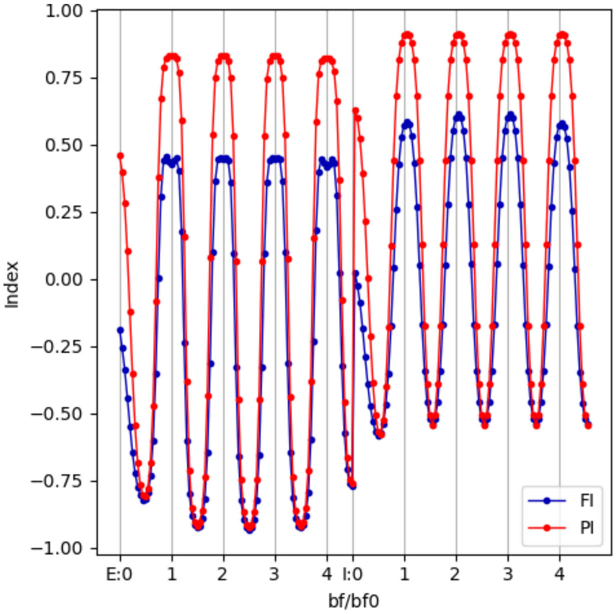


Figure 4.15: The FI and PI of each unit. We only allow for harmonics at Bf0 for the purposes of calculating FI and PI. FI>0.33 and PI>0.5 make a neuron qualify as a harmonic template unit. These are based on the neuron's *response* not the state.

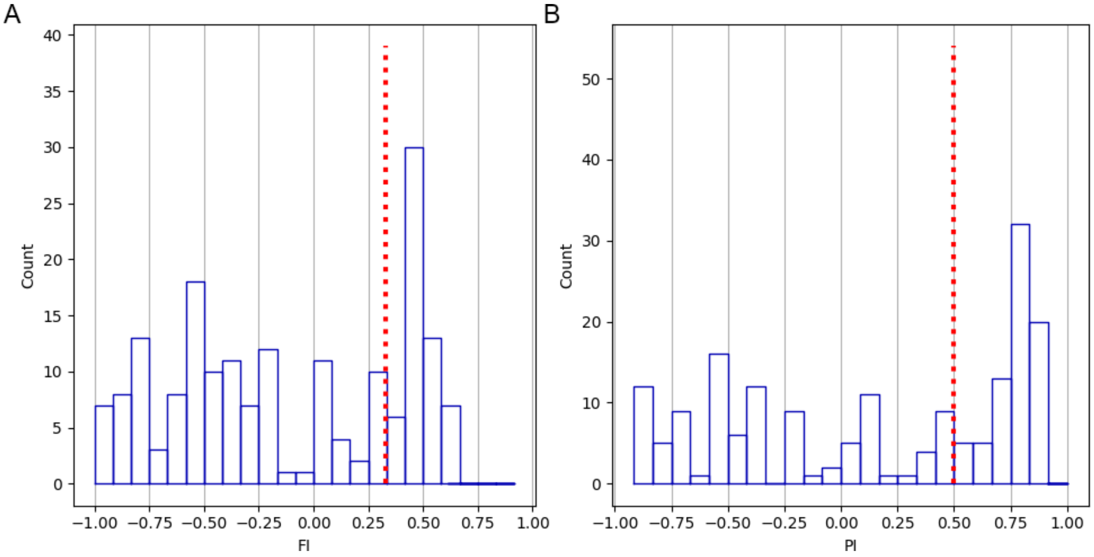


Figure 4.16: The histogram of FI and PI values. Vertical dashed lines indicate the cutoffs above which a unit is considered a “harmonic template unit”.

<u>Random harmonic stimuli and regression weights</u>

Random harmonic stimuli measure the network’s response to randomized harmonic stimuli, but with a fundamental frequency of half of Bf_0, as done in (Feng & Wang, 2017). These stimuli are shown in figure 4.17. Regression models are then ran on the vector of responses against the matrix of stimuli energies.

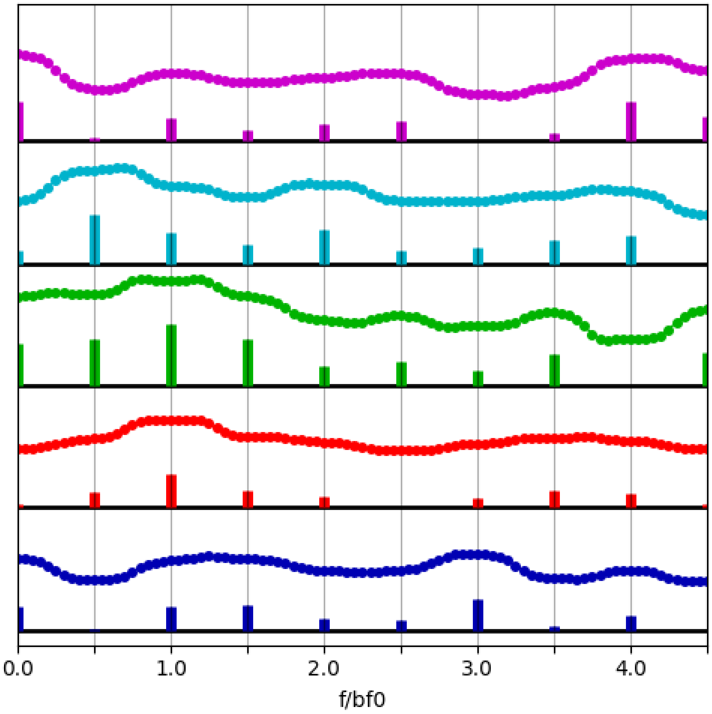


Figure 4.17: A sample of random harmonic stimuli. They are all at half of the network's b_{f0}, with randomized amplitudes per component.

The regression models estimate the linear and quadratic RHS weights of the neuron. One such neuron, which is excitatory and at an integer multiple of Bf0, is shown in figure 4.18. The linear weights form a spectral sieve.

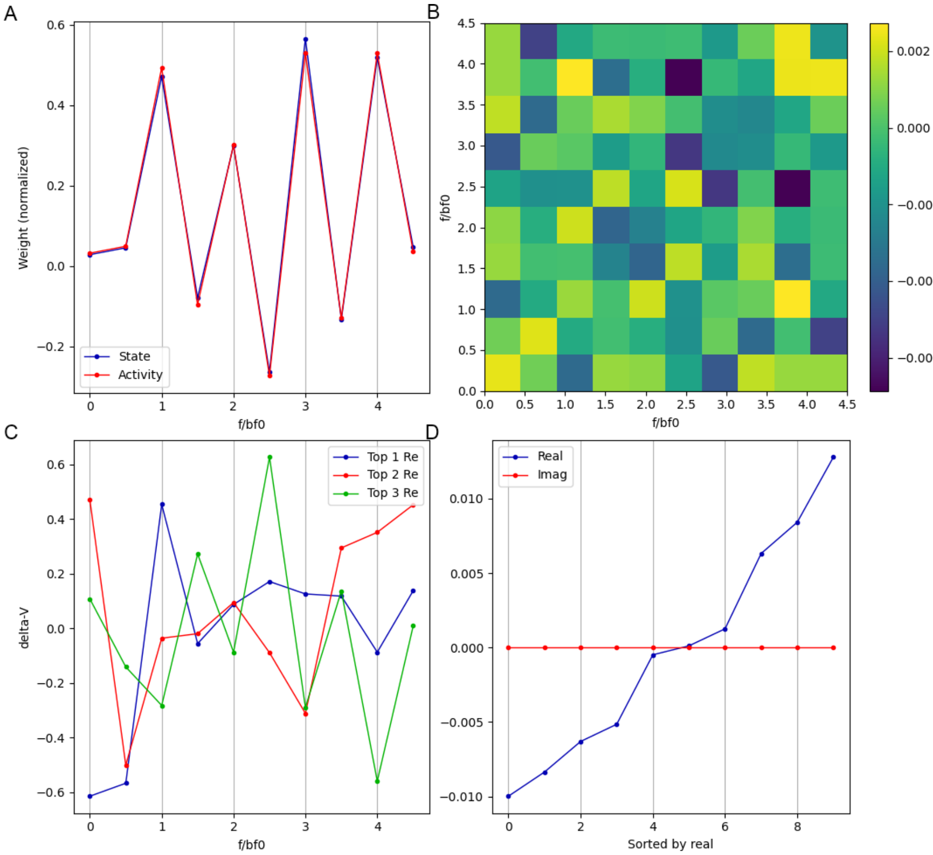


Figure 4.18: The RHS linear and quadratic weights for a single unit. A: The linear weights for both the state and activity. B: The quadratic weights for the activity (the weights for the state are very similar in shape). The top three eigenvectors of the quadratic weights are shown in (C) while (D) shows all the quadratic eigenvalues.

All of the units at integer multiples have sieve-like patterns, whether they are excitatory or inhibitory, see figure 4.19. Units at half-integer multiples usually have a center-surround response superimposed on a small “anti-sieve” like response as shown in figure 4.20. Both patterns, as well as other patterns, were found in (Feng & Wang, 2017).

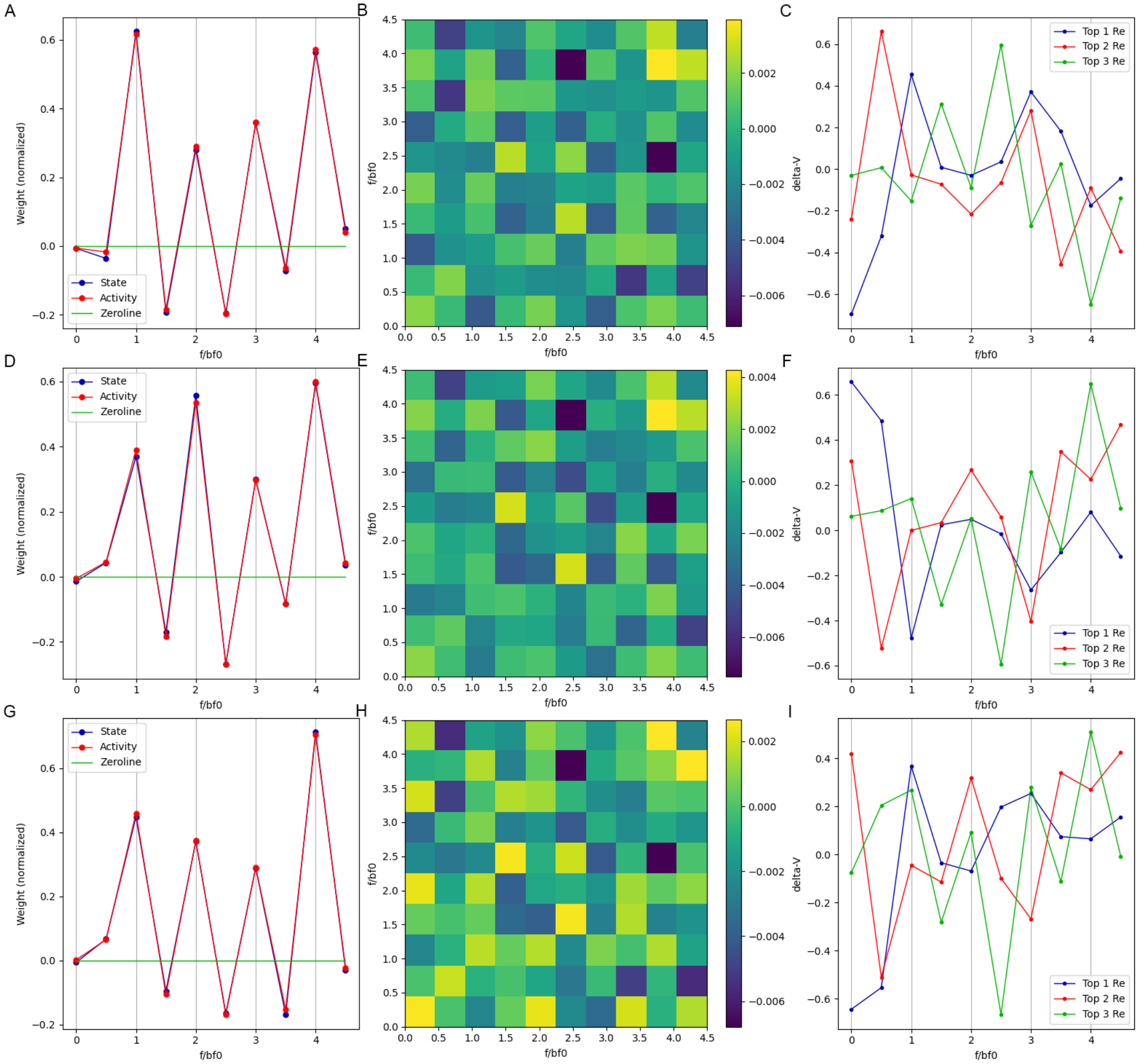


Figure 4.19: Integer-unit RHS weights and eigenanalysis. This is for units at $b_f=b_{f0}$ (A,B,C), $b_f=2b_{f0}$ (D,E,F), and $b_f=4b_{f0}$ (G,H,I). Each triplet of subfigures uses the same format as A,B, and C of the previous figure.

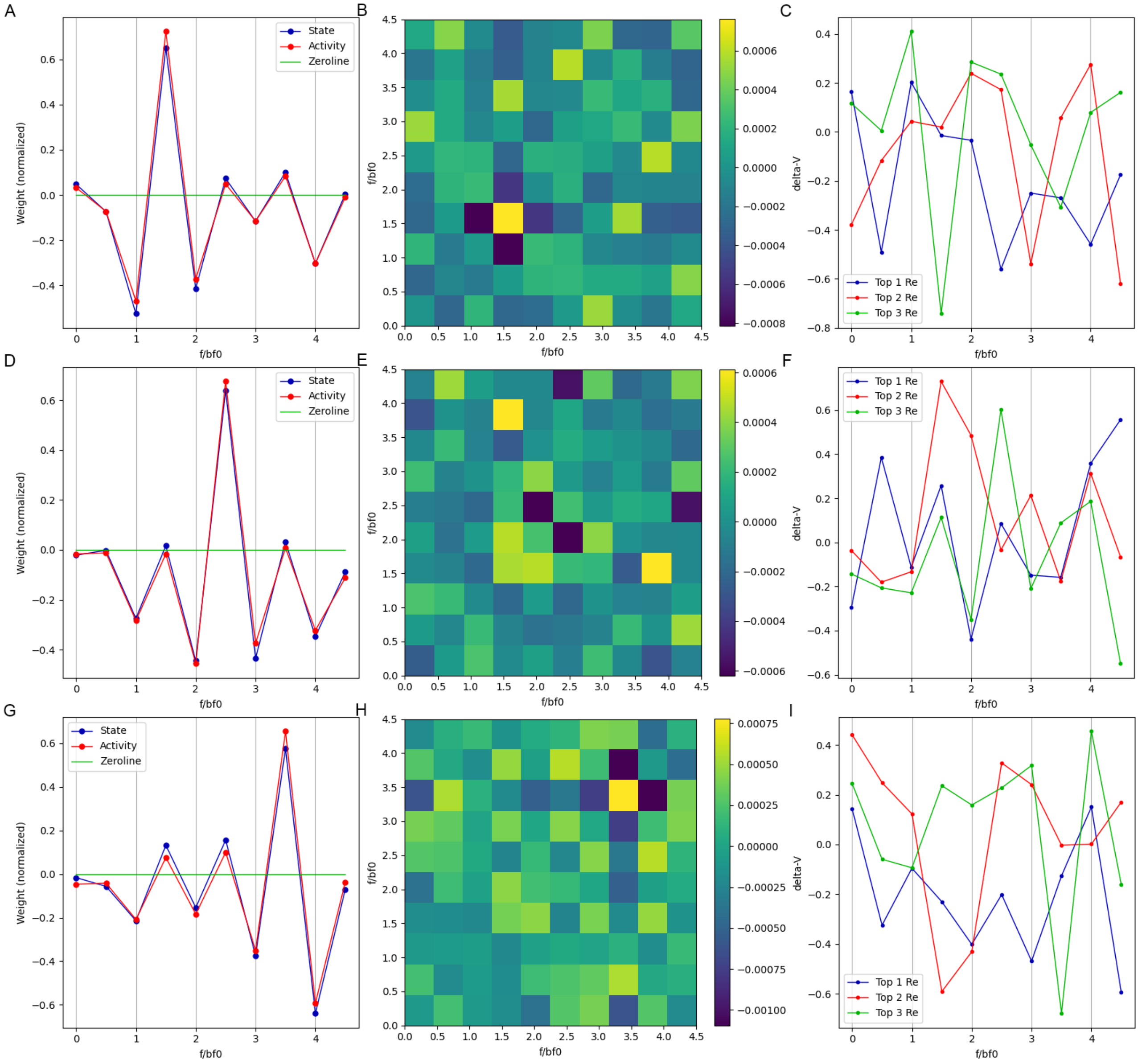


Figure 4.20: Halfinteger-unit RHS weights and eigenanalysis. This is for units at $b_f=3/2b_{f0}$ (A,B,C), $b_f=5/2b_{f0}$ (D,E,F), and $b_f=7/2b_{f0}$ (G,H,I).

The notable difference between our simulated RHS weights and the experimental weights is that we are using a linear scale rather than dB because we assume a significant amount of subcortical compression implicit in the model. The weights show a sieve-like pattern with an extra strong component at the BF, which is similar to some examples in (Feng & Wang, 2017).

Despite being a linearization, the shape of the weights did not depend strongly on the location around which RHS is taken.

Reponses to a rise-fall stimuli and eigen analysis in the hysteresis loop

A *rise-fall* stimulus is a stimulus that gradually gets stronger and weaker throughout time, but is fixed in terms of its relative spectral components. The network shows hysteresis when given this stimulus, suddenly turning on and off (no matter how slow the ramp is), as seen in figure 4.21. The level in which the network turns on is higher than the level in which it turns off.

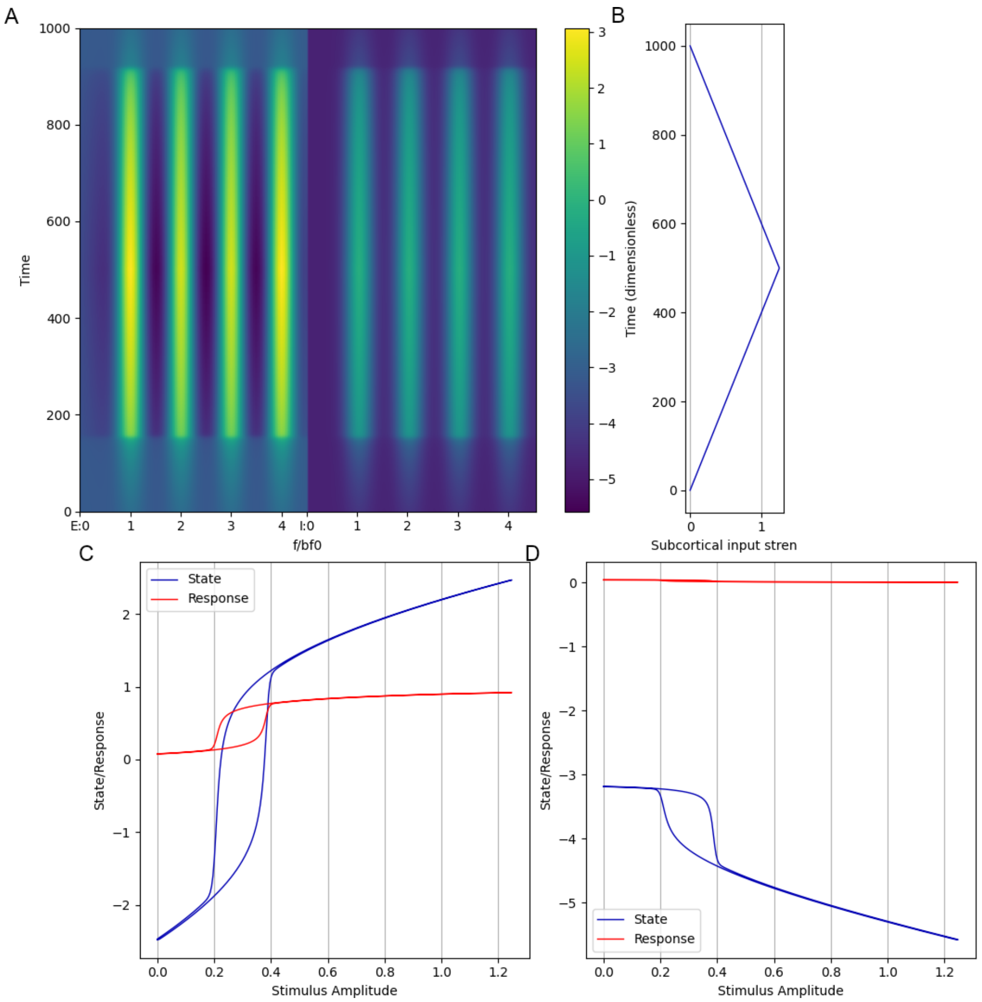


Figure 4.21: The state and response to a rise-fall stimuli. A: The state over time. B: The stimulus intensity over time (no adaptation model is applied). C: The state and response for a neuron at BF=2Bf0. D: The state and response for a neuron at BF=5/2Bf0.

In the center of the hysteresis loop there is one unstable eigenvector shown in figure 4.22. This eigenvector is a sieve for both the excitatory and inhibitory parts of the network. There is also an eigenvector that is made more stable by the recurrent weights, it is an anti-sieve in the excitatory part and a sieve in the inhibitory part. All the other eigenvectors are degenerate and represent the leakage term in the network dynamics.

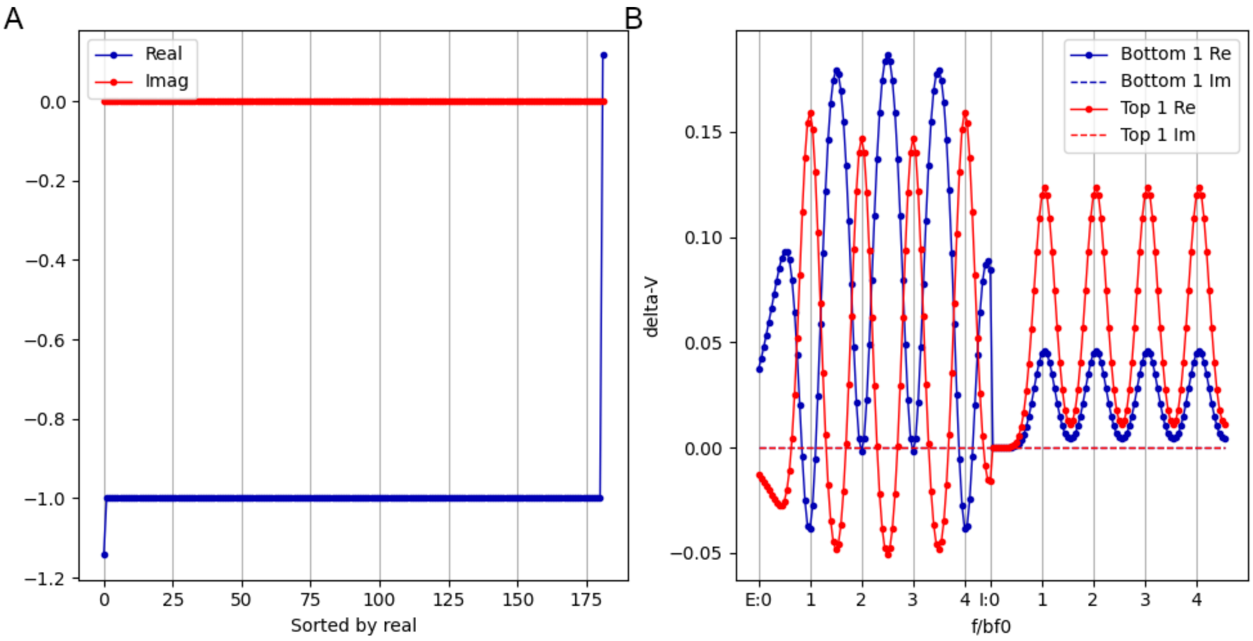


Figure 4.22: Eigen analysis in the middle of the hysteresis loop. A: The eigenvalues of the state Jacobian at the fastest change. There is one positive eigenvalue indicating first-order instability. B: The largest and smallest eigenvectors (the imaginary part is shown, but is zero for both). This is based on the *state* of the neuron, which more clearly shows the dynamics of the network in certain ways than the response.

Pattern completion

Our network is able to have similar activation patterns to stimuli missing a single component, illustrated in figure 4.23, than to stimuli with all their components present. This represents a filling-in of the missing component.

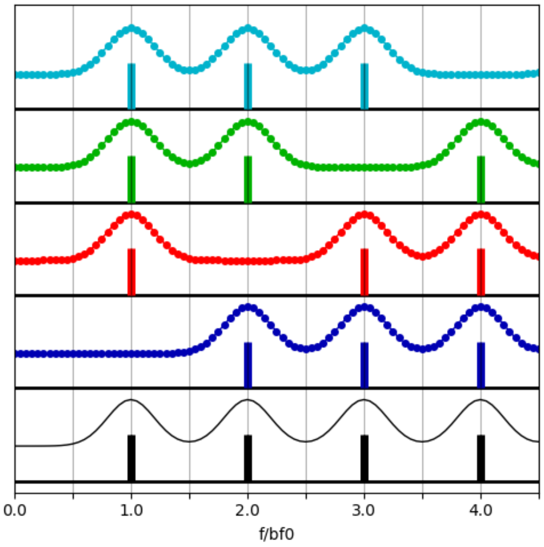


Figure 4.23. Stimuli with missing components. The black curve represents the complete harmonic stimulus, while the other curves are stimuli missing one component.

The filling in produces enough depolarization in the excitatory neurons to drive to them well above their threshold; see figure 4.24. The filling in effect is delayed by about 3 time-constants as shown figure 4.25.

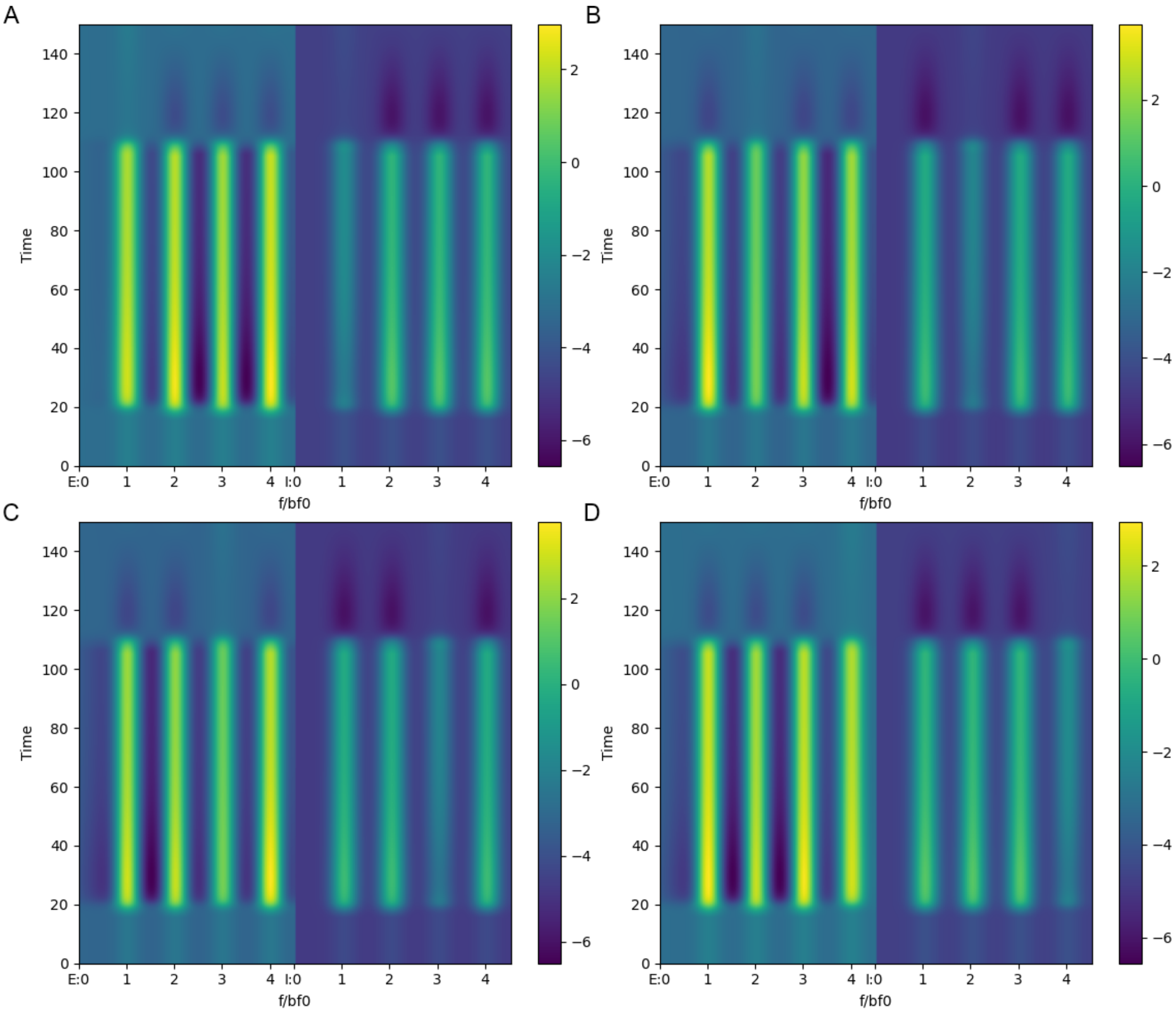


Figure 4.24: The network state for stimuli with missing components. A,B,C,D are missing the fundamental, second component, third component and fourth component respectively.

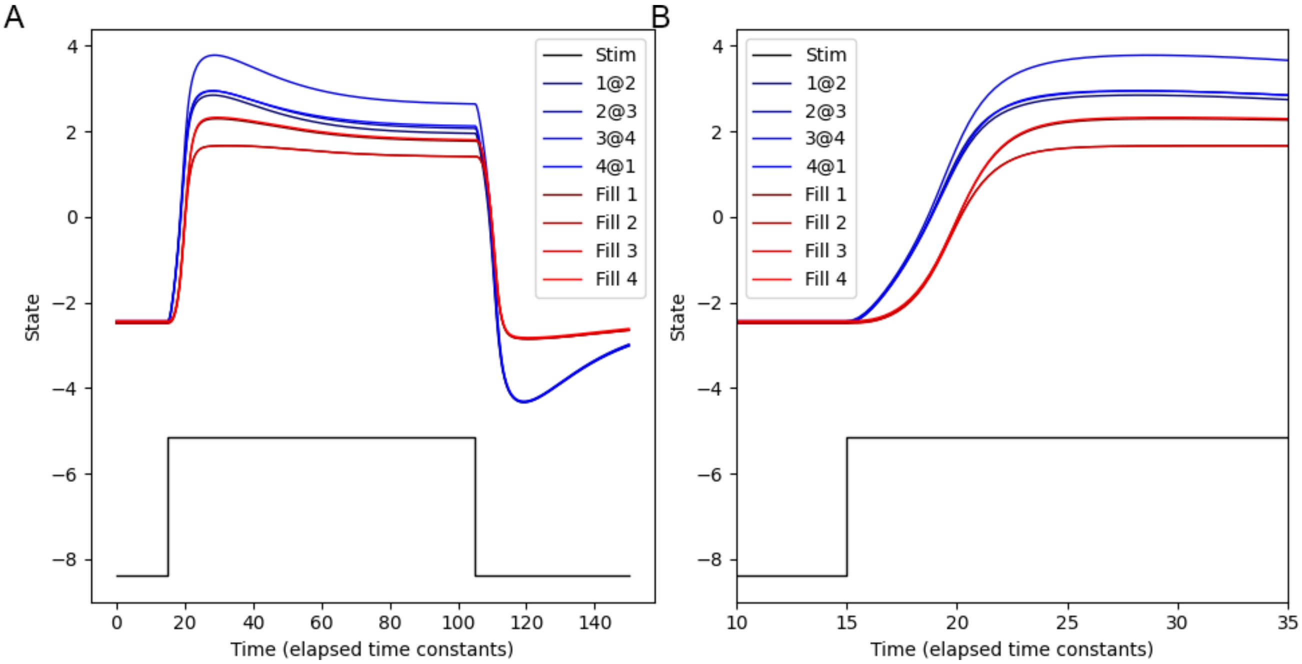


Figure 4.25: Selected neural states for stimuli with missing components. Red indicates neurons that are at the location of the missing components. Blue indicates neurons that are not at the location of the missing components. The red curve rise almost as much as the blue curves, indicating a filling in which is slightly delayed. A: The whole trajectory. B: Same as A but zoomed in, showing the rising part only in order to emphasize the delay in the red curves.

<u>Oscillation</u>

Some parameter sets showed oscillation, in particular the parameters with strong inhibition-to-excitation as well as strong excitation-to-inhibition. An example of these parameters is shown in figure 4.26. This was sometimes combined with hysteresis.

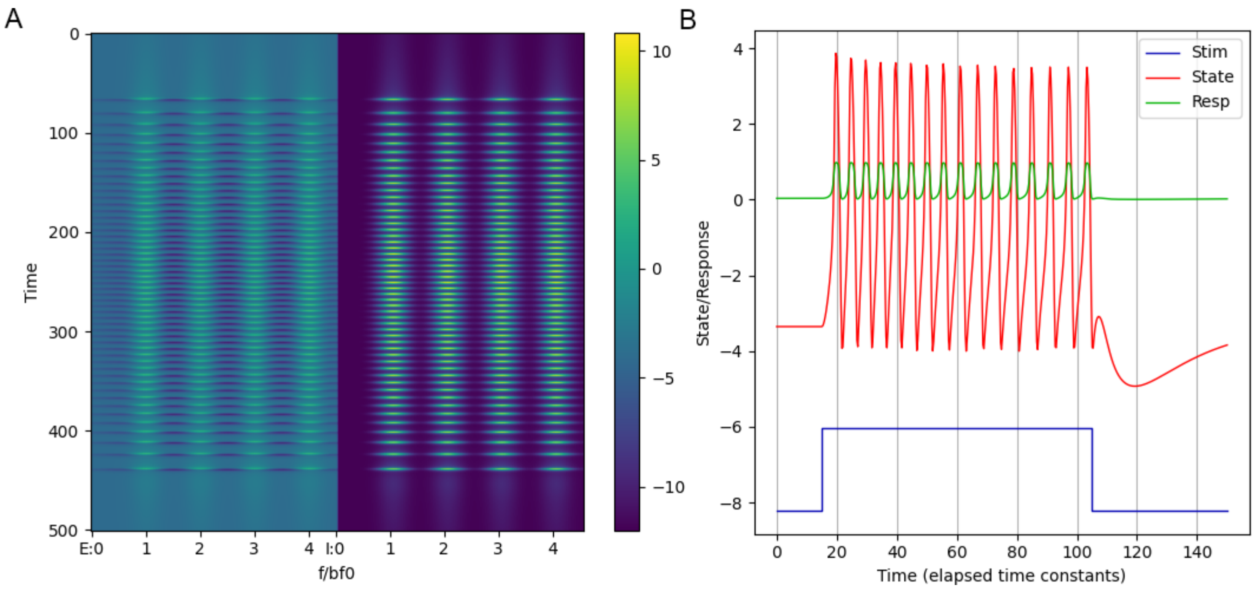


Figure 4.26: Oscillation under a different parameter set. The feedbacks from excitation to inhibition back to excitation are very high. A: The network state under a rise-fall harmonic stimulus, at higher intensities there is a shorter period of oscillation. B: The response under a fixed harmonic stimulus.

Sensitivity to parameter changes

Changing one parameter at a time has usually a broadly excitatory effect or inhibitory effect, see table 4.1. Excitatory effects also strengthen hysteresis (or even keep the network permanently activated), while inhibitory effects weaken it. A 10-20% change in many cases was sufficient to cause significantly different behaviors.

Feedback strength is the 1-norm of the weights divided by the number of neurons in the destination layer (number of channels). This is thus the *average* weight from all neurons feeding into a given neuron *averaged* over all destination neurons.

Parameter	Default value (oscillation value)	Effect of increasing (decreasing has opposite effect)
Feedback E→E	5.25 (10.0)	Strongly excitatory
Feedback E→I	5 (20.0)	Strongly inhibitory
Feedback I→E	20 (also 20)	Strongly inhibitory
Feedback I→I	7 (0.0)	Weakly excitatory
Threshold E	3 (4)	Strongly Inhibitory
Threshold I	4.75 (12)	Strongly Excitatory
Stimulus strength*	16 (12)	Weakly excitatory
Inhibitory bandwidth	25% of total range (also 25%)	Little effect for modest changes.

Table 4.1: Effects of changing individual parameters. A feedback of one means the average row 1-norm is one, i.e. $\langle \sum_j W_{quadrant,ij} \rangle = \pm 1$. The default values are used for most results, but the "oscillation value" is only used for figure 4.26. **for the default parameter set, it is important to set the stimulus strength such that the center of the hysteresis loop for the best-harmonic stimulus is about 30% of the strength used for assessing FI and PI.* A pure tone has the same strength as a single component of a harmonic stimulus, which means that pure tones have several times less total energy overall.

Secondary parameter effects

We want to separate out effects in directions that are orthogonal to overall network sensitivity/excitability: we seek to quantify the average excitability of the network, and project the parameter space down to a parameter subspace with a moderate "excitability". We then can explore this subspace.

We have seen that the network has a varying amount of hysteresis. At low values the output is a soft sigmoid. At moderate values the output begins to harden, and at high values it has a classic hysteresis loop.

Another secondary effect is how strong the center component of the RHS weights is relative to the other components. Excitatory-to-inhibitory weights make keeps the RHS weights more sievelike and less center-surround.

Eggcrate FI is limited, unless variation in population is accounted for

Pure tones need to activate it enough to trigger the positive feedback, so an FI of 0.5 was the highest reasonable value without extreme sensitivity to parameters and/or very strong hysteresis. A real neural population has a *range* of thresholds for units at any given bf. The higher threshold ones would have FI and PI up to 1, even if the lower threshold neurons are needed. Indeed, (Feng & Wang, 2017) found a continuous range of PI and FI values and thus needed to hand-select the cutoffs.

Covariance learning as an alternate rule

Our learning rule is $\langle x - x_{min}\rangle\langle x - x_{min}\rangle^T$. However, we also considered a “covariance” learning rule which is $\langle x\rangle\langle x\rangle^T\text{-}min(\langle x\rangle\langle x\rangle^T)$. The former is more realistic for Dale’s principle based learning, but the latter is also plausible. It produces a small secondary egg-crate at half integer multiples, as seen in figure 4.27, and is not as selective to harmonic stimuli.

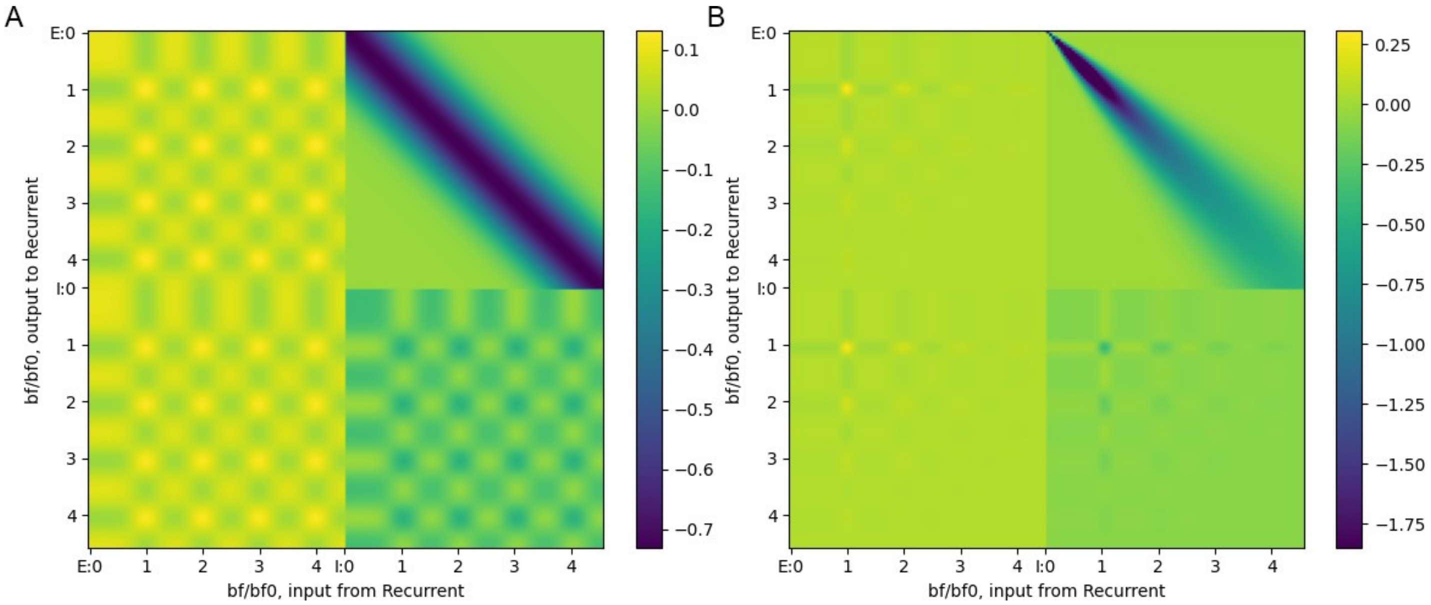


Figure 4.27: The covariance learning rule. This yields a smaller secondary egg-crate at half-integer-half-integer locations. A: Learning under fixed BW. B: Learning under fixed Q (with a restricted color range to show the less tall peaks more clearly).

Response of a combined feedforward and recurrent network

We also modelled combined feed-forward recurrent networks, where the recurrent network is used as the input to a single output unit that is weighted as a sieve. These networks perform only marginally better than pure feedforward networks for harmonicity detection, but show hysteresis, pattern completion, and oscillation like our pure recurrent networks. There is no qualitatively different behavior these "hybrid" networks have compared to the purely recurrent networks.

4.4 Discussion

We demonstrate an alternative mechanism for selectivity to harmonic sounds which is based on "eggcrates" rather than sieves. Despite this, neurons still resemble sieves in terms of estimates of the RHS weights of the neuron. The recurrent nature of the network adds a variety of effects.

Input dynamic range

Our model was evaluated for inputs of a fixed amplitude and does not work across a wide dynamic range of inputs. However, we don't include any compressive nonlinearities at the subcortical level. Fortunately, the type I units in the inferior colliculus (of which our subcortical model is based on) are usually very "flat" above a certain dB level, see (Schreiner & Winer, 2005) for an overview of the various unit types and their response properties.

Egg-crates as an alternate to sieves

Our network behaves much like a sieve in that it has high periodicity and facilitation indexes as well as sieve-like RHS weights. In a real neural network there would be a range of thresholds as well, which would produce some neurons with higher thresholds and thus even higher indexes of selectivity and sparser response overall.

Our network displays hysteresis, a fundamental feature of recurrent networks, which benefits from the subcortical transient response. Suppose a stimulus produces a sustained response for which the network is in a bi-stable state. The stimulus itself wouldn't be able to turn the network from "off" to "on" but it would be able to sustain the network in the "on" state. However, the strong onset transient response makes it likely the input exceeds the bi-stable region and forces the network to turn "on". Similarly, the offset response produces an inhibitory stimuli that may switch the network "off" in cases where the bi-stable region includes the "background" or "distractor" stimuli.

Pattern completion and oscillation are also important properties of recurrent networks. Pattern completion produces content-addressable memory typical of Hopfield networks (Hopfield & Tank, 1985) which allows reconstruction of a stimulus even if parts are missing. Oscillation may be physiological or pathological as in seizures and it does not play an obvious role in harmonicity detection.

A hybrid feed-forward and recurrent network model, in which the recurrent layer feeds into a feedforward network, was found to behave similarly to a recurrent network and thus is hard to differentiate from it. However, as the subcortical pathway is primarily monospectral in its response to tones and harmonics (Kostlan, 2015) and the cortex is highly recurrent a purely recurrent model is more parsimonious.

Experimental considerations

In a sieve, the neurons with BF values at integer multiples of the fundamental frequency are excitatory, while the half-integer-sensitive neurons are inhibitory. However, in the recurrent case they are *co-tuned*. Mistuned components will cause side-band inhibition that reduces the response to harmonic stimuli, rather than directly inhibiting the total response. For a harmonic template unit there may be “excitatory” and “inhibitory” units that feed into it and could be found electro-physiologically. If “inhibitory” units would respond preferentially respond to *half-integer* multiples of some bf_0 (or *odd-integer* multiples of ½ bf_0) this suggests a feed-forward mechanism of selectivity. However, if there is co-tuning it suggests a recurrent mechanism of selectivity. Co-tuning was found in the cortex, which increases spike-timing precision for detecting the onset response (Wehr & Zador, 2003).

Due to the tonotopy, it may be possible to trace a template neuron retrograde to see where it gets input from and use a second electrode to probe the best-frequencies of the region. Hysteresis, pattern completion, and oscillation are unique to recurrent networks and don’t occur in feed-forward networks. Hysteresis has a theoretical benefit for autocorrelated time series, which is discussed more in Chapter 5. Pattern completion is good for stimuli *identification*. Oscillation could represent seizure like state or physiological oscillations. Artificial manipulation of the amount of excitation and inhibition could potentially change the behavior by creating or removing hysteresis and/or oscillation.

CHAPTER 5:

Analysis of harmonically selective recurrent networks

5.1 Introduction

In Chapter 4 we showed that Hebbian training on harmonic stimuli could generate networks that contain harmonically selective neurons. Recurrent networks have a complex theory of which some will be analyzed and discussed in this chapter.

The richness of recurrent networks

Recurrence has several features not found in feed-forward systems. Hopfield networks can have multiple memories stored (Hopfield, 1982, 1984), which means that there will be hysteresis as the network jumps from one memory to another. Hysteresis is found commonly in auditory and other sensory psychophysical studies. Shepard tones are a circular sequence of ever “rising” exponentially-spaced harmonic sounds. Pairs of Shepard tones have a history-dependent difference-in-pitch perception (Chambers & Pressnitzer, 2014).

Another property is pattern completion, i.e. content addressable memory (Hopfield & Tank, 1985). Psychophysical correlates of this phenomenon have been observed (Cox et al., 2000). Recurrence with synaptic depression also is used as a model for stimulus specific adaptation, which is useful for directing attention to novel stimuli (Yarden & Nelken, 2017).

A much more difficult problem is the case of detecting a harmonic sound with unknown fundamental. This is similar to asking “how much of a pitch does said sound have, on a scale of 1-10” from a psychophysics perspective. Winner-take-all recurrent networks may help with this where spectral cues shift the network's guess as to the f_0 value as it hears stimuli. Such is the behavior of head-direction cells in the rat hippocampus: the network attracts to a single peak of activity pointing in some direction, which can be relocated by external stimuli (Knierim & Zhang, 2012).

As mentioned earlier, recurrent networks with symmetric weight matrixes can be understood mathematically as a neural network that is minimizing a Lyapunov function, which can be calculated analytically from the weights and thresholds (Hopfield & Tank, 1985). Said function can be interpreted statistically: is also the negative log-likelihood and the network seeks to maximize the log-likelihood. We discuss various gain functions and the corresponding Lyapunov functions, as well as the meaning of minimizing said function from an information theoretic perspective.

5.2 Analytical two-unit phase-plane analysis

In Chapter 4 we saw how the balance of excitatory and inhibitory self and cross feedback affects the dynamics. We can make a simpler model in which the excitatory neurons are lumped together as are the inhibitory neurons. This allows us to perform phase plane analysis. For piecewise-linear models most aspects of this analysis are analytic.

ReLU phase-plane analysis

The open-ended ReLU gain function is commonly used in machine learning and is similar to neurons in the cortex. This is the simplest nonlinear model in that it is piecewise linear with only two pieces per neuron. There are 10 parameters for the ReLU network are shown in table 5.1.

Parameter	Symbol	Restrictions
Capacitance	C	C > 0
Leakage	L	L>0
Threshold of excitation	T_E	None
Threshold of inhibition	T_I	None
Input to excitation	E	None
Input to inhibition	I	None
Excitatory self-feedback	W_{EE}	$W_{EE} \geq 0$
Inhibitory self-feedback	W_{II}	$W_{II} \leq 0$
Excitatory to inhibitory feedback	W_{EI}	$W_{EI} \geq 0$
Inhibitory to excitatory feedback	W_{IE}	$W_{IE} \leq 0$

Table 5.1: Parameters for the ReLU phase-plane analysis.

In addition to the parameters, there are three state variables: the voltages V_E and V_I and time *t*. This is the differential equation, which assumes equal L:

$$C\frac{dV_E}{dt} = E + W_{EE}f(V_E - T_E) + W_{IE}f(V_I - T_I) - LV_E$$

$$C\frac{dV_I}{dt} = I + W_{II}f(V_I - T_I) + W_{EI}f(V_E - T_E) - LV_I$$

For the ReLU case the gain function is:

f(*x*) = *max*(*x*, 0)

We want to make a dimensionless version of this system of equation. Start with the dimensionless voltage:

$$\nu_{i,e} \overset{\text{def}}{=} \frac{V_{i,e}}{T_{i,e}}$$

Note: $\frac{f(V_E - T_E)}{T_E} = \frac{f\left(T_E(\frac{V_E}{T_E} - 1)\right)}{T_E} = \frac{f(T_E(\nu_E - 1))}{T_E} = \frac{T_E sgn(T_E) f((\nu_E - 1) sgn(T_E))}{T_E} = sgn(T_E) f((\nu_E - 1) sgn(T_E))$

$$C\frac{d\nu_e}{dt} = \frac{E}{T_E} + W_{EE} sgn(T_E) f\big((\nu_e - 1) sgn(T_E)\big) + W_{IE} sgn(T_I)(T_I/T_E) f\big((\nu_i - 1) sgn(T_I)\big) - L\nu_E$$

$$C\frac{d\nu_i}{dt} = \frac{I}{T_I} + W_{II} sgn(T_I) f\big((\nu_i - 1) sgn(T_I)\big) + W_{EI} sgn(T_E) f\big((\nu_e - 1) sgn(T_E)\big) - L\nu_I$$

Now we introduce dimensionless time:

$$\tau \overset{\text{def}}{=} \frac{tL}{C}$$

$$\frac{d\nu_e}{d\tau} = \left(\frac{E}{LT_E}\right) + \left(\frac{W_{EE}}{L}\right) f\big((\nu_e - 1)sgn(T_E)\big)sgn(T_E) + \left(\frac{W_{IE}T_I}{LT_E}\right) f\big((\nu_i - 1)sgn(T_I)\big)sgn(T_I) - \nu_E$$

$$\frac{d\nu_i}{d\tau} = \left(\frac{I}{LT_I}\right) + \left(\frac{W_{II}}{L}\right) f\big((\nu_i - 1)sgn(T_I)\big)sgn(T_I) + \left(\frac{W_{EI}T_E}{LT_I}\right) f\big((\nu_e - 1)sgn(T_E)\big)sgn(T_E) - \nu_I$$

The equations now are dimensionless, and there are *six* independent parameters besides the choice of signs on the thresholds. Define:

$$\boldsymbol{\alpha_E \overset{\text{def}}{=} \frac{E}{LT_E}, \ \alpha_I \overset{\text{def}}{=} \frac{I}{LT_I}, \beta_E \overset{\text{def}}{=} \frac{W_{EE}}{L}, \ \beta_I \overset{\text{def}}{=} \frac{W_{II}}{L}, \gamma_E \overset{\text{def}}{=} \frac{W_{EI}T_E}{LT_I}, \ \gamma_I \overset{\text{def}}{=} \frac{W_{IE}T_I}{LT_E}}$$

The equations are now:

$$\frac{d\nu_e}{d\tau} = \alpha_E + \beta_E f\big((\nu_e - 1)sgn(T_E)\big)sgn(T_E) + \gamma_I f\big((\nu_i - 1)sgn(T_I)\big)sgn(T_I) - \nu_E$$

$$\frac{d\nu_i}{d\tau} = \alpha_I + \beta_I f\big((\nu_i - 1)sgn(T_I)\big)sgn(T_I) + \gamma_E f\big((\nu_e - 1)sgn(T_E)\big)sgn(T_E) - \nu_I$$

The pair of α parameters represent to what extent the stimulus can, on its own, bring the neuron up to threshold. The $\beta's$ represent how strong the self-feedback is in comparison to the leakage.

The $\gamma's$ are more complex. The interpretation of the $\gamma's$ is *how much our threshold can affect the other neuron*: It takes LT_I current to cause the inhibitory neuron to respond, ignoring feedback. Our threshold will reduce the excitatory output current by $W_{EI}T_E$ if there is any output current at all. This ratio determines how much the threshold of the excitatory neuron affects the inhibitory neuron's ability to turn on.

For any parameter set the previous formulas can compute these dimensionless numbers. For completeness here is a protocol for going the other way, i.e. generating a set of parameters and stimulus that satisfies a given set of dimensionless numbers. Numbers that are infinite (i.e. cases with zero threshold) can be handled with limits:

1. $L = 1, T_I = 1, T_E = 1,$
2. $W_{EE} = \beta_E, I_I = \beta_I, E = \alpha_E, \; I = \alpha_I$
3. $W_{EI} = \gamma_E, \; I_E = \gamma_I$

The differential equation is piecewise linear with 4 pieces depending on which neurons are above their threshold. Denote $\{x\}$ to be 1 if x is true and zero if false, and $\oplus$ is XOR:

$\chi_{\mathrm{E}} \overset{\mathrm{def}}{=} \left\{\{\nu_e > 1\} \oplus \{T_E < 0\}\right\}, \; \chi_{\mathrm{I}} \overset{\mathrm{def}}{=} \{\{\nu_i > 1\} \oplus \{T_I < 0\}\}$

This means, again with ReLU rectifiers: $f\big((\nu_e - 1)sgn(T_E)\big)sgn(T_E) = \; \chi_E(\nu_e - 1)$

Now pack it into a matrix form:

$$\begin{bmatrix} \dot{\nu_e} \\ \dot{\nu_\iota} \end{bmatrix} = \begin{bmatrix} \alpha_E - \beta_E \chi_{\mathrm{E}} - \gamma_I \chi_{\mathrm{I}} \\ \alpha_I - \beta_I \chi_{\mathrm{I}} - \gamma_E \chi_{\mathrm{E}} \end{bmatrix} + \begin{bmatrix} \beta_E \chi_{\mathrm{E}} - 1 & \gamma_I \chi_{\mathrm{I}} \\ \gamma_E \chi_{\mathrm{E}} & \beta_I \chi_{\mathrm{I}} - 1 \end{bmatrix} \begin{bmatrix} \nu_E \\ \nu_I \end{bmatrix}$$

The equilibrium is given by an equation with *implicit* conditionals but that is otherwise explicit:

$$\begin{bmatrix} \dot{\nu}_{eq,e} \\ \dot{\nu}_{eq,\iota} \end{bmatrix} = \frac{1}{(\beta_E \chi_E - 1)(\beta_I \chi_I - 1) - \gamma_E \chi_E \gamma_I \chi_I} \begin{bmatrix} \beta_I \chi_{\mathrm{I}} - 1 & -\gamma_I \chi_{\mathrm{I}} \\ -\gamma_E \chi_{\mathrm{E}} & \beta_E \chi_{\mathrm{E}} - 1 \end{bmatrix} \begin{bmatrix} \beta_E \chi_{\mathrm{E}} + \gamma_I \chi_{\mathrm{I}} - \alpha_E \\ \beta_I \chi_{\mathrm{I}} + \gamma_E \chi_{\mathrm{E}} - \alpha_I \end{bmatrix}$$

There are up to 4 equilibrium points, one for each possible χ truth value. For each truth value we can check existence and type. To check for existence we ensure that the predicted equilibrium location calculated for a given truth value corresponds to that value, in other words it must be self-consistent. If the equilibrium exists the type of equilibrium point will be determined by the trace and determinant:

$(\beta_E \chi_{\mathrm{E}} - 1)(\beta_I \chi_{\mathrm{I}} - 1) - \gamma_E \chi_{\mathrm{E}} - \gamma_I \chi_{\mathrm{I}} < 0 \rightarrow Saddle$

$(\beta_E \chi_{\mathrm{E}} - 1)(\beta_I \chi_{\mathrm{I}} - 1) - \gamma_E \chi_{\mathrm{E}} - \gamma_I \chi_{\mathrm{I}} > 0, \; \beta_E \chi_{\mathrm{E}} + \; \beta_I \chi_{\mathrm{I}} < 2 \rightarrow Sink$

$(\beta_E \chi_{\mathrm{E}} - 1)(\beta_I \chi_{\mathrm{I}} - 1) - \gamma_E \chi_{\mathrm{E}} - \gamma_I \chi_{\mathrm{I}} > 0, \; \beta_E \chi_{\mathrm{E}} + \; \beta_I \chi_{\mathrm{I}} > 2 \rightarrow Source$

This assumes that *CL*>0. For biologically realistic scenarios, C>0 always and L>0 except for unusual cases involving active voltage gated channels. If CL<0, we would reverse the sink/source criterion because time would flow "backwards". Thus we will use CL>0 for this analysis. The equilibrium points must also exist, truthiness of χ_E and χ_I must act for the values specified. Table 5.2 lists the conditions of equilibrium existence, stability and type. We use strict inequalities as there will always be some non-zero noise that would destroy marginal equilibria.

Case	Values	Existence	Saddle	Source (if not saddle)
$\chi_E = 0, \chi_I = 0$	$\nu_e = \alpha_E, \nu_i = \alpha_I$	$\alpha_e < 1, \alpha_i < 1$	False	False
$\chi_E = 1, \chi_I = 0$	$\nu_e = \frac{\alpha_E - \beta_I}{1 - \beta_I},$ $\nu_i = \gamma_E \frac{\alpha_E - 1}{1 - \beta_I}$	$\frac{\alpha_e - \beta_E}{1 - \beta_E} > 1$ $\gamma_E \frac{\alpha_E - 1}{1 - \beta_E} < 1$	$\beta_E > 1$	False
$\chi_E = 0, \chi_I = 1$	$\nu_e = \gamma_I \frac{\alpha_I - 1}{1 - \beta_I},$ $\nu_i = \frac{\alpha_I - \beta_I}{1 - \beta_I}$	$\gamma_I \frac{\alpha_I - 1}{1 - \beta_I} < 1$ $\frac{\alpha_I - \beta_I}{1 - \beta_I} > 1$	$\beta_I > 1$	False
$\chi_E = 1, \chi_I = 1$	$\frac{\begin{bmatrix} \beta_I - 1 & -\gamma_I \\ -\gamma_E & \beta_E - 1 \end{bmatrix} \begin{bmatrix} \beta_E + \gamma_I - \alpha_E \\ \beta_I + \gamma_E - \alpha_I \end{bmatrix}}{(\beta_E - 1)(\beta_I - 1) - \gamma_E \gamma_I}$	$\nu_{e,eq} > 1, \nu_{i,eq} > 1$	$(\beta_E - 1)(\beta_I - 1) < \gamma_E \gamma_I$	$\beta_E + \beta_I > 2$

Table 5.2: Exact values of ReLU equilibrium points.

We can solve the null clines. We need choose the independent variable in order to make the denominator only zero on a set of measure zero:

$$\begin{bmatrix} \dot{\nu}_e \\ \dot{\nu}_i \end{bmatrix} = \begin{bmatrix} \alpha_E - \beta_E \chi_E - \gamma_I \chi_I \\ \alpha_I - \beta_I \chi_I - \gamma_E \chi_E \end{bmatrix} + \begin{bmatrix} \beta_E \chi_E - 1 & \gamma_I \chi_I \\ \gamma_E \chi_E & \beta_I \chi_I - 1 \end{bmatrix} \begin{bmatrix} \nu_E \\ \nu_I \end{bmatrix}$$

$$\dot{\nu}_e = \alpha_E - \beta_E \chi_E - \gamma_I \chi_I + \nu_e(\beta_E \chi_E - 1) + \nu_i(\gamma_I \chi_I) = 0$$

$$\nu_e = \frac{\alpha_E - \beta_E \chi_E - \gamma_I \chi_I + \nu_i \gamma_I \chi_I}{1 - \beta_E \chi_E}, for\ e\ nullcline$$

We must try each possibility of χ_E

Similarly,

$$v_i = \frac{\alpha_I - \beta_I \chi_{\mathrm{I}} - \gamma_E \chi_{\mathrm{E}} + \nu_e \gamma_E \chi_E}{1 - \beta_I \chi_I}, for\ i\ nullcline$$

We want to know if, for a given set of parameters, the trajectory stays bounded for any initial conditions. There are three potential runaway cases, each sufficient to destabilize the system globally:

“E-runaway”: $\beta_E > 1, \gamma_E \leq 0$. The “E” activity explodes to infinity without “I” getting involved.

“I-runaway”: $\beta_I > 1, \gamma_I \leq 0$. Same concept as E-runaway but with the “I” running away. This requires ignoring the sign restrictions on the feedback weights.

“feedback-runaway”: The largest eigenvalue is real and positive for the $\chi_{\mathrm{E}} = \chi_{\mathrm{I}} = 1$ matrix and corresponds to an eigen vector with each component having the same sign (imaginary eigenvalues would cause the trajectory to spiral and leave the “feedback region”, as would real eigenvectors pointing the “wrong way”). This also requires ignoring the sign restrictions on the feedback weights.

The matrix in the feedback region, where both neurons are activated, is:

$$\begin{bmatrix} \beta_E - 1 & \gamma_I \\ \gamma_E & \beta_I - 1 \end{bmatrix}$$

We have:

$$trace = \beta_E + \beta_I - 2$$

$$det = (\beta_E - 1)(\beta_I - 1) - \gamma_E \gamma_I$$

$$discr = \frac{trace^2}{4} - det$$

$$\lambda_+ = \frac{trace}{2} + \sqrt{discr}$$

Runaway: $discr > 0 \quad and \quad \lambda_+ > 0 \quad and \quad (\lambda_+ - \beta_I + 1)\gamma_E > 0$

If *any* of these three conditions apply, we don't have global stability. If none do we have global stability. Global stability without stable equilibrium points means we have a *limit cycle*. This combination is possible with no sinks or saddles and a single source.

In fact, (for positive signs on the threshold) there are 13 combinations of sinks sources saddles and global stability conditions (less conditions are available if we consider the sign restrictions in the weights). Some possible sign-respecting phase-plane behaviors are shown in figure 5.1.

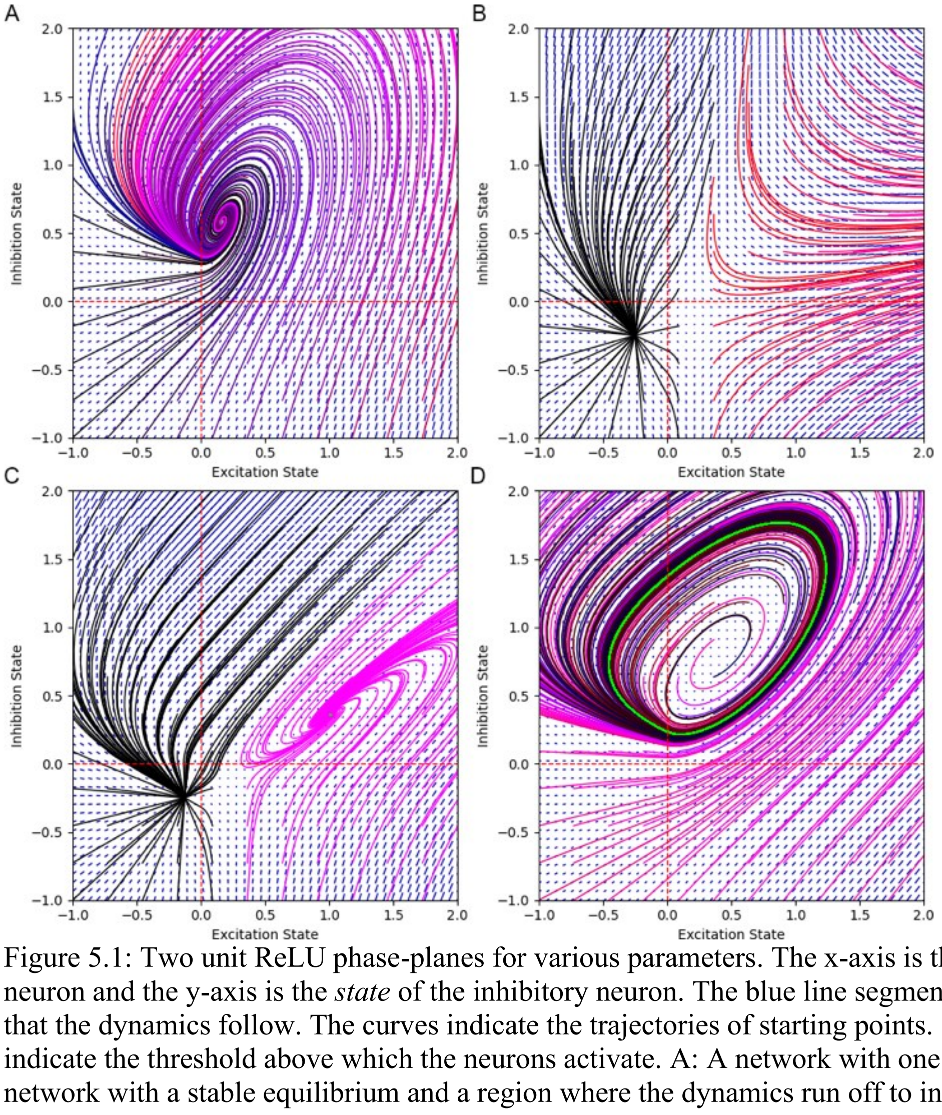


Figure 5.1: Two unit ReLU phase-planes for various parameters. The x-axis is the *state* of the excitatory neuron and the y-axis is the *state* of the inhibitory neuron. The blue line segments indicate the vector field that the dynamics follow. The curves indicate the trajectories of starting points. The red dotted lines indicate the threshold above which the neurons activate. A: A network with one stable equilibrium. B: A network with a stable equilibrium and a region where the dynamics run off to infinity. There is a saddle equilibrium on the boundary between the two basins of attraction. C: A network with two stable equilibria and a saddle between them. D: A network with no stable equilibrium but it has a limit cycle and no region that runs off to infinity. Limit cycles in these systems always surround an unstable spiral source. The instability in the linear dynamics of the "center" of the limit cycle gets stabilized by the nonlinearities in the gain function, in this case the deactivation of excitatory neural activity.

Zrect phase-plane analysis

ReLU was the simplest piecewise nonlinear model to analyze analytically and is realistic for individual cortical neurons which rarely fire strongly enough to hit refractory periods. But a real cortical neural network displays inhibition that prevents the neurons from getting near saturation. We assume that this is approximated by a sigmoid gain, which in turn can be approximated by a piecewise linear function, in the shape of a backwards Z. We call this the “Zrect” function. This has sigmoid-like properties of thresholding and saturation while still being piecewise linear. There are 12 parameters for the ReLU network are shown in table 5.3, note that we allow different amounts of dynamic range.

Parameter	Symbol	Restrictions
Capacitance	C	C > 0
Leakage	L	L>0
Threshold of excitation	T_E	None
Threshold of inhibition	T_I	None
Input to excitation	E	None
Input to inhibition	I	None
Excitatory self-feedback	W_{EE}	$W_{EE} \geq 0$
Inhibitory self-feedback	W_{II}	$W_{II} \leq 0$
Excitatory to inhibitory feedback	W_{EI}	$W_{EI} \geq 0$
Inhibitory to excitatory feedback	W_{IE}	$W_{IE} \leq 0$
Saturation level of excitation	S_E	$S_E \geq 0$
Saturation level of inhibition	S_I	$S_I \geq 0$

Table 5.3: Parameters for the Zrect phase-plane analysis.

The Z case is the same dynamics as the previously analyzed ReLU case except for the gain function:

$f(x, \mathrm{s}) = min(max(x, 0), \mathrm{s})$, where s is S_E or S_I.

As in the ReLU case, we want to extract dimensionless parameters from this equation.

First we make time dimensionless:

$$\tau \stackrel{\text{def}}{=} \frac{tL}{C}$$

$$\frac{dV_E}{d\tau} = \frac{E}{L} + \frac{W_{EE}}{L} f(V_E - T_E,\ \ S_E) + \frac{W_{IE}}{L} f(V_I - T_I, S_I) - V_E$$

$$\frac{dV_I}{d\tau} = \frac{I}{L} + \frac{W_{II}}{L} f(V_I - T_I, S_I) + \frac{W_{EI}}{L} f(V_E - T_E, S_E) - V_I$$

Then shift the thresholds to zero:

$V_{e0} \overset{\text{def}}{=} V_E - T_E, V_{i0} \overset{\text{def}}{=} V_I - T_I,\ E_0/L \overset{\text{def}}{=} E/L - T_E,\ I_0/L \overset{\text{def}}{=} I/L - T_I$

$$\frac{dV_{e0}}{d\tau} = \frac{E}{L} + \frac{W_{EE}}{L} f((V_{e0} + T_E) - T_E,\ \ S_E) + \frac{W_{IE}}{L} f((V_{i0} + T_I) - T_I,\ S_I) - (V_{e0} + T_E)$$

$$\frac{dV_{e0}}{d\tau} = \frac{E_0}{L} + \frac{W_{EE}}{L} f(V_{e0},\ \ S_E) + \frac{W_{IE}}{L} f(V_{i0},\ S_I) - V_{e0}$$

$$\frac{dV_{i0}}{d\tau} = \frac{I_0}{L} + \frac{W_{II}}{L} f(V_{i0},\ \ S_I) + \frac{W_{EI}}{L} f(V_{e0},\ S_E) - V_{i0}$$

Make the rectifiers dimensionless:

$$\frac{dV_{e0}}{d\tau} = \frac{E_0}{L} + \frac{W_{EE}S_E}{L} f(V_{e0}/S_E,\ \ 1) + \frac{W_{IE}\ S_I}{L} f(V_{i0}/\ S_I, 1) - V_{e0}$$

$$\frac{dV_{i0}}{d\tau} = \frac{I_0}{L} + \frac{W_{II}S_I}{L} f(V_{i0}/\ S_{tI},\ \ 1) + \frac{W_{EI}S_E}{L} f(V_{e0}/S_E, 1) - V_{i0}$$

We can now divide by the saturation levels to make the equations dimensionless:

$$\frac{dV_{e0}}{S_{tE}d\tau} = \frac{E_0}{LS_{tE}} + \frac{E_E}{L} f(V_{e0}/S_{tE},\ \ 1) + \frac{I_E\ S_{tI}}{S_{tE}L} f(V_{i0}/\ S_{tI}, 1) - \frac{V_{e0}}{S_{tE}}$$

$$\frac{dV_{i0}}{S_{tI}d\tau} = \frac{I_0}{S_{tI}L} + \frac{I_I}{S_{tI}L} f(V_{i0}/\ S_{tI},\ \ 1) + \frac{E_I S_{tE}}{S_{tI}L} f(V_{e0}/S_{tE}, 1) - \frac{V_{i0}}{S_{tI}}$$

Define: $\boldsymbol{\alpha_e} \overset{\text{def}}{=} \frac{\boldsymbol{E_0}}{\boldsymbol{LS_{tE}}}$, $\boldsymbol{\alpha_i} \overset{\text{def}}{=} \frac{\boldsymbol{I_0}}{\boldsymbol{LS_{tI}}}$, $\boldsymbol{\beta_e} \overset{\text{def}}{=} \frac{\boldsymbol{E_E}}{\boldsymbol{L}}$, $\boldsymbol{\beta_i} \overset{\text{def}}{=} \frac{\boldsymbol{I_I}}{\boldsymbol{L}}$, $\boldsymbol{\gamma_e} \overset{\text{def}}{=} \frac{\boldsymbol{I_E\ S_{tI}}}{\boldsymbol{S_{tE}L}}$, $\boldsymbol{\gamma_i} \overset{\text{def}}{=} \frac{\boldsymbol{E_I\ S_{tE}}}{\boldsymbol{S_{tI}L}}$, $\boldsymbol{\nu_e} \overset{\text{def}}{=} \frac{\boldsymbol{V_{e0}}}{\boldsymbol{S_{tE}}}$, $\boldsymbol{\nu_i} \overset{\text{def}}{=} \frac{\boldsymbol{V_{i0}}}{\boldsymbol{S_{tI}}}$,

$$\frac{d\nu_e}{d\tau} = \alpha_e + \beta_e g(\nu_e) + \gamma_e g(\nu_i) - \nu_e$$

$$\frac{d\nu_i}{d\tau} = \alpha_i + \beta_i f g(\nu_i) + \gamma_i g(\nu_e) - \nu_i$$

$$g_z(x) \overset{\text{def}}{=} \min\,(\max(x, 0)\,, 1)$$

For any parameter set it is straightforward to compute these dimensionless numbers, by first computing intermediate parameters. For completeness here is a protocol for going the other way, i.e. generating a set of parameters and stimulus that satisfies *any* given set of dimensionless numbers:

4. $L = 1,\ C = 1, T_I = 0, T_E = 0, S_{tE} = 1, S_{tI} = 1$
5. $E_E = \beta_E, I_I = \beta_I\ , E = \alpha_E,\ I = \alpha_I$
6. $E_I = \gamma_E,\ I_E = \gamma_I$

Our Zrect gain *g* is piecewise linear, so we can write it in terms of a matrix with bracketed Heaviside function if-statements. There are 9 pieces depending on which neurons are above their threshold. Denote $\{x\}$ to be 1 if *x* is true and zero if false:

$\chi_{\mathrm{E}} \overset{\mathrm{def}}{=} \{0 \le \nu_e < 1\},\ \chi_{\mathrm{I}} \overset{\mathrm{def}}{=} \{0 \le \nu_i < 1\}, \kappa_E \overset{\mathrm{def}}{=} \{\nu_e \ge 1\}, \kappa_I \overset{\mathrm{def}}{=} \{\nu_i \ge 1\}$

$$\frac{d\nu_e}{d\tau} = \alpha_e + \beta_e\nu_e\chi_E + \beta_e\kappa_E + \gamma_e\nu_i\chi_I + \gamma_e\kappa_I - \nu_e$$

$$\frac{d\nu_i}{d\tau} = \alpha_i + \beta_i\nu_i\chi_I + \beta_i\kappa_I + \gamma_i\nu_e\chi_E + \gamma_e\kappa_I - \nu_i$$

Now pack it into a matrix form:

$$\begin{bmatrix}\dot{\nu_e}\\ \dot{\nu_i}\end{bmatrix} = \begin{bmatrix}\alpha_e + \beta_e\kappa_e + \gamma_e\kappa_I\\ \alpha_i + \beta_i\kappa_i + \gamma_i\kappa_E\end{bmatrix} + \begin{bmatrix}\beta_e\chi_{\mathrm{E}} - 1 & \gamma_e\chi_{\mathrm{I}}\\ \gamma_i\chi_{\mathrm{E}} & \beta_i\chi_{\mathrm{I}} - 1\end{bmatrix}\begin{bmatrix}\nu_e\\ \nu_i\end{bmatrix}$$

The equilibrium is given by an equation with *implicit* conditionals but that is otherwise explicit:

$$\begin{bmatrix}\beta_e\chi_{\mathrm{E}} - 1 & \gamma_e\chi_{\mathrm{I}}\\ \gamma_i\chi_{\mathrm{E}} & \beta_i\chi_{\mathrm{I}} - 1\end{bmatrix}\begin{bmatrix}\nu_{eq,e}\\ \nu_{eq,i}\end{bmatrix} = -\begin{bmatrix}\alpha_e + \beta_e\kappa_e + \gamma_e\kappa_I\\ \alpha_i + \beta_i\kappa_i + \gamma_i\kappa_E\end{bmatrix}$$

$$\begin{bmatrix}\nu_{eq,e}\\ \nu_{eq,i}\end{bmatrix} = \frac{1}{(\beta_e\chi_E - 1)(\beta_i\chi_I - 1) - \gamma_e\chi_E\gamma_e\chi_I}\begin{bmatrix}1 - \beta_i\chi_{\mathrm{I}} & \gamma_e\chi_{\mathrm{I}}\\ \gamma_i\chi_{\mathrm{E}} & 1 - \beta_e\chi_{\mathrm{E}}\end{bmatrix}\begin{bmatrix}\alpha_e + \beta_e\kappa_e + \gamma_e\kappa_I\\ \alpha_i + \beta_i\kappa_i + \gamma_i\kappa_E\end{bmatrix}$$

There are up to 9 equilibrium points, one for each possible χ and κ truth (only one of these can be true at a time), however getting all 9 requires violating the sign restrictions on excitation or inhibition. For each truth value we can check existence and type. To check for existence we ensure that the predicted equilibrium location calculated for a given truth value corresponds to that value, in other words it must be self-consistent.

If the equilibrium exists the type of equilibrium point will be determined by the trace and determinant:

$(\beta_E \chi_{\mathrm{E}} - 1)(\beta_I \chi_{\mathrm{I}} - 1) - \gamma_E \chi_{\mathrm{E}} - \gamma_I \chi_{\mathrm{I}} < 0 \rightarrow Saddle$

$(\beta_E \chi_{\mathrm{E}} - 1)(\beta_I \chi_{\mathrm{I}} - 1) - \gamma_E \chi_{\mathrm{E}} - \gamma_I \chi_{\mathrm{I}} > 0,\ \beta_E \chi_{\mathrm{E}} + \beta_I \chi_{\mathrm{I}} < 2 \rightarrow Sink$

$(\beta_E \chi_{\mathrm{E}} - 1)(\beta_I \chi_{\mathrm{I}} - 1) - \gamma_E \chi_{\mathrm{E}} - \gamma_I \chi_{\mathrm{I}} > 0,\ \beta_E \chi_{\mathrm{E}} + \beta_I \chi_{\mathrm{I}} > 2 \rightarrow Source$

This assumes that $L/C>0$. For biologically realistic scenarios, $C>0$ always and $L>0$ except for unusual cases involving active voltage gated channels. If $L/C<0$, we would reverse the sink/source criterion because time would flow "backwards". We use $L/C>0$ for this analysis.

For the equilibrium points to exist, the truthiness of χ_{E} and χ_{I} at an equilibrium must be self-consistent with the truthiness that has been specified to calculate said equilibrium.

We can make a table of stability and type. We use strict inequalities as there will always be some non-zero noise that would destroy marginal equilibria.

Case	Values	Existence	Saddle	Source (if not saddle)
All zero	$\nu_e = \alpha_e, \nu_i = \alpha_I$	$\alpha_e < 0, \alpha_i < 0$	False	False
$\chi_E = 1, \chi_I = \kappa_I = 0$	$\nu_e = \frac{\alpha_e}{\beta_e - 1}$, $\nu_i = \alpha_i + \gamma_i \frac{\alpha_e}{\beta_e - 1}$	$0 < \frac{\alpha_e}{\beta_e - 1} < 1$ $\alpha_i + \gamma_i \frac{\alpha_e}{\beta_e - 1} < 1$	$\beta_e > 1$	False
$\chi_E = \kappa_E = 0, \chi_I = 1$	$\nu_e = \alpha_e + \gamma_e \frac{\alpha_i}{\beta_i - 1}$, $\nu_i = \frac{\alpha_i}{\beta_i - 1}$	$\alpha_e + \gamma_e \frac{\alpha_i}{\beta_i - 1} < 1$ $0 < \frac{\alpha_i}{\beta_i - 1} < 1$	$\beta_i > 1$	False
$\chi_E = 1, \chi_I = 1$	$\frac{\begin{bmatrix}\alpha_e(1-\beta_i) + \alpha_i\gamma_e \\ \alpha_i(1-\beta_e) + \alpha_e\gamma_i\end{bmatrix}}{(\beta_e - 1)(\beta_i - 1) - \gamma_e\gamma_i}$	$0 < \nu_{e,eq} < 1$, $0 < \nu_{i,eq} < 1$,	$(\beta_e - 1)(\beta_i - 1) < \gamma_e\gamma_i$	$\beta_E + \beta_I > 2$
$\kappa_E = 1, \chi_I = \kappa_I = 0$	$\nu_e = \alpha_e + \beta_e$ $\nu_i = \alpha_i + \gamma_i$	$\alpha_e + \beta_e > 1$ $\alpha_i + \gamma_i < 1$	False	False
$\kappa_E = 1, \chi_I = 1$	$\nu_e = \alpha_e + \beta_e + \frac{\gamma_e\alpha_i + \gamma_e\gamma_i}{1 - \beta_i}$ $\nu_i = \frac{\alpha_e + \gamma_i}{1 - \beta_i}$	$\nu_{e,eq} > 1$, $0 < \frac{\alpha_e + \gamma_i}{1 - \beta_i} < 1$	$\beta_i > 1$	False
$\chi_E = \kappa_E = 0, \kappa_I = 1$	$\nu_e = \alpha_e + \gamma_e$ $\nu_i = \alpha_i + \beta_i$	$\alpha_e + \gamma_e < 1$ $\alpha_i + \beta_i > 1$	False	False
$\chi_E = 1, \kappa_I = 1$	$\nu_i = \frac{\alpha_i + \gamma_e}{1 - \beta_e}$ $\nu_i = \alpha_i + \beta_i + \frac{\gamma_i\alpha_e + \gamma_e\gamma_i}{1 - \beta_i}$	$0 < \frac{\alpha_i + \gamma_e}{1 - \beta_e} < 1$ $\nu_{i,eq} > 1$	$\beta_e > 1$	False
$\kappa_E = 1, \kappa_I = 1$	$\nu_e = \alpha_E + \beta_e + \gamma_e$ $\nu_i = \alpha_i + \beta_i + \gamma_i$	$\alpha_E + \beta_e + \gamma_e > 1$ $\alpha_i + \beta_i + \gamma_i > 1$	False	False

Table 5.4: Exact values of "Z" equilibrium points. Note that the "Z" gain function is a three-linear-piece approximation of the sigmoid gain function.

We can solve for the null clines in a similar way to the ReLU case. We need choose the independent variable in order to make the denominator only zero on a set of measure zero:

$$\dot{v_e} = \alpha_e + \beta_e v_e \chi_E + \beta_e \kappa_E + \gamma_e v_i \chi_I + \gamma_e \kappa_I - v_e = 0$$

$$v_e = \frac{\alpha_e + \beta_e \kappa_E + \gamma_e v_i \chi_I + \gamma_e \kappa_I}{1 - \beta_e \chi_E}, for\ E\ nullcline$$

For each value of v_i, we must try each possibility of χ_E and κ_E and discard any that don't work. There may be multiple values of $v_e$for a given v_i. Similarly,

$$v_i = \frac{\alpha_i + \beta_i \kappa_I + \gamma_i v_e \chi_E + \gamma_i \kappa_E}{1 - \beta_i \chi_I}, for\ I\ nullcline$$

We trivially have a global attractor because in the limit of large positive or large negative responses the leak term is proportional to the voltages, and brings the system toward the origin. Everything else saturates to a finite number. There are still cases with no stable equilibria, they all have a limit cycle. Some Zrect phase-plane behaviors are shown in figure 5.2

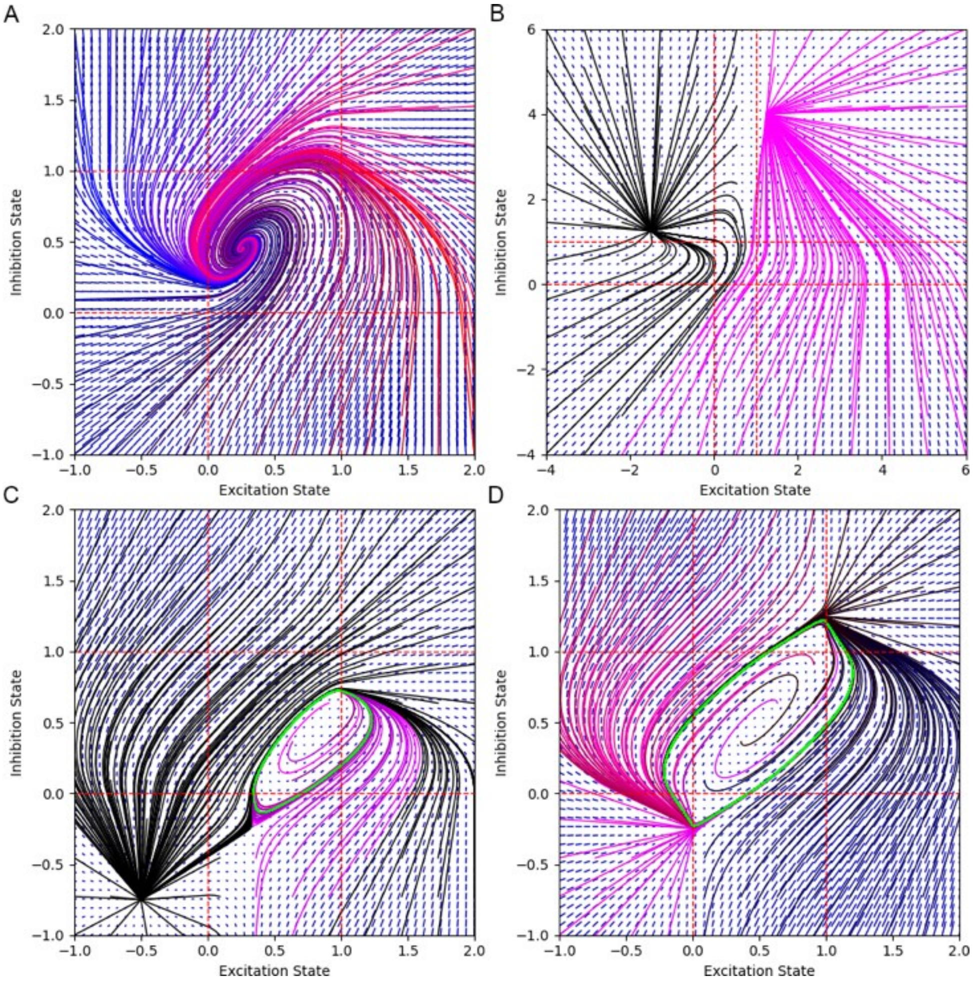


Figure 5.2: Two unit Zrect phase-planes for various parameters. These are presented in the same format as figure 5.1 except that there are two pairs of red dotted lines indicating the activation and the saturation. A network with a single stable equilibrium. B: A network with two stable equilibria, with a saddle between them. C: A network with a stable equilibrium, saddle, and limit cycle. D: A network with a limit cycle but no stable equilibria.

5.3 Converting neural networks into a two-unit networks

In the previous Section of this chapter we have shown that two-unit neural networks can be analytically approximated in terms of phase-plane features. In this section we show how to approximate a more complex neural network as a two-unit network.

We assume that the time constants of both models are 1, thus there is no need to scale time. We seek to lump all excitatory and inhibitory neurons Together.

$$v_{\overline{exc0}} \overset{\text{def}}{=} \frac{1}{k}\sum even(i)\, v_i\,, \qquad v_{\overline{inh0}} \overset{\text{def}}{=} \frac{1}{k+1}\sum odd(i)\, v_i$$

Define the average total excitatory and inhibitory input:

$$\alpha_{\overline{exc}} \overset{\text{def}}{=} \frac{1}{k}\sum even(i)\, \alpha_i\,, \qquad \alpha_{\overline{inh}} \overset{\text{def}}{=} \frac{1}{k+1}\sum odd(i)\, \alpha_i$$

Define the average gain inputs:

$$g_{\overline{EE}} \overset{\text{def}}{=} \frac{1}{k}\sum even(i) g(v_i),\ g_{\overline{II}} \overset{\text{def}}{=} \frac{1}{k+1}\sum odd(i) g(v_i),$$

$$g_{\overline{EI}} \overset{\text{def}}{=} \frac{2k}{k+1} g_{\overline{EE}},\ \ g_{\overline{IE}} \overset{\text{def}}{=} \frac{1}{k}\left(2\left(\sum odd(i) g(v_i)\right) - g(v_1) - g(v_N)\right)$$

Now we can rewrite the equations in terms of averages:

$$\frac{dv_{\overline{exc0}}}{d\tau} = \alpha_{\overline{exc}} + W_{EE} g_{\overline{EE}} + W_{IE} g_{\overline{IE}} - v_{\overline{exc0}}$$

$$\frac{dv_{\overline{inh0}}}{d\tau} = \alpha_{\overline{inh}} + W_{EI} g_{\overline{EI}} + W_{II} g_{\overline{II}} - v_{\overline{inh0}}$$

The first approximation is to reverse the order of averaging and applying the gain function:

$$g_{\overline{EE}} \approx g(v_{\overline{exc0}}), \qquad g_{\overline{II}} \approx g(v_{\overline{inh0}}), \qquad g_{\overline{EI}} \approx \frac{2k}{k+1} g(v_{\overline{exc0}}), \qquad g_{\overline{IE}} \approx \frac{2k-2}{k} g(v_{\overline{inh0}})$$

These approximations are exact if all excitatory are at the same voltage and so are all inhibitory neurons, and they are accurate to $O(\epsilon^2)$ for small ϵ deviations away from uniformity. However, for large deviations they can be considerably inaccurate. *Since this is the main source of error,* *deviations from the Zrect model that do not rely on fine-tuning are likely do to non-uniform activation*.

We approximate the sigmoid gain function with a Z function of the same maximum slope, minimum and maximum value, and argument of midpoint. A minimum least-square difference approximation would have a slightly higher maximum slope, but the maximum slope is very important for setting the bifurcation dynamics, and we are still *close* to a least squares fit. This gives us: $g(x) \approx g_z(\frac{x}{4} + \frac{1}{2})$

Putting this all together we approximate the harmonic equations in terms of the nonharmonic equations:

$$\frac{dv_{\overline{exc}}}{d\tau} \approx \alpha_{\overline{exc}} + W_{EE} g_z\left(\frac{v_{\overline{exc}}}{4} + \frac{1}{2}\right) + W_{IE} \frac{2k-2}{k} g_z\left(\frac{v_{\overline{\imath nh}}}{4} + \frac{1}{2}\right) - v_{\overline{exc0}}$$

$$\frac{dv_{\overline{\imath nh}}}{d\tau} \approx \alpha_{\overline{\imath nh}} + W_{EI} \frac{2k}{k+1} g_z\left(\frac{v_{\overline{exc}}}{4} + \frac{1}{2}\right) + W_{II} g_z\left(\frac{v_{\overline{\imath nh}}}{4} + \frac{1}{2}\right) - v_{\overline{\imath nh0}}$$

Define: $v_{\overline{exc}} \overset{\text{def}}{=} \frac{v_{\overline{exc0}}}{4} + \frac{1}{2}$, $\qquad v_{\overline{\imath nh}} \overset{\text{def}}{=} \frac{v_{\overline{\imath nh0}}}{4} + \frac{1}{2}$

Helpful: $v_{\overline{exc0}} = 4v_{\overline{exc}} - 2$

$$4\frac{dv_{\overline{exc}}}{d\tau} \approx \alpha_{\overline{exc}} + W_{EE} g_z(v_{\overline{exc}}) + W_{IE} \frac{2k-2}{k} g_z(v_{\overline{\imath nh}}) - 4v_{\overline{exc}} + 2$$

$$4\frac{dv_{\overline{\imath nh}}}{d\tau} \approx \alpha_{\overline{\imath nh}} + W_{EI} \frac{2k}{k+1} g_z(v_{\overline{exc}}) + W_{II} g_z(v_{\overline{\imath nh}}) - 4v_{\overline{\imath nh}} + 2$$

Substitute the alphas:

$$\frac{dv_{\overline{exc}}}{d\tau} \approx \left(\frac{2 + \frac{1}{k}\sum even(i)\, \alpha_i}{4}\right) + \left(\frac{W_{EE}}{4}\right) g_z(v_{\overline{exc}}) + \left(W_{IE} \frac{k-1}{2k}\right) g_z(v_{\overline{\imath nh}}) - v_{\overline{exc}}$$

$$\frac{dv_{\overline{\imath nh}}}{d\tau} \approx \left(\frac{2 + \frac{1}{k+1}\sum odd(i)\, \alpha_i}{4}\right) + \left(W_{EI} \frac{k}{2k+2}\right) g_z(v_{\overline{exc}}) + \left(\frac{W_{II}}{4}\right) g_z(v_{\overline{\imath nh}}) - v_{\overline{\imath nh}}$$

The final conversion formula:

$$\alpha_{exc} = \frac{1}{2} + \frac{\sum even(i)\,\alpha_i}{4k}, \alpha_{inh} = \frac{1}{2} + \frac{\sum odd(i)\,\alpha_i}{4(k+1)}, \beta_{exc} = \frac{W_{EE}}{4}, \beta_{inh} = \frac{I_I}{4},$$

$$\gamma_{exc} = W_{IE}\frac{k-1}{2k}, \gamma_{inh} = W_{EI}\frac{k}{2k+2}$$

$$\nu_{exc} = \frac{1}{2} + \frac{1}{4k}\sum even(i)\,\nu_i, \quad \nu_{inh} = \frac{1}{2} + \frac{1}{4k+4}\sum odd(i)\,\nu_i$$

Like the larger networks, two-unit networks are able to have hysteresis and/or oscillation, depending on the exact parameters.

Phase offsets

Suppose a 2x2 phase-plane matrix is *barely* a spiral source, so that it is *almost* linear and the nonlinear boundaries are pushing ever so gently to stabilize it. Thus we can do linear analysis.

Phase plane constraints: The trace of the matrix must be zero, with determinate > 0.

E-I constraints (matrix is to-from indexed, x axis is E, y axis is I):

$$\frac{dV}{dt} = \begin{bmatrix} W_{EE}-1 & W_{IE} \\ W_{EI} & W_{II}-1 \end{bmatrix}(V - V_{eq})$$

$$W_{EE} + W_{II} = 2$$

$$(W_{EE}-1)(W_{II}-1) - (W_{EI})(W_{IE}) > 0$$

$$W_{EE}, W_{EI}, W_{IE}, W_{II} \geq 0$$

This will give us an elliptical orbit.

Time lag: $t = argmax_{\eta}(((x - \langle x \rangle) - \eta) * (y - \langle y \rangle))$

For a circular orbit the phase difference is 90 degrees, but it can be less or more for an elliptical orbit, and in the almost linear regime can be analyzed with matrix exponential.

5.4 Lyapunov function analysis

Gains and Lyapunov functions

Consider a recurrent dynamical system with symmetric weight matrix W:

$$\tau \frac{dv}{dt} = W^T g(v-b) - v + j$$

Where j is the input current v is the voltage, and b is the threshold.

The activity of the network is $s = g(v-b)$. It is easier to compute the Lyapunov function in activity space is (Hopfield & Tank, 1985):

$$L(s) = \sum \int_0^{s_i} (g^{-1}(\mathrm{s_i}))ds + b^T s - j^T s - \frac{1}{2} s^T W s$$

$$s_i = g(v_i - b_i)$$

We derived L for four separate gain functions and summarize the results in table 5.5.

Gain	Formula	Lyapunov function
Linear (this is a "control" of sorts)	$g(x) = x$	$L(s) = (b-j)^T s - \frac{1}{2} s^T (W-I) s$
Sigmoid	$g(x) = \frac{1}{1+\exp(-kx)}$	$L(s) = \frac{1}{k}(s \ln s + (1-s)\ln(1-s)) + (b-j)^T s - \frac{1}{2} s^T W s$
Step	$g(x) = u(x)$	$L(s) = (b-j)^T s - \frac{1}{2} s^T W s$
ReLU	$g(x) = \int u(x)$	$L(s) = (b-j)^T s - \frac{1}{2} s^T (W-I) s$

Table 5.5: Four Lyapunov functions for four commonly used neural network gains.

This is quadratic for three of the 4 cases, but there are boundary conditions for the nonlinear gains. It is nearly quadratic for the sigmoid case, where it gets very steep very near the boundaries where s approaches zero or one, preventing the neuron from getting to the actual boundary. For the quadratic cases this changes the problem from one of nonlinear optimization to one of bounded quadratic optimization.

Lyapunov functions as objective functions

If there is no feedback from the excitatory to inhibitory units there is a Lyapunov function for the inhibitory subnetwork, and the excitatory subnetwork using the inhibitory network's stable state as input. This guarantees global stability if the gain is bounded as are our sigmoid gains used in Chapter 4. But even given two-way feedback we can still discuss the effects of the functions, it is just not *guaranteed* to have a Lyapunov function which gets minimized.

For a sigmoid gain in activity space (bounded between s=0 and s=1) and symmetrical weights we have:

$$L(s) = \frac{1}{k}(s\ln(s) + (1-s)\ln(1-s)) + (b-j)^T s - \frac{1}{2}s^T W s$$

Where b is the threshold, k is the gain (which is one without loss of generality), j is the current, and W is the weight matrix for the recurrent network.

There is a statistical interpretation to this function. The function is trying to find *directions* that are strongly distinguished from s=0.5 while maintaining entropy.

The first term is the binary entropy function and thus extreme values with low entropy are penalized. Indeed, the infinite slope of the function prevents the limits from ever being reached.

The term $\frac{-1}{2}s^T W s$ is also part of the prior. One can consider W as an inverse covariance matrix of a normal distribution, and this term is to maximize the contrast (or *un*likelihood), which is the opposite goal as to what most Bayesian priors do. For a sign-agonistic network with a given stimulus, the simplest Hebbian rule is $-W = -xx^T$. This is a *negative* semidefinite term with one negative eigenvalue. It is destabilizing, favoring $s \rightarrow \pm x$ and is minimized at infinity.

If the destabilization (most negative eigenvalue) exceeds the second derivative of the entropy term (which is a diagonal matrix) then there are two stable equilibrium so long as $b - j$ is small enough. Otherwise there is only one stable equilibria.

$(b - j)^T$ is the *posterior* term, that updates the "Bayesian inference" process. It states that the log likelihood increases in a given direction, no matter how far along said direction one travels. If this term aligns with x it will act in synergy with the quadratic term. Thus the overall likelihood optimization is to maximize entropy while finding the directions that make the most "sense".

5.5 Bf-agnostic harmonicity detection

Suppose we *aren't* restricted to a single bf0, but instead want to know if sounds are harmonic or not; i.e. a bf0-agnostic detector. Suppose the goal of a network is to detect a harmonic sound given by x = K*x0 where x0 is the input vector and the filter K is a fixed-width gaussian convolution. We want to solve this problem.

A highly nonlinear problem

Assume that we have harmonic stimuli with randomized component strength and randomized f0. Each of these stimuli corresponds to a subcortical input in $\mathbb{R}^N$ for *N* channels. Our network must respond to these subcortical inputs while ignoring background inputs. What is the *shape* of the set of harmonic inputs in this high dimensional space?

For simplicity, we assume that the harmonic stimuli is well-resolved (sigma(K) << f0). The makes the x a collection of equally-spaced equal-shape and height gaussian peaks that don't overlap significantly.

This stimulus space is a high dimensional *open generalized cone*. Given a cone in 3 dimensions with its vertex at the origin, points on it can be parameterized by theta and r. theta is a non-linear parameter in that it adjusting it causes the point to make a non-straight curve in the 3 dimensional space. The parameter r is the distance from the origin and acts as a linear multiplier: doubling r doubles the point's location. Furthermore, moving along r is perpendicular to moving along theta.

A generalized cone in three dimensions can have any arbitrary shaped base. In this case it can be parameterized by arclength s instead of theta, while r is unchanged. The arclength is taken *at unit r*. Our case is a higher-dimensional version of the generalized cone. All directions along A are orthogonal to the direction along f0 and to each-other. Cones and generalized cones don't have *intrinsic* curvature but they are curved in the higher dimensional space (i.e they have non-zero *principle curvature*). An *open* generalized cone has an unbounded *s* rather than a parameter that loops back on itself.

A cone or generalized cone in 3D space is the union of a 1-parameter locus of lines through the origin. Our stimuli space unionizes a 1-parameter collection of a hyperplanes through the origin. If we have *N* components across our harmonic complex, each hyper-plane is *N* dimensional and the overall surface is *N*+1 dimensional. For a given f0, each single convolved component corresponds to a line. These lines are orthogonal to each-other for well-resolved harmonics and define a hyperplane of possible stimuli for said f_0. This surface lives in an infinite dimensional Euclidian space that becomes finite-dimensional when we discretize. However, our locus of hypersurfaces is *still* a single-parameter locus, parameterized by f_0.

The curvature along the cone, the coordinate acceleration, as we change the arclength is *constant*. A harmonic complex is given by:

$$x = \sum_i A_i K * \delta(f_0 i)$$

Arclength can be parameterized by (with discretization the integral would be a sum):

$$ds = \sqrt{\int \left(\frac{dx}{ds}\right)^2 df_0}, \qquad A = [1,1, \dots 1]$$

Due to the translational symmetry, as long as the harmonic complex stays well-resolved, we have $s \sim f_0$. Thus f_0 is proportional to arclength parametrization *we will use it instead of arclength*.

Moving along a given component in A takes the stimulus closer and further from the origin. The velocity vector is given by: $\frac{dx}{dA_i} = K * \delta(f_0 i)$. There is no curvature/acceleration along any A direction.

Along the $f_0$direction the velocity is:

$\frac{dx}{d_{f0}} \sim \sum_i iA_i \, (\frac{dK}{dx} * \delta(f_0 i))$.

The appearance if *i* in the sum is because higher components move faster as we change f_0.

The curvature (acceleration vector) is given by:

$$a = \frac{d^2x}{df_0^2}$$

This acceleration is:

$$a = \sum_i iA_i \left(\frac{d^2K}{dx^2} * \delta(f_0 i)\right), \qquad |a| \sim \sum_i iA_i$$

Thus our curve (at a constant A) has a constant acceleration amount (but not a constant acceleration *direction*).

When we relax the assumption that of well-resolved harmonics, we still have a generalized hyper-cone because *x* is still linear in *A*. However, the velocity (relationship between *s* and *f0*) and acceleration magnitude are no longer is independent of *f0*. At moderate *f0*, the tails of the components start interacting with the head of the next component and velocity lowers. At very low f0's, the curves overlap so much velocity increases. Also, the single-component lines that generate the hyper-plane are no-longer orthogonal.

What kinds of discrimination problems require non-linear decision problems? Consider a discrimination problem where there are N=8 components. Non-harmonic sounds have the first 4 components shifted slightly downward, and the last 4 shifted upward. A non-harmonic sound can be parameterized by *f0*, A, as well as epsilon, with epsilon determining the shift:

$$x_{back} = \sum_{i=1\ to\ 8} A_i K * \delta\left(f_0 i - \frac{\epsilon}{i}[i \leq 4] + \frac{\epsilon}{i}[i > 4]\right)$$

For small ϵ, the negative shift of the first 4 components is like decreasing *f0* while the positive shift is like increasing *f0* . These cancel out and we have orthogonality: $\left(\frac{dx}{d_{f0}}\right) . \left(\frac{dx_{back}}{d\epsilon}\right) = 0$.

Thus the direction away from harmonicity is always perpendicular to velocity. A discrimination function that is proportional to *A* would have to *follow* this changing velocity.

Solving the agnostic detection of harmonics problem

In Chapter 3 our test with a variable range of *f0* values performed poorly on feed-forward networks; with the same parameters as that variable-f0 problem where the background stimuli are jittered versions of the foreground stimuli, except with the random height at 0.25 and the log-range of the fundamental of 0.8.

A natural way to solve this problem is a *max of sieves*. For each *f0* value we build a linear classifier (which was shown to be very similar to the single-*f0* solutions in Chapter 3). The classifier for a given *f0* is:

$$\frac{(x-\overline{x_{back}})^T(\overline{x_{f0}}-\overline{x_{back}})}{|\overline{x_{f0}}-\overline{x_{back}}|^2}+b_{unbias}>\frac{1}{2}$$

When the subcortical input *x* is equal to the mean of the background stimuli (for all *f0* values), $\overline{x_{back}}$, this classifier returns zero. Without (the scalar) b_{unbias}, when *x* is equal to the mean of the subset of the foreground stimuli with the fundamental is our given *f0,* $\overline{x_{f0}}$, the classifier returns one. The cutoff is half-way in-between these points.

We consider a sound harmonic iff any of the classifiers for any of the allowed *f0* frequencies returns true, i.e. we compute the maximum activation over our sieve-bank network. b_{unbias} is set so that the average of all background and foreground training stimuli is ½ (the background is below ½ by the same amount that the foreground is above ½). The maximum operation will tend to return values that are too high and b_{unbias} corrects for this effect.. The problem is easily converted into a recurrent neural network as shown in figure 5.3.

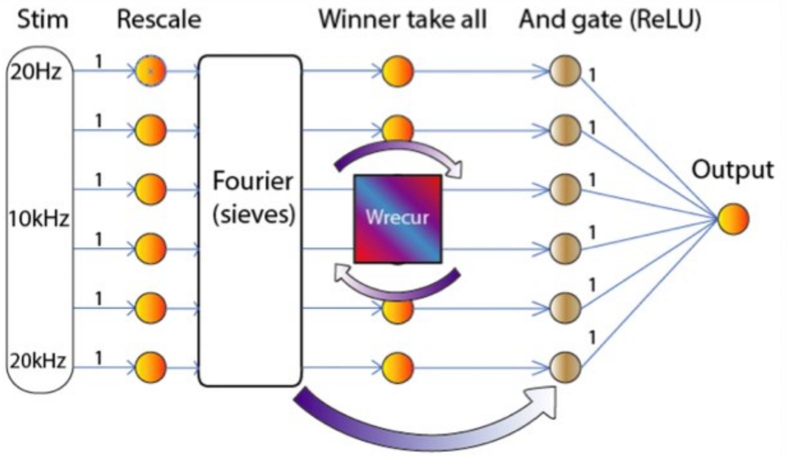


Figure 5.3: A recurrent network approximation of max-of-sieves. It has similar performance to a max-of-sieves algorithm.

The “Fourier” weight matrix shown in figure 5.4 stores these sieves and resembles a Fourier transform in that it approximately measures how strongly sinusoidal the network and has it’s thresholds set so that units downstream 50% activation at the thresholds in the max-of-sieves.

The weights feeding into the recurrent layer from the Fourier layer are diagonal and activate the corresponding recurrent units. The recurrent weight matrix with number of channels N_{ch} is excitatory with global inhibition:

$R_{ij},= \frac{|i-j|}{N_{ch}-1},\ R_{wrap} = R[R < 0.5] + (1 - R)[R \geq 0.5]$; R varies from -1 to 1 corner to corner and is zero down the diagonal, R_{wrap} has a periodic boundary condition that is not natural but makes the attractor peak location unbiased.

$Exc_{unnorm} = \exp\left(-\frac{0.5R^2}{\sigma^2}\right),\ \ Exc = Exc_{unnorm} diag\left(\sum_j Exc_{unnorm,ij}\ \right),$

$Inh = \frac{k_I}{N_{ch}},\quad W = Exc + Inh$

The parameters are: $k_E = 1, k_E = 1, \sigma = 0.25$

This layer behaves as a “winner take all” in that there is an attractor that inhibits the formation of other locations while reinforcing itself. The location of the attractor will be pulled toward the strongest region of inputs from the Fourier layer so it will act like an argmax function.

The and-gate layer is the only layer with non-sigmoid gain and it has an ReLU gain. It gets identity-matrix inputs from the sievebank and recurrent layer. Units only activate if they are in the vicinity of the active region in the recurrent network and have above-50% activation of the Fourier layer. Finally, the single output neuron is very easily activated.

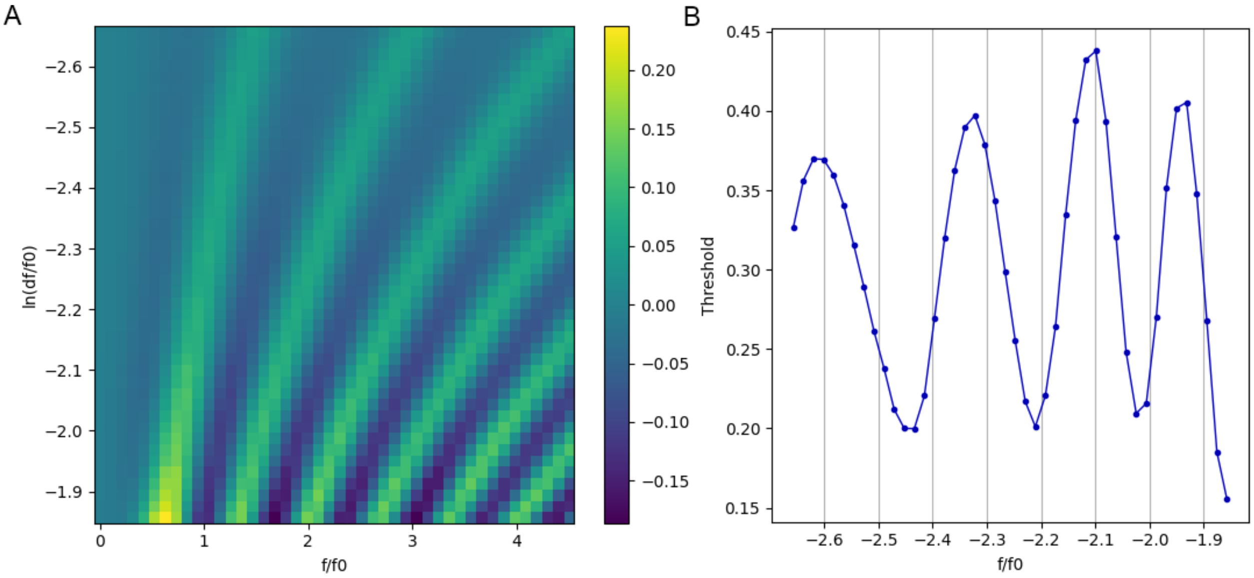


Figure 5.4: The sieve layer in the winner-take all network. A: The bank of sieves, each row is a sieve. B: The thresholds at each sieve.

It is possible to solve this particular problem without a recurrent network by making the ReLU neurons get direct input form the Fourier layer and have a threshold of ½. However, the use of a “winner take all” network offers several computational advantages.

Firstly, it acts like a noise-filter: a single “false alarm” neuron from the Fourier layer will not pull the active region toward its location as much as a collection of nearby neurons with moderate activity. A simple the max-of-sieves method, which can be implemented in a feedforward network, doesn’t have. The recurrent network also allows a graded determination of the maximum value rather than a simple binary “on-off” classifier. Finally, the recurrent network allows programmable selective attention to be added: an extra layer of neurons feeding diagonally into the recurrent layer could select where the attention is drawn. Thus recurrence has several advantages that would be difficult to implement with purely feed-forward networks. The results of a run with a recurrent network are shown in figure.

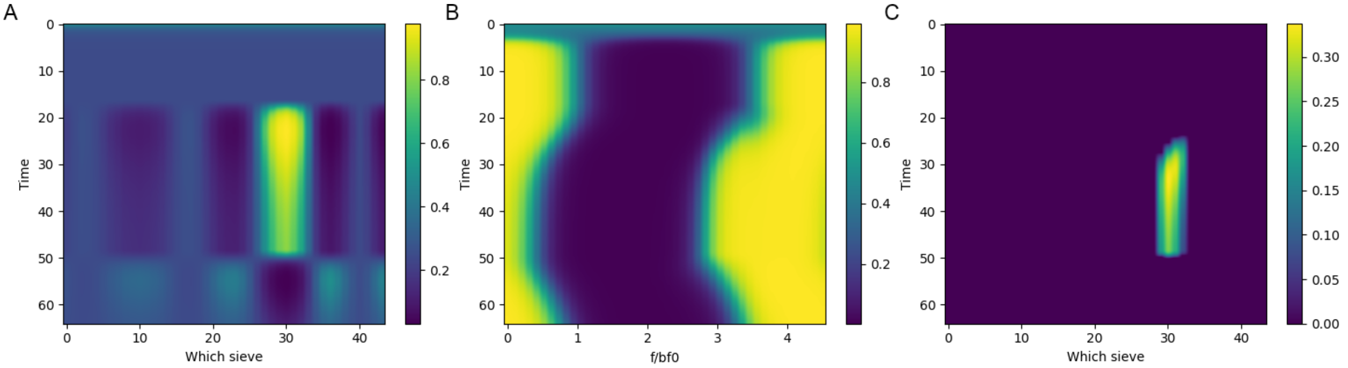


Figure 5.5: The recurrent network responses to a fixed harmonic stimulus. The stimulus turns on at time 18 and turns off at time 50. A: The bank of sieves. B: The winner-take-all network. The activation “pulls” the excitation toward it. C: The and-gate that requires activity from both A and B to activate.

Feedforward networks perform miserably without special training methods, as summarized in table 5.6. The network is set up in the same way as Chapter 3. However, a max-of-sieves performs well.

Network	Training %	Test %	Linearness
Zero hidden layers	56.89	54.69	0.9663
One hidden layer	51.29	50.58	0.9608
Two hidden layers	49.93	49.93	Highly nonlinear
Maximum of sieves	79.1	78.52	Highly nonlinear
Win take all recurrent	81.49	81.64	Highly nonlinear

Table 5.6: The training and test performance data of the various networks. The maximum of sieves well outperforms the feed-forward networks and is comparable to the winner-take-all network. 1024 stimuli in each category were used with the corresponding binomial error margins of about ± 3%.

5.6 Optimal hysteresis parameters for autocorrelated time-series

Perceptual hysteresis is found throughout multiple sensory modalities. For the auditory system tone perception vs absence shows a range of a few dB around threshold (Pickles, 2013).

Hysteresis is useful when past stimuli are likely to be the same label as the present stimulus. We can calculate analytically the advantage of hysteresis in simplified cases. Consider a neural network that projects stimuli into $\mathbb{R}^1$ which represents the output label. Call this random variable X. The neural network must decide whether a stimulus is background or foreground.

Assume that the background stimuli have the same standard deviation as the foreground stimuli, and that it is a normal distribution. Also assume we penalize both type 1 and type 2 errors equally and over a long time period we get a 50:50 stimuli presentation. Without loss of generality, background stimuli average $-\mu$, foreground average μ, and the standard deviation is 1. If $\mu \gg 1$ we can cleanly distinguish them, but if $\mu < 2$ or so there will be a significant error rate.

Since we expect stimuli not to change between background and foreground all that often, we store our guess *G* of the previous state and use it to bias our decision. The guess can be "background" or "foreground". We have some hysteresis: If *G* is "background", we will switch *G* to foreground (we still use the background *G* for this stimulus, but next stimulus will use foreground) when $x > m$. If our guess is foreground, we will switch it when $x < -m$.
A combination of x and *G* is used to make a decision. Without hysteresis it is best to set the threshold to zero (due to the symmetry in the problem), but with threshold we set it to $\pm g$ depending on the state of *G*.

Stimuli switch between background and foreground stimuli with a switching probability p. If $p \ll 1$, one would expect to be able to use hysteresis to more accurately inform the answer.

This can be modelled as a 4 state system, with an accuracy percentage given by cumulative normal distributions. Let G- indicate background G, and S- indicate background stimuli, while "+" indicates foreground. Let Φ indicate the cumulative unit normal distribution; it is also possible to replace Φ with a different distribution.

State	Accuracy	Next = G-/S-	Next = G-/S+	Next = G+/S-	Next = G+/S+
G-/S-	$\Phi(\mu+g)$	$\Phi(m+\mu)(1-p)$	$\Phi(m+\mu)p$	$\Phi(-\mu-m)(1-p)$	$\Phi(-\mu-m)p$
G-/S+	$\Phi(\mu-g)$	$\Phi(m-\mu)p$	$\Phi(m-\mu)(1-p)$	$\Phi(\mu-m)p$	$\Phi(\mu-m)(1-p)$
G+/S-	$\Phi(\mu-g)$	$\Phi(\mu-m)(1-p)$	$\Phi(\mu-m)p$	$\Phi(m-\mu)(1-p)$	$\Phi(m-\mu)p$
G+/S+	$\Phi(\mu+g)$	$\Phi(-\mu-m)p$	$\Phi(-\mu-m)(1-p)$	$\Phi(m+\mu)p$	$\Phi(-\mu-m)(1-p)$

Table 5.7: Hysteresis switching rules.

We can calculate the equilibrium probabilities by equating how much goes into vs out of each. By grouping them into "bad" (accuracy $\Phi(\mu - g)$) and "good" (accuracy $\Phi(\mu + g)$) cases we get:

$$\Pr(bad)\left(\Phi(m-\mu)p + \Phi(\mu - m)(1-p)\right) = \Pr\,(good)(\Phi(m+\mu)p + \Phi(-\mu - m)(1-p))$$

Since the probabilities add up to unity we have:

$$\Pr(bad) = \frac{\Phi(m+\mu)p+\Phi(-\mu-m)(1-p)}{\Phi(m-\mu)p+\Phi(\mu-m)(1-p)+\Phi(m+\mu)p+\Phi(-\mu-m)(1-p)}$$

If p is small and m and μ are large and comparable, we are much less likely to be stuck in a "bad" situations.

The average accuracy is:

$$A = \frac{\left(\Phi(m+\mu)p + \Phi(-\mu-m)(1-p)\right)\Phi(\mu-g) + \left(\Phi(m-\mu)p + \Phi(\mu-m)(1-p)\right)\Phi(\mu+g)}{\Phi(m-\mu)p + \Phi(\mu-m)(1-p) + \Phi(m+\mu)p + \Phi(-\mu-m)(1-p)}$$

For a given μ we can find m and g to get the best accuracy.

Consider the case of vanishing p:

$$A_{P\to 0} = \frac{\left(\Phi(-\mu-m)\right)\Phi(\mu-g) + \left(\Phi(\mu-m)\right)\Phi(\mu+g)}{\Phi(\mu-m)+\Phi(-\mu-m)}$$

Simplify:

$$A_{P\to 0} = \frac{\Phi(-\mu-m)\Phi(\mu-g)}{\Phi(\mu-m)+\Phi(-\mu-m)} + \frac{\Phi(\mu-m)\Phi(\mu+g)}{\Phi(\mu-m)+\Phi(-\mu-m)}$$

When there is autocorrelation, a moderate amount of hysteresis helps with the classification accuracy as shown in figure 5.6.

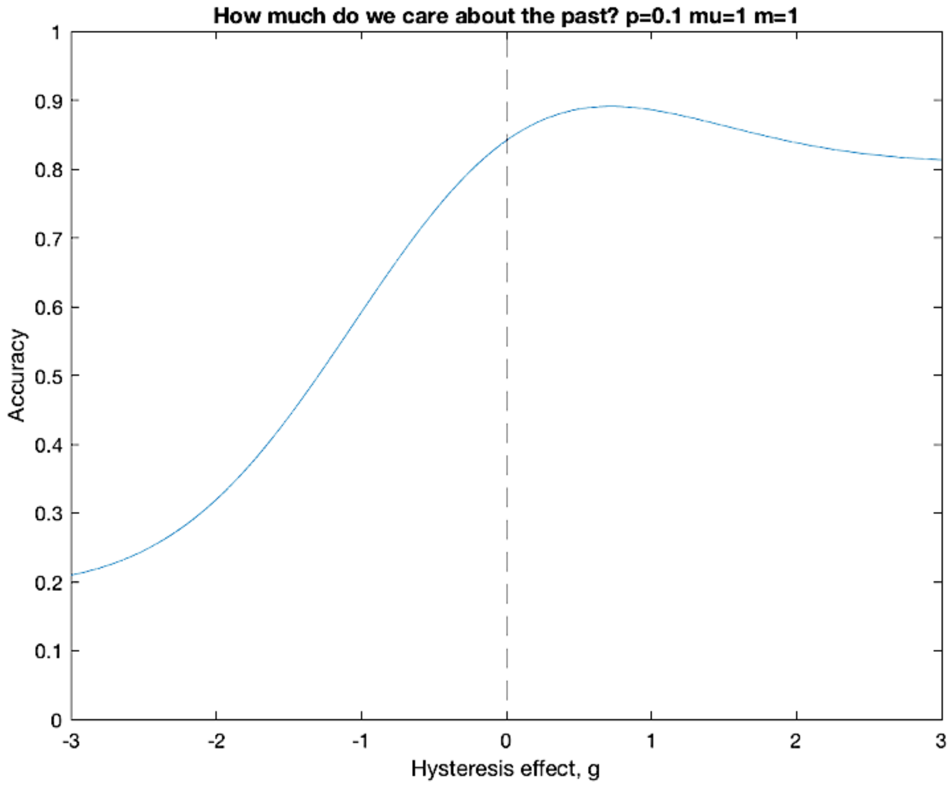


Figure 5.6: Accuracy as a function of g; g is how strong the threshold shift of hysteresis is. This is the "height" of the hysteresis loop. Other parameters p=0.1, m=μ=1. We assume normally distributed Φ.

Accuracy also depends on where we set the hysteresis threshold and degrades with higher switching probability if we have hysteresis, as shown in figure 5.7.

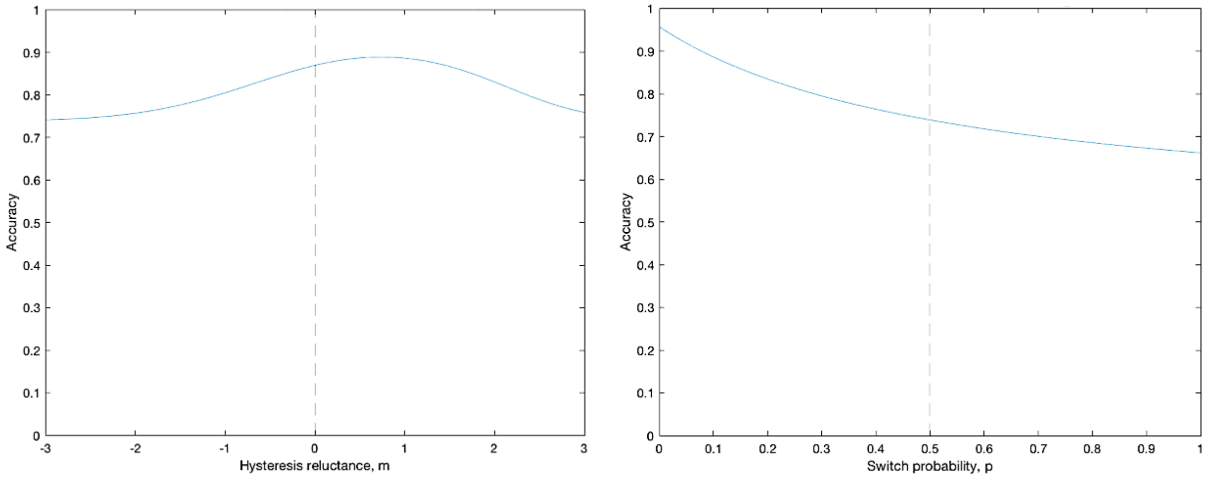


Figure 5.7: Accuracy as a function of m and p. A: Accuracy as a function of m; m is how strong the stimuli have to deviate from the mean to change the hysteresis. This is the "width" of the hysteresis loop. Other parameters: p=0.1, μ=1, g=1. B: Accuracy as a function of how often the stimuli switch p, other parameters m=μ=g=1. We assume normally distributed Φ.

What parameters in the problem maximizes the benefit of hysteresis? Small but non vanishing p does (p>0.5 actually would favor *negative* m). For optimal g and m, μ~0.7 is a sweet-spot that maximizes the benefit from hysteresis, for a maximum accuracy benefit of about 0.0584, as summarized in figure 5.8.

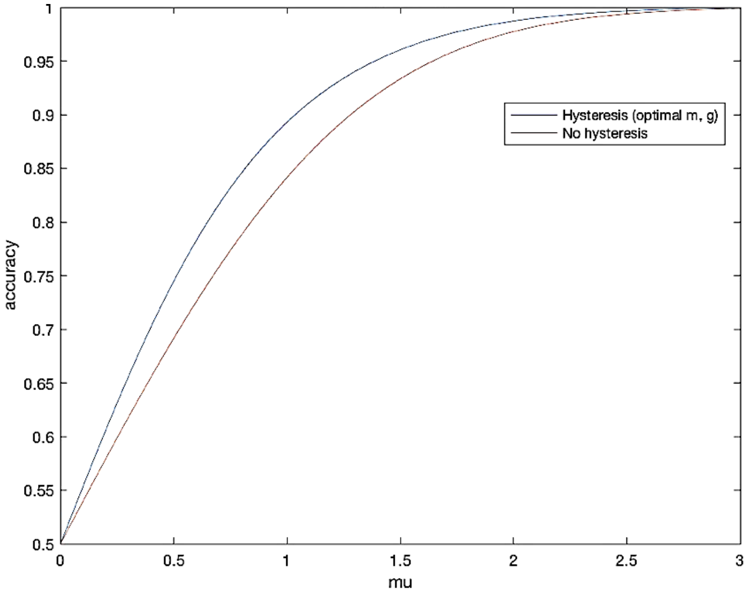


Figure 5.8: Accuracy of the optimal hysteresis parameters when p is small. The two parameters, m and g, are numerically optimized for each μ for each p vanishingly close to zero. This is compared to a model without hysteresis. We see that strong autocorrelation in the time series produces a slightly higher accuracy in the hysteresis. We assume normally distributed Φ.

The effect can be improved further with more complex systems. The <6% accuracy benefit is small. Having more variables (and more hysteresis bits) will "amplify" the benefit. Also, having an ultra-thin tail distribution (i.e. a truncated normal distribution) also was found to increase the benefit. Finally, averaging several network stimuli would increase the benefit of hysteresis as long as p stays very low.

Noise is inherent in any real neural process. A recurrent network that attempts to store its own firing rate, such as though a series of 1-neuron RELU autofeedbacks; neural integrators have been found in the oculomotor system (Cannon & Robinson, 1985); will be prone to drift. In order to store information robustly a network must either store it in the weights itself (which occurs during training) or be recurrent with multiple attracting patterns of activity. Said memory may need to be stored for a period of time while the animal processes the meaning of the sound stimuli.

The latter case is observed in a Hopfield network, and can be used to store multiple vectors of information; each training vector, if they aren't too numerous or degenerate, creates a basin of attraction in the Lyapunov function (Hopfield, 1982, 1984). Hysteresis is observed when a varying external forcing function is applied to a system in such a way that it pushes the state from one basin of attraction to another and back. Thus for robust memory that is stored in the firing patterns (most likely this is a form of short-term memory) hysteresis is inevitable.

5.7 Other potential recurrent network advantages

Recurrent networks may have other benefits beyond what was aforementioned. These could potentially be addressed experimentally.

Locality constraints as biological budgeting

Axoplasm takes space and energy, so the brain tries to minimize distances between connected neurons by wiring neurons locally when possible. If a recurrent network can send information through nearby channels it could avoid long-distance connections.

Robustness to damage

A recurrent network with n units will have O(n ^2) network connections. This is in contrast to a feedforward network with a single neuron as output which has a layer with O(n) connections. Damage in terms of a small fraction of connections ϵ set to zero will have O ($n\, exp(\epsilon^{-1})$) inputs that are completely ignored while the feedforward case has O($n\epsilon$). Indeed Hopfield networks have found to be robust to damage (Schonfeld, 1993).

CHAPTER 6:

Optimizing stimulus design for neural response and model identification

6.1 Introduction

Optimal stimulus design is the problem of selecting stimuli that maximizes relevant information about the neurons in question. It is an adaptive method, using the neural response to inform the best stimuli.

The space of perception

The naïve space of sound stimuli is the space of waveforms as a vector space. However, sampling fairly from that space gives white noise, which does not carry meaningful differences waveform to waveform.

Perceptual “just noticeable differences” (JNDs) provide a biologically relevant definition of a stimulus space. Mathematically, a JND is the “ds” term in differential geometry. The distance between any two stimuli is a geodesic: a sequence of intermediate stimuli that travels the least number of JNDs. From this, given infinite experimental resources, we could hypothetically find the metric at each point and build up a global picture of the stimulus space. In practice, the space is too far high dimensional to work with, so it must be simplified. One example is “random spectral shapes” stimuli developed in (Slee & Young, 2013) in which sounds are composed of pure tones spaced logarithmically and of randomized dB levels. However, this does not take into account any temporal structure. The sound textures in (McDermott &

Simoncelli, 2011) allow a perceptually rich range of sounds that coverers multiple linear and nonlinear statistics.

Model-free approach vs model-informed approach

A *model-free* approach of optimal design is applied when there is no underlying model or family of models concerning the neural network. One such approach is maximizing the response, which must be done through a gradient free method such as genetic optimization. The stimuli that gives the maximum response may be the most relevant information.

A *model-informed* approach of optimal design is used when there is a model with unknown parameters. The parameters of the model (as well as which of several models is best) are usually the relevant pieces of information. The stimuli designed do not maximize the response of the neuron, but instead maximize of the model, but instead maximize how much the entropy in the parameter distribution function is reduced. This approach is made mathematically rigorous in (DiMattina & Zhang, 2011) for feed-forward networks.

6.2 Online, model-free determination of optimal stimulus

Optimal stimulus design is a challenge to apply to cortical experiments due to a combination of the recurrent nature of the cortex (which makes it harder to try to combine the neurons "one level" upstream into a model neuron) and the sparsity. For a simple one-layer network, the optimal driving stimuli is very close to the weights (nonlinearities in the gain function make the shape slightly different, but this effect is small). However, harmonically selective units are likely more complex than just feed-forward sieves because of the ubiquity of recurrent connections in the cortex and the success of recurrent networks as shown in Chapter 4. The sound textures in

(McDermott & Simoncelli, 2011) capture a much richer repertoire of sound stimuli. They compute the envelopes across 32 frequency channels and various statistics in each band as well as modulation power spectra and cross-correlations. This is summarized in figures 6.1, 6.2, and table 6.1. These textures allow a perceptually rich range of sounds that are more likely to drive selective units but the inverse problem of computing a sound from a texture is computationally expensive.

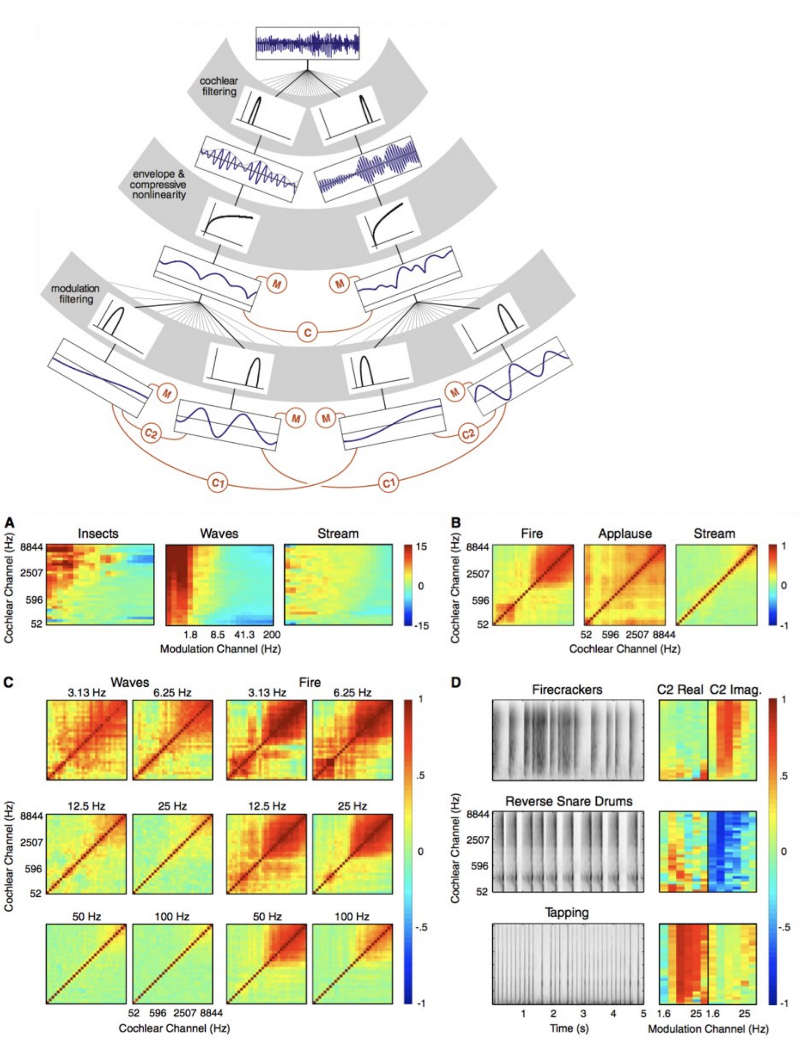


Figure 6.1: Schematic of how the sound textures are generated. Top: summary of sound texture calculation in McDermott, 2011. Bottom: Plots of textures for various sounds. *Figure from (McDermott & Simoncelli, 2011).*

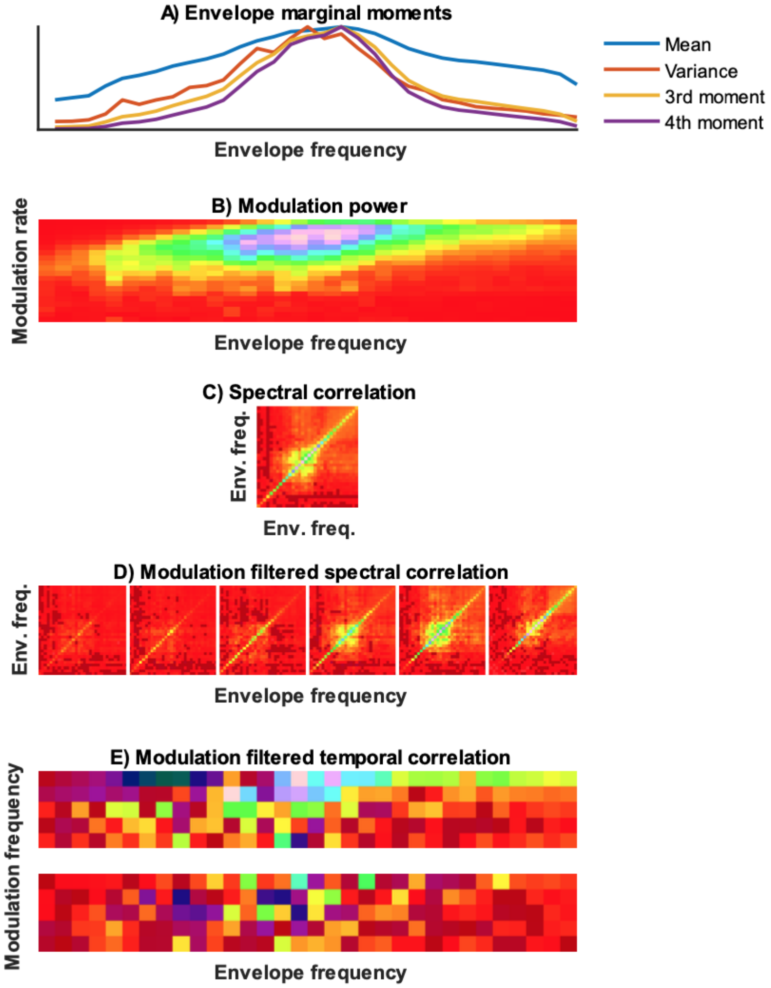
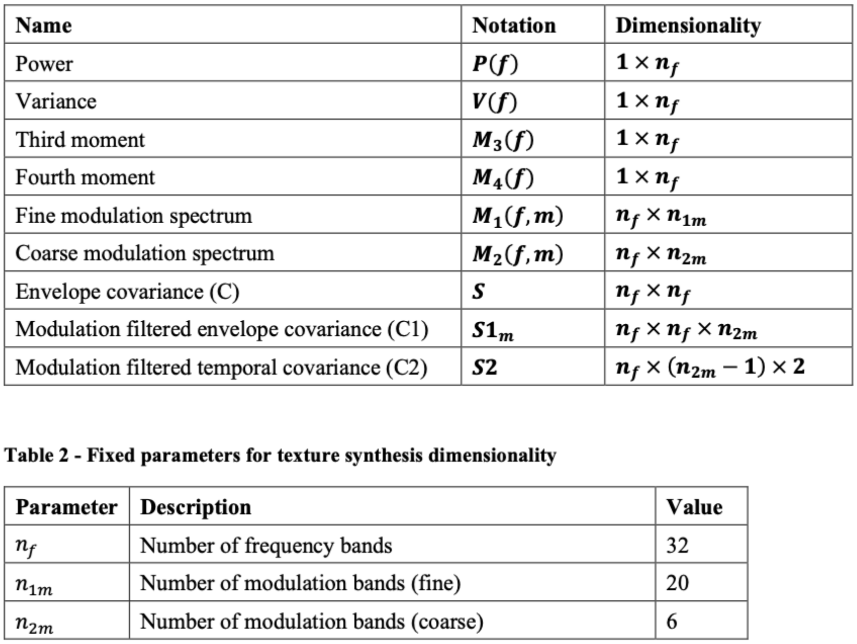


Figure 6.2: An example of summary statistics for a sound texture. *Figure from (Gamble & others, 2020).*



Name	Notation	Dimensionality
Power	$\boldsymbol{P(f)}$	$\boldsymbol{1 \times n_f}$
Variance	$\boldsymbol{V(f)}$	$\boldsymbol{1 \times n_f}$
Third moment	$\boldsymbol{M_3(f)}$	$\boldsymbol{1 \times n_f}$
Fourth moment	$\boldsymbol{M_4(f)}$	$\boldsymbol{1 \times n_f}$
Fine modulation spectrum	$\boldsymbol{M_1(f,m)}$	$\boldsymbol{n_f \times n_{1m}}$
Coarse modulation spectrum	$\boldsymbol{M_2(f,m)}$	$\boldsymbol{n_f \times n_{2m}}$
Envelope covariance (C)	$\boldsymbol{S}$	$\boldsymbol{n_f \times n_f}$
Modulation filtered envelope covariance (C1)	$\boldsymbol{S1_m}$	$\boldsymbol{n_f \times n_f \times n_{2m}}$
Modulation filtered temporal covariance (C2)	$\boldsymbol{S2}$	$\boldsymbol{n_f \times (n_{2m} - 1) \times 2}$

Table 2 - Fixed parameters for texture synthesis dimensionality

Parameter	Description	Value
n_f	Number of frequency bands	32
n_{1m}	Number of modulation bands (fine)	20
n_{2m}	Number of modulation bands (coarse)	6

Table 6.1: All the statistics and their dimensionality. *Table from (Gamble & others, 2020)*.

Accelerating the inverse problem

We increased the performance ~8 fold of solving the inverse problem of generating a sound from a texture by solving for the envelopes. This problem is also nonlinear and is solved iteratively by minimizing the difference between the desired texture and the actual texture generated by the soundwaves or envelopes using the conjugate gradient method, starting from a random signal.

This optimization addresses the main computational expense. There are 32 spectral filters used in (McDermott & Simoncelli, 2011) which means computing the envelopes form the wave requires computing 32 Hilbert transforms at the sound's sample rate. However, each envelope can be safely down sampled by about 10 fold and these Hilbert transforms can be skipped.

The conjugate gradient method needs to constrain the signs of the envelopes. This is done by adding a penalty term: $E_{sign} \sim \sum x_i^2 [x_i < 0]$. All the other terms penalize the square-norms of how different one block of the summary statistics in the texture is form the desired texture. The various terms in the objective function are weighted with the weights manually adjusted so that each term is usually optimized at the same speed and the sign error is around 1% during optimization. Despite the complexity of the constraint, the speed up is still significant: From about 25 minutes to about 3 minutes for each sound, allowing a bank of sounds to be computed in this shorter time frame simultaneously on different cores.

Filling the wave in under the envelopes

The filling in of the waveform is computed once, after the iterative process is complete. For each spectral filter we fill in the fine-structure such that the resulting waveform has the same range of frequencies as the filter, at least approximately.

Suppose F is the filter in the spectral domain; we will only consider real-valued F. We can *sample* from F by representing F as a discrete vector, sampling uniformly on the rectangle enclosing the filter, and accepting points only if they are under the curve. This process is illustrated in figure 6.3. We call X the random variable of this sampling process.

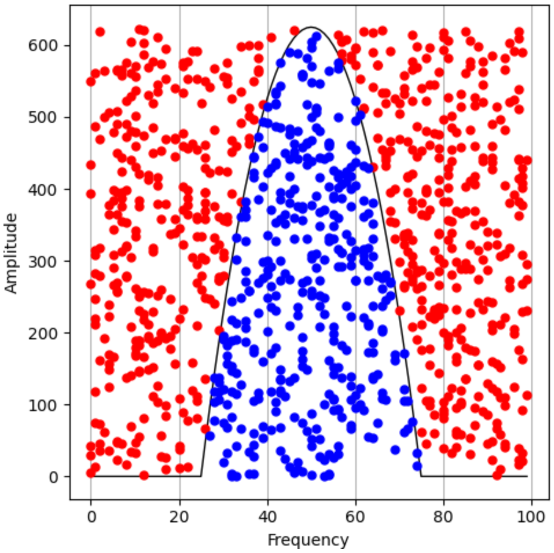


Figure 6.3: Sampling from arbitrary distributions. Any distribution that is represented by a probability density vector can be achieved by sampling uniformly from the rectangle enclosing the distribution. The x-axis location of the blue points are our samples.

Once we have our samples of F we generate a frequency function $f(t)$ which is defined at the sample times as: $f(t_i) = X_i$. Intermediate values of t use a linear interpolation. The sample times are linearly spaced, with the spacing given by:

$$\Delta_t = t_i - t_{i-1} = \frac{1}{2\pi\sigma}, \qquad \sigma = E(X^2) - E(X)^2$$

Here σ represents the bandwidth of the filter, which we convert to radians per second in order to get the "time uncertainty" of the filter Δ_t. Broader filters will have broader σ values and will have the instantaneous frequency in the fine structure changing more frequently. The final fine-structure function is $s(t) = \cos\left(2\pi \int_0^t f(\eta) d\eta\right)$. This integral is discretized to the sound's sample rate and solved numerically, as illustrated in figure 6.4.

The soundwave for the k'th channel is: $y_k(t) = s_k(t) Env_k(t)$ where $Env(t)$ is the up-sampled envelope, as illustrated in figure 6.5. The total waveform is the sum of all channels: $y(t) = \sum_k s_k(t) Env_k(t)$

The sum of the square-amplitudes in the filterbank in (McDermott & Simoncelli, 2011) is equal for all relevant frequencies. This means there is no frequency under or over representation that must be corrected for, so the simple sum works.

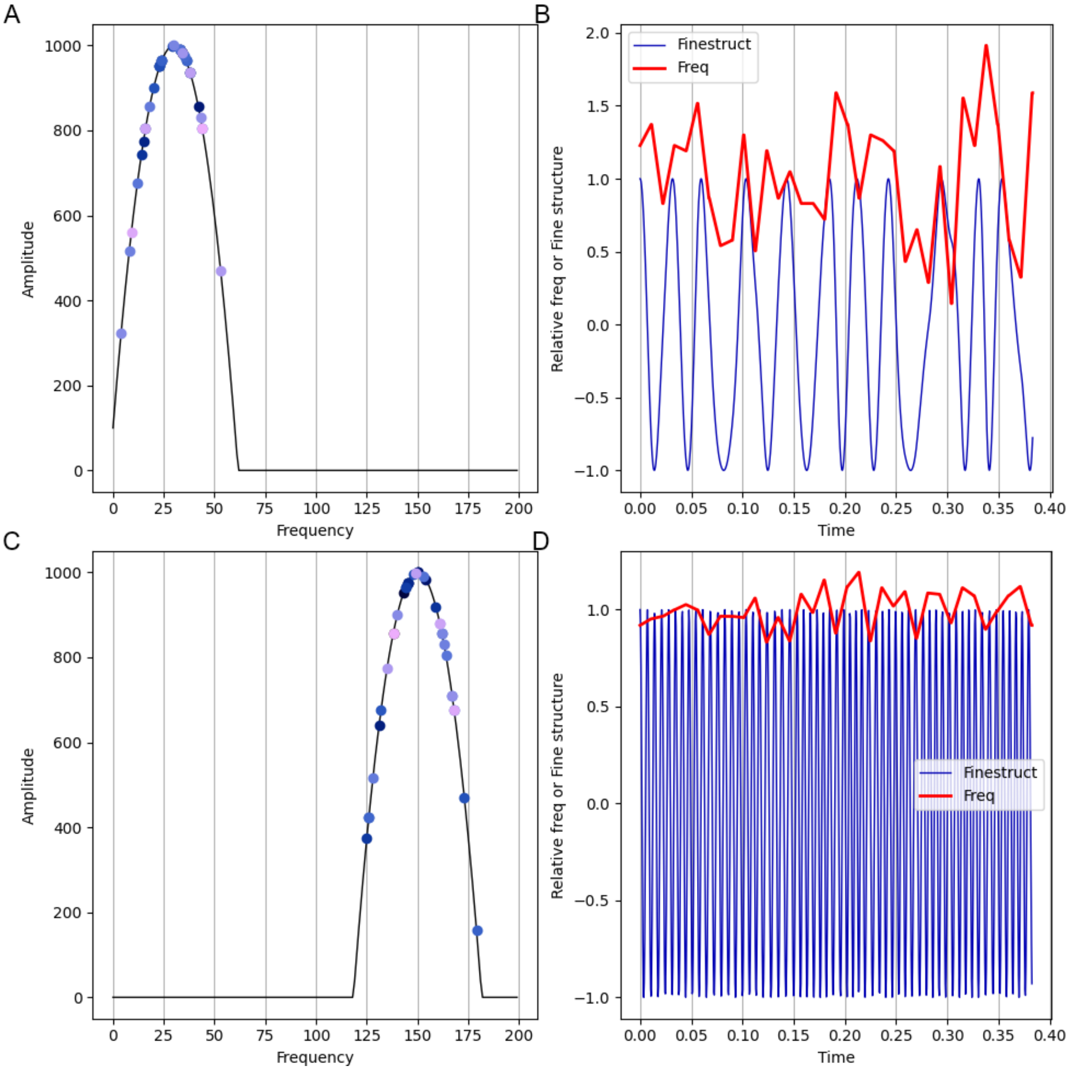


Figure 6.4: The process of calculating the fine structure. A,C: The frequencies are sampled from the envelope filter, and are more likely to be near the middle of the filter. B,D: To get the fine structure, the sampled frequencies are linearly interpolated (red curve) and the frequency is integrated over time and the cosine is taken. At higher points on the red curve, the blue curve will tend to change faster. The Q-factor is much higher in C than in A, making the fine-structure in D much more regular than in B.

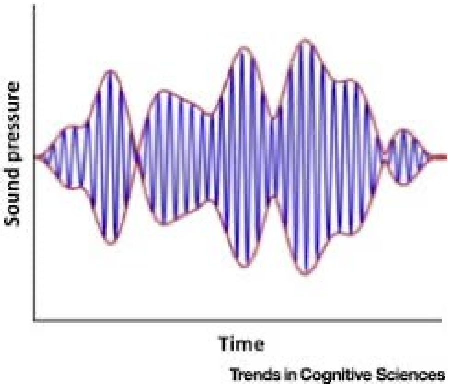


Figure 6.5: Generating the soundwave the fine structure and envelope. The process is simply to component-wise multiply them. The red curve is the envelope and the blue curve is the result. *Figure from (Shannon, 2016).*

Results

The process of generating envelopes from a complex sound and reconstructing the fine structure as mentioned previously resulted in a hardly noticeable perceptual difference in the sound or sound-from-texture process.

The speedup allowed a collaborator, Darik Gamble, to run a genetic algorithm using the response of an awake marmoset A1 unit as it's fitness function. The collaborator plans to publish these results, which have been successful in climbing the fitness landscape for most neurons. He expects these neurons to be background detectors that inhibit a foreground detector.

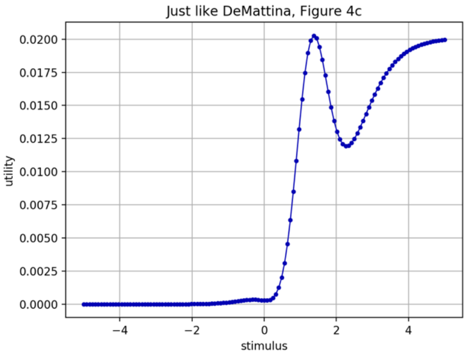
6.3 Computational model-based optimal stimulus design for recurrent networks

Stimuli can also be designed to maximize the information about a model neuron or neural network. I implemented the optimization method in (DiMattina & Zhang, 2011), which can be applied to any feed-forward network. The model uses a Gaussian mixture to encode the uncertainty of the parameter weights. It finds the stimulus that generates the expected maximum increase of the Fisher information, which amounts to the log of the determinant of the covariance matrixes of the parameters. It uses random-starts and resets to deal with non-linearities and local minimums.

This should accelerate experiments provided the model is accurate, one could compare the number of trails needed to estimate parameters to a given level of prediction accuracy to verify that in fact it does make a significant improvement in the stimulus design. The generated sounds may oscillate between several sounds each specialized to detect a single parameter, such *bf*$_0$ vs maximum response.

Reproduction of DiMattina and Zhang, 2011

We were able to reproduce the model in (DiMattina & Zhang, 2011). Briefly, we maximize the determinate of the fisher information matrix which is, under our assumptions of normal, the inverse covariance matrix of the parameter estimation, which is summarized in figure 6.6.



$$\sum_{i=1}^{N}\left(\frac{1}{\nu}\nabla_{\boldsymbol{\theta}} f(\mathbf{x},\boldsymbol{\theta}_i)^{\mathrm{T}}\mathbf{F}_n^{-1}(\boldsymbol{\theta}_i)\nabla_{\boldsymbol{\theta}} f(\mathbf{x},\boldsymbol{\theta}_i)\right) p_n(\boldsymbol{\theta}_i),$$

$$L(\mathcal{D}_n \mid \boldsymbol{\theta}) = -\sum_{i=1}^{n}\frac{1}{2r_i}(r_i - f(\mathbf{x},\boldsymbol{\theta}_i))^2 - \ln\sqrt{2\pi r_i},$$

Figure 6.6: The maximum information stimulus design figure reproduction. Here, we reproduce (DiMattina & Zhang, 2011) in which we find the stimulus ***x*** that maximizes the above equation (where *f* is the network's output), and then updates its estimate of the parameter $\boldsymbol{\theta}$ by maximizing the bottom function.

Optimal stimulus design for recurrent networks

Two-unit recurrent networks can be evaluated at the equilibrium state for optimal stimulus design purposes (Doruk & Zhang, 2019). We adapt this method for many-unit networks with Hebbian weights trained with a harmonic stimulus. We outline a way to adapt the optimal stimulus model to a recurrent network with a sigmoid gain and symmetric weights. From Chapter 5, the Lyapunov function in activity-space s a unit sigmoid gain and symmetric W is:

$$L(s) = \sum(slns + (1-s)\ln(1-s)) + (b-j)^T s - \frac{1}{2}s^T W s$$

The network minimized this function locally to find s_{eq}. For simplicity, we assume that there is additive Gaussian noise in the activity with variance ν, which may be proportional to the mean if there is a Poisson noise model. We also assume that our estimate covariance matrix is fairly tight so that we don't have to integrate over the space of possible parameters but can instead evaluate the equation at the mean parameters.

We have a stimulus x and we want to evaluate how useful it is in improving our Fisher information matrix F, by maximizing the determinate of the updated Fisher matrix. For any stimuli x, we compute the subcortical input j by convolving it with a gaussian kernel of fixed width, with the same parameters as in Chapter 3: $j = K * x$. We do not parameterize the stimulus with a smaller subset of parameters, but it would be simple to do so in order to restrict ourselves to harmonic stimuli.

For any stimulus and s value we can compute the first and second derivatives of L:

$$\frac{dL}{ds} = \ln s - \ln(1-s) + b - j - Ws, \quad \frac{d^2L}{ds^2} = diag\left(\frac{1}{s} + \frac{1}{1-s}\right) - W$$

W and b aren't known, but we can use the men of our parameter estimate to calculate them. We solve for s_{eq} by using the conjugate gradient method (128 iterations) which minimizes a modified function with a penalty for the edges and better numerical stability:

$$s_{clamp} = \max\left(\epsilon_{edge}, \min\left(s, 1 - \epsilon_{edge}\right)\right), \qquad \epsilon_{edge} = 1e - 5$$

$$L_{opt} = L\left(s_{clamp}(s)\right) + \sum \frac{1}{2} k_{edge}\left((s-1)[s > 1] - s[s < 0]\right)^2$$

$$\frac{dL_{opt}}{ds} = \frac{dL}{ds} .* \left(s_{clamp}(s) = s\right) + \sum k_{edge}\left((s-1)[s > 1] - s[s < 0]\right)$$

The penalty weight is extremely high at $k_{edge} = 2000$. It must be this high for the optimization to be driven back to the allowed region of s.

We then compute how sensitive s_{eq} is to W and b. We define a locally quadratic expansion of L:

$$L(s) = \mathrm{L}(\mathrm{s}^*) + \frac{dL}{ds_{s=\mathrm{s}^*}}^{T} (s - \mathrm{s}^*) + \frac{1}{2}(s - \mathrm{s}^*)^T \frac{d^2L}{ds^2_{s=\mathrm{s}^*}} (s - \mathrm{s}^*) + O|s - \mathrm{s}^*|^3$$

If $\mathrm{s}^*$is near the equilibrium we can calculate the equilibrium accuracy by setting the slope of this to zero:

$$\frac{dL}{ds_{s=s*}} + 2\frac{d^2L}{ds^2_{s=\mathrm{s}^*}}\left(s_{eq} - \mathrm{s}^*\right) + O\left|s_{eq} - s^*\right|^2 = 0$$

$$s_{eq} - s^* = -\frac{1}{2}\left(\frac{d^2L}{ds^2_{s=\mathrm{s}^*}}\right)^{-1} \frac{dL}{ds_{s=s*}} + O\left|s_{eq} - s^*\right|^2$$

We are solving for $\frac{d\left(s_{eq} - s^*\right)}{dW,b}$ at the equilibrium point, namely at $s^* = s_{eq}$. This makes these approximate equations exact. Also, since $\frac{dL}{ds_{s=s_{eq}}}$ is zero the chain rule term that differentiates $\left(\frac{d^2L}{ds^2_{s=\mathrm{s}^*}}\right)^{-1}$ is zero, simplifying the equation significantly:

$$\left(\frac{d\left(s_{eq}\right)}{dW}\right)_{ijk} = \left(\frac{d^2L}{ds^2}_{s=\mathrm{s_{eq}}}\right)^{-1}_{ki} s_{eq,j}$$

In this case, the i index represents the index on s_{eq} and the *j,k* indexes are the index on W.

Also, $\frac{d(s_{eq})}{db} = -\left(\frac{d^2L}{ds^2}_{s=\mathrm{s}^*}\right)^{-1}$ is a symmetric matrix.

We use chain rule to compute how the equilibrium point would move under a small change in the parameters:

$$\left(\frac{ds_{eq}}{d\theta}\right)_{il} = \left(\frac{ds_{eq}}{dW}\right)_{ijk}\left(\frac{dW}{d\theta}\right)_{jkl} + \left(\frac{ds_{eq}}{db}\right)_{ij}\left(\frac{db}{d\theta}\right)_{jl}$$

In this case, for $\frac{dW}{d\theta}$ the j, k indexes represent the element of the (symmetric) weights W, and the l index is the index on the parameters θ. For $\frac{db}{d\theta}$ the *j* index represents elements of *b* and the l index represents elements of θ. For $\frac{ds_{eq}}{d\theta}$ the i index represents elements of the equilibrium activity s_{eq} and the *j* index represents the elements of θ.

Suppose we are recording the k'th neuron in the network. We can adapt equation 2.5 in (DiMattina & Zhang, 2011):

$$F_{n+1} = F_n + \frac{1}{\nu_k}\frac{ds_{eq,k}}{d\theta}\left(\frac{ds_{eq,k}}{d\theta}\right)^T$$

If we are recording from all the neurons, given independent noise for each neuron, we have:

$$F_{n+1} = F_n + \sum_k \frac{1}{\nu_k}\frac{ds_{eq,k}}{d\theta}\left(\frac{ds_{eq,k}}{d\theta}\right)^T$$

The utility function, for our D-optimal design, is $u(j)$ =ln($|F_{n+1}|$)

This allows us to estimate the utility function for any stimulus j.

It is possible, in principle, to optimize the utility function through conjugate gradient by computing $\frac{du}{dj}$. However, it risks error accumulation in s_{eq} drifting away from the true equilibrium due to nonlinearities in the problem. Note that this would not be a problem in feed-forward methods that have explicit response. Also, gradient methods are local methods and we want to test a wider range of j values. Thus we optimize j through CMA-ES, which is a genetic algorithm that is relatively robust to.

Egg-crate weight parameterizations

We need to design a parameterization, namely we need to specify $W(\theta)$ and $b(\theta)$, and then compute $\frac{dW(\theta)}{d\theta}$ and $\frac{db(\theta)}{d\theta}$. For our network θ has four parameters and guarantees a symmetric weight matrix (which forbids using Dale's principle but guarantees stable equilibrium for our sigmoid gain). The third parameter is the number of components represented, which is 4.5. This is used to generate an "eggcrate" which is equivalent *up to a shift and scale* to the three (anti)Hebbian quadrants in the weight matrix shown in figure 4.9. The first two parameters are the mean and standard deviation of the elements of the weight matrix, respectively, and are multiplied by the number of channels (to keep the parameters scale-invariant). These are set to 0.5 and 0.75 respectively. Finally, the fourth and last parameter is the threshold of each neuron (which does not vary from neuron to neuron).

The derivatives $\frac{dW(\theta)}{d\theta}$, and $\frac{db(\theta)}{d\theta}$ are third-rank tensors and matrixes, respectively, as defined previously. They are calculated with difference quotients, which is efficient in this case because we are differentiating with respect to a scalar for each parameter.

Optimal stimulus design for recurrent networks, results

For our parameter sets, the equilibrium state to various stimuli is sieve-like, with the sensitivity not depending strongly on the choice of stimuli, see figure 6.7 and figure 6.8.

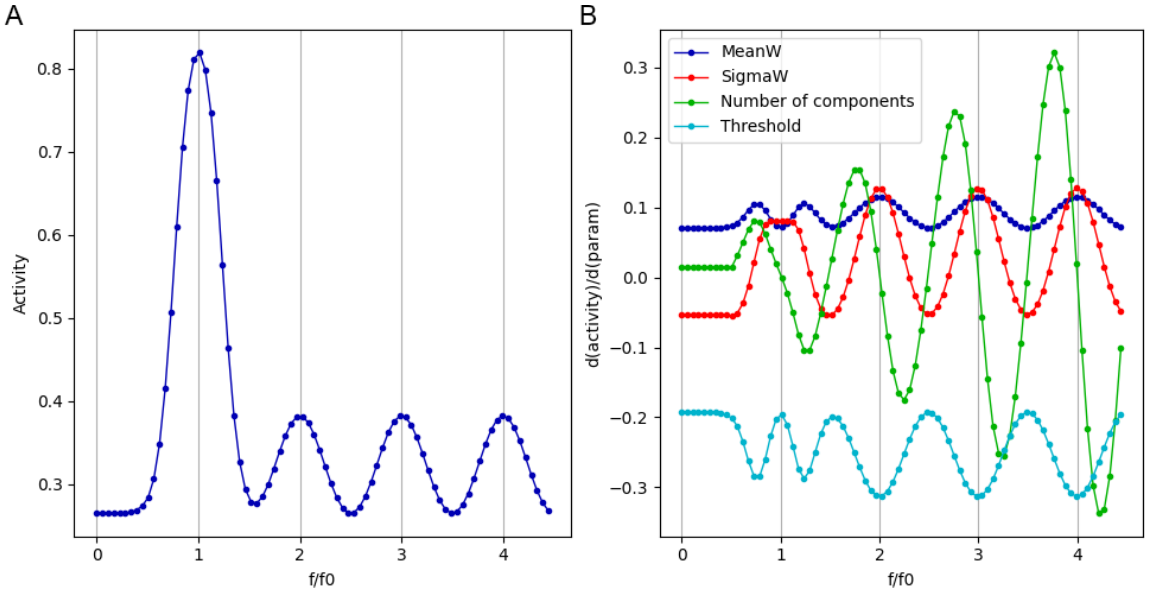


Figure 6.7: The equilibrium state and sensitivity given a tone stimulus. A: The equilibrium state. B: the sensitivity of this equilibrium to each parameter of θ.

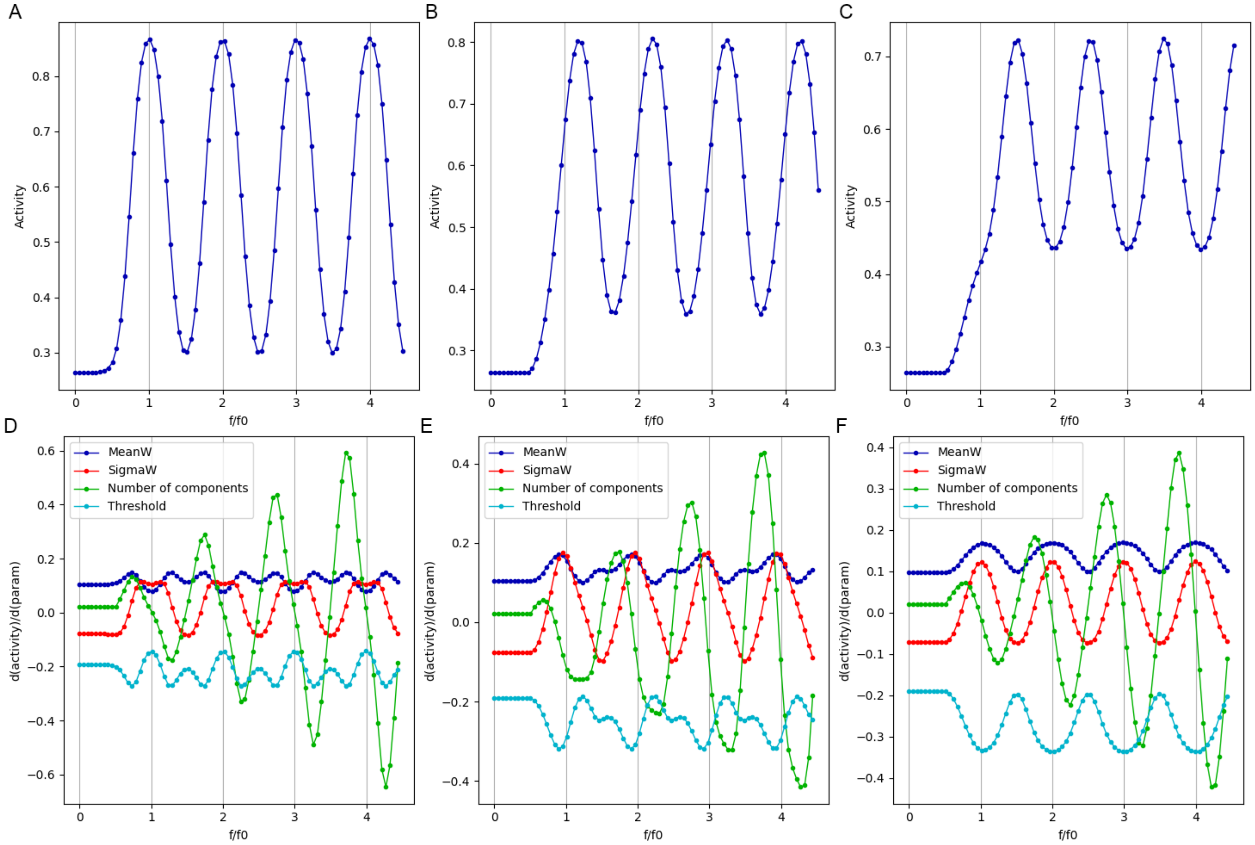


Figure 6.8: The equilibrium state and sensitivity given harmonic/mistuned stimuli. All stimuli have f0=Bf0 with the Bf0 determined by the number of components which is the third element of θ. A,B,C: The equilibrium states. D,E,F: The sensitivities. A,D: 0% mistuning. B,E: 25% mistuning. C,F: 50% mistuning.

On the other hand, the optimal stimulus depends strongly on the relative amount of information we are trying to find for each θ. For an identity covariance matrix, the stimulus is approximately harmonic with f0 = ½ Bf0 and is mistuned 50%, see figure 6.9. For deducing each element of θ the stimulus tends to be either harmonic or mistuned with f0=Bf0, see figure 6.10.

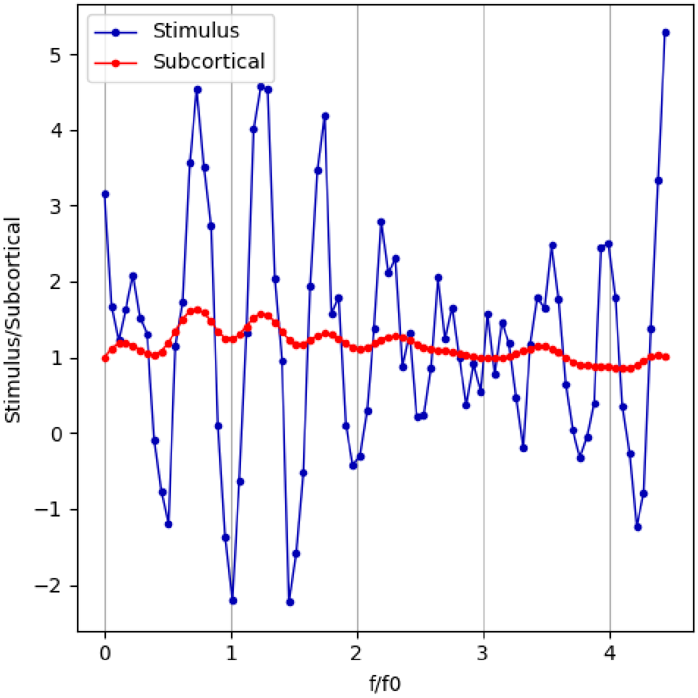


Figure 6.9: The optimal stimulus when given an identity covariance matrix of θ. The subcortical input is shown as well, it is a blurred version of the stimulus.

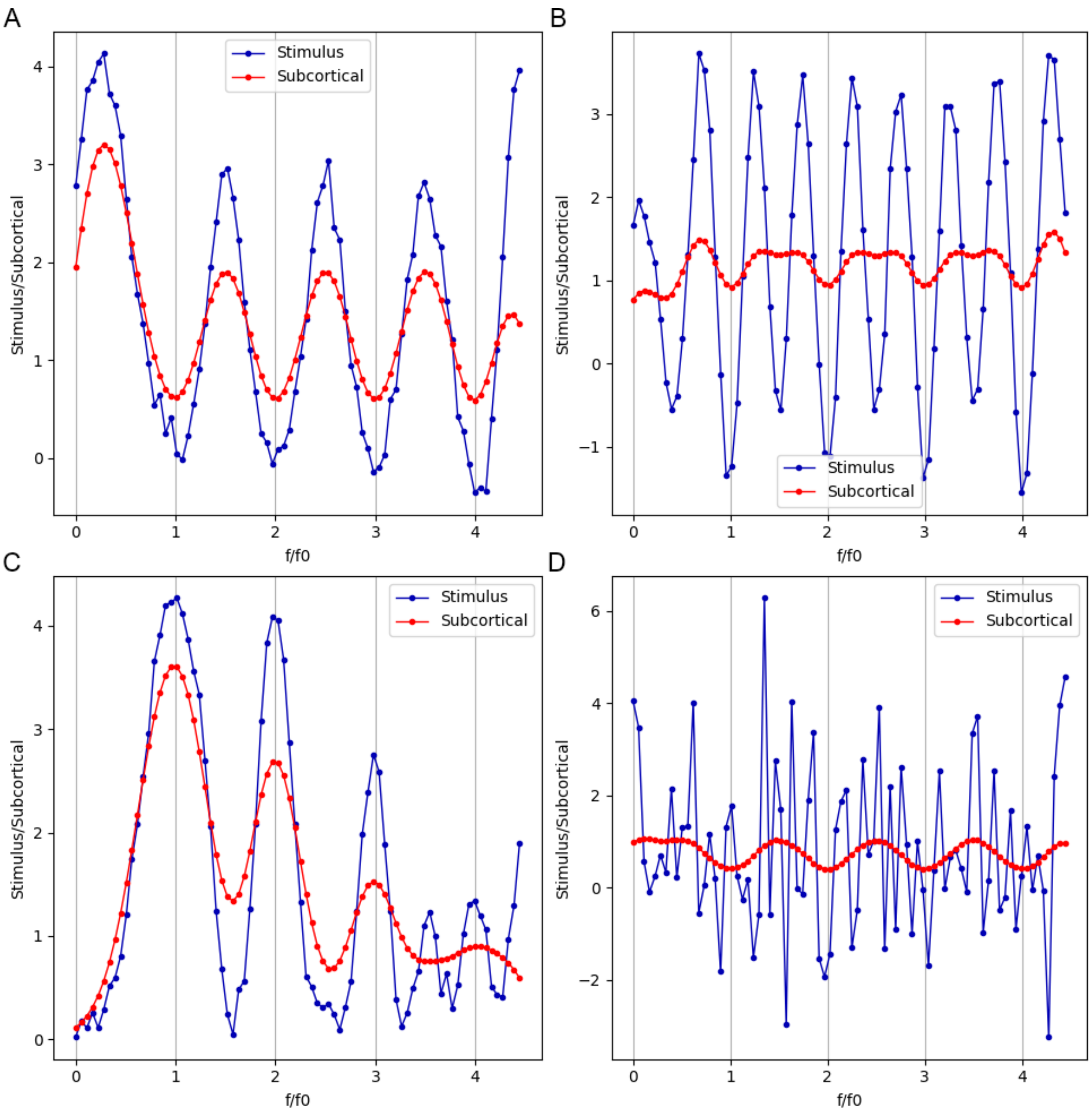


Figure 6.10: The optimal stimuli to deduce the individual elements of θ. A: For deducing the mean value of W. B: For deducing the standard deviation of W. C: For deducing the number of components. D: for deducing the threshold b.

6.4 Discussion

We were able to improve and extend two optimal stimulus design approaches. The model-free approach was computationally expensive, but we accelerated it by solving for the envelopes rather than the waveforms in the conjugate gradient process. The model-informed approach was extended to handle symmetric recurrent networks with analytically calculated Lyapunov functions.

Both models could be extended further. The model-free approach could try to measure the range of responses rather than simply finding the maximum. The model-informed approach could be extended to asymmetric networks, in particular ones that respect Dale's principle. Both of these are potential for future work.

CHAPTER 7:

Conclusion

This thesis presents a theoretical study of possible computational mechanisms for harmonic selectivity in the auditory cortex. From a functional point of view, our analysis of natural sounds from auditory nerve responses suggests that harmonic selectivity in the auditory cortex may emerge as an efficient representation of natural sounds that contain harmonic components in their statistics. Neural network models developed in this thesis show how harmonic selectivity similar to the *harmonic template units* found in the marmoset auditory cortex might be generated by combining subcortical inputs that are tuned to single frequency bands. In feedforward networks, the most robust mechanism as revealed by supervised learning for harmonics detection is a spectral sieve with alternating excitatory and inhibitory weights along the frequency axis. In recurrent networks, harmonic selectivity can be achieved by recurrent connection weights established by unsupervised Hebbian learning. Other issues addressed in the thesis include how time-domain selectivity, but not frequency-based selectivity, could be related to synaptic time constants, and how stimuli can be designed to most efficiently drive cortical neurons or to reduce uncertainty about models. Detailed summary and conclusion of the key results are as follows.

7.1 Temporal harmonic selectivity mechanisms

Harmonic sounds are associated with a temporal regularity, which means that at very low fundamental frequencies the cortex should be able to detect this periodicity. Integrate-and-fire models have a benefit here in that they directly track membrane and synaptic time constants. By adding synaptic depletion into both the excitatory and inhibitory inputs of an integrate-and-fire model neuron, it expands the range of behaviors to include most of the modulation-sensitive click-train responses observed in the awake marmoset cortex: model neurons could also respond in either a phasic or tonic manner and sometimes responded *less* to click trains that were too high of a frequency. However, the exquisite periodicity selectivity in pitch-sensitive neurons has yet to be modelled. Crucially, this model has no recurrent connections. Adding short-loop recurrent microcircuits, perhaps with the help of future experiments tailored to find them, may generate more periodicity selectivity.

7.2 Subcortical models

Most cortical models need a supporting subcortical model. We found that we could produce specialized models that capture the most famous aspects of the time-response of the basic types of cochlear nuclear (CN) neurons, including the chopper neurons by using a smoothed integrate-and-fire model. However, at the inferior colliculus (IC) there was too little data and too much of a combinatorial explosion in the possible network topologies to generate a comprehensive model.

Fortunately, we only needed spectral-domain models to feed into our cortical models. In this case the most important properties of the IC neurons are the spectral receptive fields, of which a fair amount of data has been collected. We model type I neurons, the most frequency-selective neurons which have pure-tone receptive fields that are usually close to a gaussian response centered on the best frequency (Schreiner & Winer, 2005). Although two-tone sharpening and saturation effects were not captured by this simple Gaussian-convolution model, the model still served us well and we deemed it reasonable as these effects weren’t particularly strong in harmonic/mistuned stimuli given to the Marmoset IC (Kostlan, 2015).

7.3 Natural sound statistics generate some harmonically selective units

At higher frequencies, when f_0 is above about 100 Hz, the cortex cannot follow the temporal periodicity of the sound waveform and thus operates primarily in the spectral domain (Besser, 1967). Also, Q-factors in the cochlea are higher at moderate and high frequencies which makes spectral sieving easier (Zilany et al., 2013). Our analysis is focused in this spectral sieving regime.

A natural bank of sounds contains many harmonic sounds as well as many non-harmonic sounds or sounds with a fleeting harmonic-like structure. Applying either PCA or ICA to an auditory nerve (AN) model generates a series of spectral filters that can be thought of a neural population that extracts the most salient features of a sound. Both can produce a reasonable proportion of harmonically selective neurons but are undoubtedly extracting other properties of the sound as well. ICA is more biologically relevant as it generates sparser filters with center-surround and Gabor-like patterns which better resemble experimental receptive fields. Also, ICA has a strong theoretical footing in terms of information maximization (in contrast to minimizing mean-square error as does PCA). This allows it to better capture higher order statistics which are relevant in most natural sounds.

The harmonically selective units generated by PCA and ICA were not stereotypical sieves. However, 2-3 well-positioned excitatory and inhibitory components is enough to meet the harmonic template unit criteria in terms of facilitation index (FI) and periodicity index (PI) in (Feng & Wang, 2017) if we consider these weights to feed into a zero-threshold ReLU unit. If the threshold is above zero, "iceberg" effects can make the FI and PI arbitrarily close to unity.

7.4 Training harmonic template units produces sieves

Sound statistics can generate harmonic selectivity, but they do not address what receptive fields make the most harmonic selective neurons. A spectral sieve, which is excited by integer multiples of a fundamental frequency and inhibited by half-integer multiples, is the most straightforward way to generate harmonic selectivity. This is indeed the only supervised learning solution that was found. Sieves persisted across a wide variety of different sets of background and foreground stimuli, noise models, and other training conditions, including when several biological considerations are taken into account. The sieve that was trained is very similar to an ideal linear classifier in most cases.

Dale's principle states that no neurons are both excitatory and inhibitory, which means it is realistic to include two sets of neurons that get subcortical input, one excitatory and one inhibitory in terms of its effect on the output neuron. Doing this splits the positive and negative part of the sieve between the excitatory and inhibitory group, producing very similar responses to the sign-agnostic case. Another biological consideration is sparsity. Adding in a sparsity constraint to the weights made the sieve spikier, but it was still a sieve. Adding one or more hidden layers didn't make the neural network behave much differently than a simple sieve unless it failed to produce selectivity at all. Thus sieves are both the most straightforward way and the most robust and only way to produce harmonic selectivity.

7.5 Recurrent network harmonic selectivity

Cortical neurons are highly dependent on recurrent connections. This throws into question whether a feedforward paradigm is the way harmonic selectivity is achieved. Recurrent networks are most commonly trained with a Hebbian rule. The sign-agnostic Hebbian recurrent network that is generated from harmonic training stimuli is an eggcrate pattern.

It is more realistic to invoke Dale's principle and separate the network into excitatory and inhibitory banks of neurons that both get subcortical input from each neuron's best frequency. This gives us four blocks in the weight matrix depending on whether the origin and destination is excitatory or inhibitory. Each block can have a different rule. We found the best results when the within-excitatory block is Hebbian and the inhibitory block is anti-Hebbian, but the inhibitory-to-excitatory connections are a simple surround-inhibition "trough", and a mild Hebbian excitatory to inhibitory connection acts to reduce the overall activity. This network features co-tuning of excitation and inhibition, unlike the sieve case. Indeed, anti-Hebbian learning among inhibitory neurons has been experimentally measured and represents inhibitory synapse getting *stronger* under coincidental activation much like the excitatory Hebbian rule. However, there is still a large variety of realistic learning models and parameters to choose from so one may have even better selectivity.

7.6 Recurrent network fingerprints

We found that our recurrent networks, depending on the training parameters used, have a variety of behaviors that feedforward networks do not have that mirror certain experimental results. Psychophysical experiments have found that hysteresis is nearly universal among all the senses. Indeed, hysteresis is the most easily found property of our networks. It occurs when mutual connections between excitatory weights are strong enough to allow for a positive feedback loop.

Results from psychophysics also indicate that the auditory system has a “filling in” effect: it perceives pitch at the fundamental frequency of a harmonic complex even when said fundamental component is missing. This was seen in our network, where a single missing component did not stop the neurons there from activating upon hearing a harmonic stimulus because it was supported by excitatory connections form neurons where there was still energy.

Finally, recurrent networks can oscillate when the excitation and inhibition both strongly feedback onto each-other. EEG’s of the brain show several frequency bands of oscillation during physiological functioning as well during seizures. Our network displayed oscillation when the excitatory-to-inhibitory and inhibitory-to-excitatory connections are both strong.

Dale’s principle is thus crucial in the recurrent network. This stays in contrast to the feed-forward case in which separation of excitation and inhibition did not change the networks behavior. This suggests that machine learning algorithms that make use of recurrent networks may benefit from having separate excitatory and inhibitory regions.

7.7 Variable *f0* harmonicity detection

The harmonic selectivity gets much more nonlinear when the goal is to select for *all* harmonic sounds rather than just a single one. This corresponds to the perceptual question of “how much of a *pitch* does this sound have?”. A max-of-sieves was found to be the best solution to this problem (with about 80% accuracy), which can be approximately implemented with a winner-take-all attractor recurrent network.

7.8 Model-free stimulus design with sound textures

In many cases there are too many possible models of the cortical neurons under study to be able to find the best model and parameters. In these cases a simple criteria could be individual or population response of neurons.

The space of sound stimuli is the space of possible waveforms. This gets mapped down into the space of perception, which can be defined by neural activity or by constructing a manifold by using just-noticeable-differences as a distance metric. The mapping is highly non-linear: the vast majority of the space of possible sounds gets mapped to “white noise” and thus fails to sample the perceptual space.

Random sound textures provide one of the best currently known ways to generate a wide variety of perceptual constructs, however the algorithm to generate sounds form the texture involves nonlinear optimization and is slow enough to make online experimental design difficult. Fortunately, sounds can be generated from envelops by filling-in a randomized band-appropriate fine-structure; converting from natural sound to envelope bank and back to natural sound was found to alter perception very little. This enabled solving for the envelopes instead of the waveforms. As the envelopes can be down-sampled about 10 times we can improve our algorithm’s speed by about 8 fold (the waveform would otherwise need a full-length Hilbert transform for each envelope for each iteration). This sped up the algorithm to ~3 minutes per core and allowed a successful demonstration of firing-rate optimization.

7.9 Model-based stimulus design

Some experiments have a fairly narrow set of possible models for the neuron they are testing. This means that stimuli can be designed to maximally constrain the parameter set, which in turn depends on how constrained the parameter set is in the first place. Future work is needed to assess the effectiveness of this approach for probing harmonically selectivity networks.

REFERENCES

Abeles, M., & Goldstein, M. H. (1972). Responses of single units in the primary auditory cortex of the cat to tones and to tone pairs. *Brain Research*.

Agamaite, J. A., Chang, C.-J., Osmanski, M. S., & Wang, X. (2015). A quantitative acoustic analysis of the vocal repertoire of the common marmoset (Callithrix jacchus). *The Journal of the Acoustical Society of America*, *138*(5), 2906–2928.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *ArXiv Preprint ArXiv:1803.08375*.

Agnati, L. F., Zoli, M., Strömberg, I., & Fuxe, K. (1995). Intercellular communication in the brain: Wiring versus volume transmission. *Neuroscience*, *69*(3), 711–726.

Agostinelli, F., Hoffman, M., Sadowski, P., & Baldi, P. (2014). Learning activation functions to improve deep neural networks. *ArXiv Preprint ArXiv:1412.6830*.

Aitkin, L., & Park, V. (1993). Audition and the auditory pathway of a vocal New World primate, the common marmoset. *Progress in Neurobiology*, *41*(3), 345–367.

Amit, D. J., & Amit, D. J. (1992). *Modeling brain function: The world of attractor neural networks*. Cambridge university press.

Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, *88*(2), 680–697.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*(3), 241–253.

Barlow, H. B. & others. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, *1*(01).

Barzelay, O., Furst, M., & Barak, O. (2017). A new approach to model pitch perception using sparse coding. *PLoS Computational Biology*, *13*(1), e1005338.

Bates, M. E., Simmons, J. A., & Zorikov, T. V. (2011). Bats use echo harmonic structure to distinguish their targets from background clutter. *Science*, *333*(6042), 627–630.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Bendor, D. (2015). The role of inhibition in a computational model of an auditory cortical neuron during the encoding of temporal information. *PLOS Comput Biol*, *11*(4), e1004197.

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

Bendor, D., & Wang, X. (2010). Neural coding of periodicity in marmoset auditory cortex. *Journal of Neurophysiology*, *103*(4), 1809–1822.

Bernstein, J. G., & Oxenham, A. J. (2003). Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? *The Journal of the Acoustical Society of America*, *113*(6), 3323–3334.

Bernstein, J. G., & Oxenham, A. J. (2008). Harmonic segregation through mistuning can improve fundamental frequency discrimination. *The Journal of the Acoustical Society of America*, *124*(3), 1653–1667.

Besser, G. (1967). Auditory flutter fusion as a measure of the actions of centrally acting drugs: Modification of the threshold for fusion and the influence of adapting stimuli. *British Journal of Pharmacology and Chemotherapy*, *30*(2), 329.

Bezerra, B. M., & Souto, A. (2008). Structure and usage of the vocal repertoire of Callithrix jacchus. *International Journal of Primatology*, *29*(3), 671.

Bidelman, G. M., & Khaja, A. S. (2014). Spectrotemporal resolution tradeoff in auditory processing as revealed by human auditory brainstem responses and psychophysical indices. *Neuroscience Letters*, *572*, 53–57.

Bieser, A., & Müller-Preuss, P. (1996). Auditory responsive cortex in the squirrel monkey: Neural responses to amplitude-modulated sounds. *Experimental Brain Research*, *108*(2), 273–284.

Blättler, F., & Hahnloser, R. H. (2011). An efficient coding hypothesis links sparsity and selectivity of neural responses. *PloS One*, *6*(10), e25506.

Bloomfield, P. (2004). *Fourier analysis of time series: An introduction*. John Wiley & Sons.

Boer, E. de. (1956). Pitch of inharmonic signals. *Nature*, *178*(4532), 535–536.

Borra, T., Versnel, H., Kemner, C., van Opstal, A. J., & van Ee, R. (2013). Octave effect in auditory attention. *Proceedings of the National Academy of Sciences*, *110*(38), 15225–15230.

Bota, M., & Swanson, L. W. (2007). The neuron classification problem. *Brain Research Reviews*, *56*(1), 79–88.

Braus, I. (1995). Retracing one's steps: An overview of pitch circularity and Shepard tones in european music, 1550–1990. *Music Perception*, *12*(3), 323–351.

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

Brito, C. S., & Gerstner, W. (2016). Nonlinear Hebbian learning as a unifying principle in receptive field formation. *PLoS Computational Biology*, *12*(9), e1005070.

Brosch, M., Budinger, E., & Scheich, H. (2013). Different synchronization rules in primary and nonprimary auditory cortex of monkeys. *Journal of Cognitive Neuroscience*, *25*(9), 1517–1526.

Brosch, M., & Schreiner, C. E. (2000). Sequence sensitivity of neurons in cat primary auditory cortex. *Cerebral Cortex*, *10*(12), 1155–1167.

Brosch, M., Schulz, A., & Scheich, H. (1999). Processing of sound sequences in macaque auditory cortex: Response enhancement. *Journal of Neurophysiology*, *82*(3), 1542–1559.

Brunstrom, J. M., & Roberts, B. (2001). Effects of asynchrony and ear of presentation on the pitch of mistuned partials in harmonic and frequency-shifted complex tones. *The Journal of the Acoustical Society of America*, *110*(1), 391–401.

Bürck, M., & van Hemmen, J. L. (2009). Neuronal identification of signal periodicity by balanced inhibition. *Biological Cybernetics*, *100*(4), 261–270.

Campbell, M., Greated, C. A., Myers, A., & others. (2004). *Musical instruments: History, technology, and performance of instruments of western music*. Oxford University Press on Demand.

Cannon, S. C., & Robinson, D. A. (1985). An improved neural-network model for the neural integrator of the oculomotor system: More realistic neuron behavior. *Biological Cybernetics*, *53*(2), 93–108.

Caporale, N., & Dan, Y. (2008). Spike timing–dependent plasticity: A Hebbian learning rule. *Annu. Rev. Neurosci.*, *31*, 25–46.

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

Carlin, M. A., & Elhilali, M. (2013). Sustained firing of model central auditory neurons yields a discriminative spectro-temporal representation for natural sounds. *PLoS Comput Biol*, *9*(3), e1002982.

Carlson, N. L., Ming, V. L., & DeWeese, M. R. (2012). Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput Biol*, *8*(7), e1002594.

Carlyon, R. P. (1996). The effect of mean rate cues on the pitch of filtered pulse trains. *The Journal of the Acoustical Society of America*, *99*(4), 2489–2500.

Carlyon, R. P., & Shackleton, T. M. (1994). Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? *The Journal of the Acoustical Society of America*, *95*(6), 3541–3554.

Chambers, C., & Pressnitzer, D. (2014). Perceptual hysteresis in the judgment of auditory pitch shift. *Attention, Perception, & Psychophysics*, *76*(5), 1271–1279.

Chimoto, S., Kitama, T., Qin, L., Sakayori, S., & Sato, Y. (2002). Tonal response patterns of primary auditory cortex neurons in alert cats. *Brain Research*, *934*(1), 34–42.

Clarkson, M. G., & Rogers, E. C. (1995). Infants require low-frequency energy to hear the pitch of the missing fundamental. *The Journal of the Acoustical Society of America*, *98*(1), 148–154.

Cohen, M. A., & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, *5*, 815–826.

Cohen, M. A., Grossberg, S., & Wyse, L. L. (1995). A spectral network model of pitch perception. *The Journal of the Acoustical Society of America*, *98*(2), 862–879.

Comon, P., & Jutten, C. (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Cox, I. J., Miller, M. L., Minka, T. P., Papathomas, T. V., & Yianilos, P. N. (2000). The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, *9*(1), 20–37.

Cynx, J., & Shapiro, M. (1986). Perception of missing fundamental by a species of songbird (Sturnus vulgaris). *Journal of Comparative Psychology*, *100*(4), 356.

Darwin, C., & Carlyon, R. (1995). *Auditory grouping, i in hearing. Handbook of perception and cognition, bcj moore*. San Diego: Academic Press.

Day, M. L., & Delgutte, B. (2016). Neural population encoding and decoding of sound source location across sound level in the rabbit inferior colliculus. *Journal of Neurophysiology*, *115*(1), 193–207.

De Boer, E. (1976). On the “residue” and auditory pitch perception. In *Auditory System* (pp. 479–583). Springer.

De Cheveigné, A. (1998). Cancellation model of pitch perception. *The Journal of the Acoustical Society of America*, *103*(3), 1261–1271.

De Cheveigne, A. (2005). Pitch perception models. In *Pitch* (pp. 169–233). Springer.

de Cheveigné, A., McAdams, S., Laroche, J., & Rosenberg, M. (1995). Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement. *The Journal of the Acoustical Society of America*, *97*(6), 3736–3748.

Decharms, R. C., & Merzenich, M. M. (1996). Primary cortical representation of sounds by the coordination of action-potential timing. *Nature*, *381*(6583), 610–613.

Deutsch, D., & Boulanger, R. C. (1984). Octave equivalence and the immediate recall of pitch sequences. *Music Perception*, *2*(1), 40–51.

DiMattina, C., & Wang, X. (2006). Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. *Journal of Neurophysiology*, *95*(2), 1244–1262.

DiMattina, C., & Zhang, K. (2011). Active data collection for efficient estimation and comparison of nonlinear neural models. *Neural Computation*, *23*(9), 2242–2288.

Doruk, R. O., & Zhang, K. (2019). Adaptive stimulus design for dynamic recurrent neural network models. *Frontiers in Neural Circuits*, *12*, 119.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K., & Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*(5226), 981–985.

Duifhuis, H., Willems, L. F., & Sluyter, R. (1982). Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *The Journal of the Acoustical Society of America*, *71*(6), 1568–1580.

Eccles, J. C., Fatt, P., & Koketsu, K. (1954). Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurones. *The Journal of Physiology*, *126*(3), 524–562.

Ehret, G., & Merzenich, M. (1988). Neuronal discharge rate is unsuitable for encoding sound intensity at the inferior-colliculus level. *Hearing Research*, *35*(1), 1–7.

Epple, G. (1968). Comparative studies on vocalization in marmoset monkeys (Hapalidae). *Folia Primatologica*, *8*(1), 1–40.

Evans, E., & Whitfield, I. (1964). Classification of unit responses in the auditory cortex of the unanaesthetized and unrestrained cat. *The Journal of Physiology*, *171*(3), 476–493.

Feng, A. S., Narins, P. M., Xu, C.-H., Lin, W.-Y., Yu, Z.-L., Qiu, Q., Xu, Z.-M., & Shen, J.-X. (2006). Ultrasonic communication in frogs. *Nature*, *440*(7082), 333–336.

Feng, L., & Wang, X. (2017). Harmonic template neurons in primate auditory cortex underlying complex sound processing. *Proceedings of the National Academy of Sciences*, *114*(5), E840–E848.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6*(4), 559–601.

Fishman, Y. I., Micheyl, C., & Steinschneider, M. (2013). Neural representation of harmonic complex tones in primary auditory cortex of the awake monkey. *Journal of Neuroscience*, *33*(25), 10312–10323.

Fishman, Y. I., Reser, D. H., Arezzo, J. C., & Steinschneider, M. (1998). Pitch vs. Spectral encoding of harmonic complex tones in primary auditory cortex of the awake monkey. *Brain Research*, *786*(1–2), 18–30.

Fishman, Y. I., & Steinschneider, M. (2010). Neural correlates of auditory scene analysis based on inharmonicity in monkey primary auditory cortex. *Journal of Neuroscience*, *30*(37), 12480–12494.

Fitzpatrick, D. C., Kanwal, J. S., Butman, J. A., & Suga, N. (1993). Combination-sensitive neurons in the primary auditory cortex of the mustached bat. *Journal of Neuroscience*, *13*(3), 931–940.

Gamble, D. W. & others. (2020). *Physiological Characterization Of Parabelt Auditory Cortex In The Awake, Behaving Marmoset Monkey* [PhD Thesis]. Johns Hopkins University.

Gao, L., Kostlan, K., Wang, Y., & Wang, X. (2016). Distinct subthreshold mechanisms underlying rate-coding principles in primate auditory cortex. *Neuron*, *91*(4), 905–919.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, *3*(Aug), 115–143.

Gilbert, C. D. (1998). Adult cortical dynamics. *Physiological Reviews*, *78*(2), 467–485.

Gilbert, C. D., & Wiesel, T. N. (1979). Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. *Nature*, *280*(5718), 120–125.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315–323.

Gold, T. (1948). Hearing. II. The physical basis of the action of the cochlea. *Proceedings of the Royal Society of London. Series B-Biological Sciences*, *135*(881), 492–498.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *The Journal of the Acoustical Society of America*, *54*(6), 1496–1516.

Gutnick, M. J., & Mody, I. (Eds.). (1995). *The Cortical Neuron*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195083309.001.0001

Hartmann, W. M., McAdams, S., & Smith, B. K. (1990). Hearing a mistuned harmonic in an otherwise periodic complex tone. *The Journal of the Acoustical Society of America*, *88*(4), 1712–1724.

Heffner, H., & Whitfield, I. C. (1976). Perception of the missing fundamental by cats. *The Journal of the Acoustical Society of America*, *59*(4), 915–919.

Hefner, H., & Heffner, R. S. (1986). Effect of unilateral and bilateral auditory cortex lesions on the discrimination of vocalizations by Japanese macaques. *Journal of Neurophysiology*, *56*(3), 683–701.

Hengen, K. B., Lambo, M. E., Van Hooser, S. D., Katz, D. B., & Turrigiano, G. G. (2013). Firing rate homeostasis in visual cortex of freely behaving rodents. *Neuron*, *80*(2), 335–342.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Himberg, J., Hyvärinen, A., & Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage*, *22*(3), 1214–1222.

Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational abilities. *Proc. Natl. Acad. Sci. USA*, *79*, 2554–2558.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*(10), 3088–3092.

Hopfield, J. J., & Tank, D. W. (1985). "Neural" computation of decisions in optimization problems. *Biological Cybernetics*, *52*(3), 141–152.

Horn, B. K., Hilden, H. M., & Negahdaripour, S. (1988). Closed-form solution of absolute orientation using orthonormal matrices. *JOSA A*, *5*(7), 1127–1135.

Houtgast, T., & Steeneken, H. Jm. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acta Acustica United with Acustica*, *28*(1), 66–73.

Houtsma, A. J., & Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *The Journal of the Acoustical Society of America*, *87*(1), 304–310.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., & Xie, L. (2020). Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. *ArXiv Preprint ArXiv:2008.00264*.

Huang, C., & Rinzel, J. (2016). A neuronal network model for pitch selectivity and representation. *Frontiers in Computational Neuroscience*, *10*, 57.

Huang, W., Liu, K., & Zhang, K. (2017). Unsupervised learning via maximizing mutual information in neural population coding. *2017 AAAI Spring Symposium*, 575–579.

Hubel, D. H., Henson, C. O., Rupert, A., & Galambos, R. (1959). " Attention" units in the auditory cortex. *Science*, *129*(3358), 1279–1280.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634.

Hyvärinen, A., & Oja, E. (1997). One-unit learning rules for independent component analysis. *Advances in Neural Information Processing Systems*, 480–486.

Jackson, H. M., & Moore, B. C. (2013). The dominant region for the pitch of complex tones with low fundamental frequencies. *The Journal of the Acoustical Society of America*, *134*(2), 1193–1204.

Jane, J. Y., & Young, E. D. (2000). Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus. *Proceedings of the National Academy of Sciences*, *97*(22), 11780–11786.

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences*, *97*(22), 11793–11799.

Kadia, S. C., & Wang, X. (2003). Spectral integration in A1 of awake primates: Neurons with single-and multipeaked tuning characteristics. *Journal of Neurophysiology*, *89*(3), 1603–1622.

Kadia, S., Liang, L., Wang, X., Doucet, J., & Ryugo, D. (1999). Horizontal connections within the primary auditory cortex of cat. *Assoc. Res. Otolaryngol. Abstr*, *22*, 34.

Kalluri, S., Depireux, D. A., & Shamma, S. A. (2008). Perception and cortical neural coding of harmonic fusion in ferrets. *The Journal of the Acoustical Society of America*, *123*(5), 2701–2716.

Kanwal, J. S., Fitzpatrick, D. C., & Suga, N. (1999). Facilitatory and inhibitory frequency tuning of combination-sensitive neurons in the primary auditory cortex of mustached bats. *Journal of Neurophysiology*, *82*(5), 2327–2345.

Katsiamis, A. G., Drakakis, E. M., & Lyon, R. F. (2007). Practical gammatone-like filters for auditory processing. *EURASIP Journal on Audio, Speech, and Music Processing*, *2007*, 1–15.

Kaur, S., Lazar, R., & Metherate, R. (2004). Intracortical pathways determine breadth of subthreshold frequency receptive fields in primary auditory cortex. *Journal of Neurophysiology*, *91*(6), 2551–2567.

Khalil, H. K., & Grizzle, J. W. (2002). *Nonlinear systems* (Vol. 3). Prentice hall Upper Saddle River, NJ.

Kiang, N., Sachs, M. B., & Peake, W. (1967). Shapes of tuning curves for single auditory-nerve fibers. *The Journal of the Acoustical Society of America*, *42*(6), 1341–1342.

Klein, D. J., König, P., & Körding, K. P. (2003). Sparse spectrotemporal coding of sounds. *EURASIP Journal on Advances in Signal Processing*, *2003*(7), 1–9.

Knierim, J. J., & Zhang, K. (2012). Attractor dynamics of spatially correlated neural activity in the limbic system. *Annual Review of Neuroscience*, *35*, 267–285.

Kostlan, K. J. (2015). *Responses to harmonic and mistuned complexes in the awake marmoset inferior colliculus* [Masters Thesis]. Johns Hopkins University.

Kozlov, A. S., & Gentner, T. Q. (2016). Central auditory neurons have composite receptive fields. *Proceedings of the National Academy of Sciences*, *113*(5), 1441–1446.

Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology*, *11*(3), 346–374.

Kudoh, M., Nakayama, Y., Hishida, R., & Shibuki, K. (2006). Requirement of the auditory association cortex for discrimination of vowel-like sounds in rats. *Neuroreport*, *17*(17), 1761–1766.

Kurt, S., Deutscher, A., Crook, J. M., Ohl, F. W., Budinger, E., Moeller, C. K., Scheich, H., & Schulze, H. (2008). Auditory cortical contrast enhancing by global winner-take-all inhibitory interactions. *PLoS One*, *3*(3), e1735.

Langner, G., Albert, M., & Briede, T. (2002). Temporal and spatial coding of periodicity information in the inferior colliculus of awake chinchilla (Chinchilla laniger). *Hearing Research*, *168*(1–2), 110–130.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363.

Lewicki, M. S., & Konishi, M. (1995). Mechanisms underlying the sensitivity of songbird forebrain neurons to temporal order. *Proceedings of the National Academy of Sciences*, *92*(12), 5582–5586.

Licklider, J. C. R. (1956). Audio frequency analysis. *Information Theory*, 253–268.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*(3), 105–117.

Little, W. A. (1974). The existence of persistent states in the brain. *Math. Biosci.*, *19*, 101–120.

Makhoul, J. (1981). On the eigenvectors of symmetric Toeplitz matrices. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(4), 868–872.

Malmberg, C. F. (1918). The perception of consonance and dissonance. *Psychological Monographs*, *25*(2), 93.

Margoliash, D. (1983). Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *Journal of Neuroscience*, *3*(5), 1039–1057.

Markram, H., & Tsodyks, M. (1996). Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature*, *382*(6594), 807–810.

Marsh, R. A., Nataraj, K., Gans, D., Portfors, C. V., & Wenstrup, J. J. (2006). Auditory responses in the cochlear nucleus of awake mustached bats: Precursors to spectral integration in the auditory midbrain. *Journal of Neurophysiology*, *95*(1), 88–105.

Maskos, U., Kissa, K., Cloment, C. S., & Brûlet, P. (2002). Retrograde trans-synaptic transfer of green fluorescent protein allows the genetic mapping of neuronal circuits in transgenic mice. *Proceedings of the National Academy of Sciences*, *99*(15), 10120–10125.

Matsubara, J. A., & Phillips, D. (1988). Intracortical connections and their physiological correlates in the primary auditory cortex (AI) of the cat. *Journal of Comparative Neurology*, *268*(1), 38–48.

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2010). Individual differences reveal the basis of consonance. *Current Biology*, *20*(11), 1035–1041.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*(5), 926–940.

McGuire, B. A., Gilbert, C. D., Rivlin, P. K., & Wiesel, T. N. (1991). Targets of horizontal connections in macaque primary visual cortex. *Journal of Comparative Neurology*, *305*(3), 370–392.

McLachlan, N. (2011). A neurocognitive model of recognition and pitch segregation. *The Journal of the Acoustical Society of America*, *130*(5), 2845–2854.

Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *The Journal of the Acoustical Society of America*, *89*(6), 2866–2882.

Mlynarski, W., & McDermott, J. (2017). Lossy compression of uninformative stimuli in the auditory system. *The Journal of the Acoustical Society of America*, *141*(5), 3897–3897.

Moeller, C. K., Kurt, S., Happel, M. F., & Schulze, H. (2010). Long-range effects of GABAergic inhibition in gerbil primary auditory cortex. *European Journal of Neuroscience*, *31*(1), 49–59.

Moerel, M., De Martino, F., Santoro, R., Yacoub, E., & Formisano, E. (2015). Representation of pitch chroma by multi-peak spectral tuning in human auditory cortex. *Neuroimage*, *106*, 161–169.

Moore, B. C., Glasberg, B. R., & Peters, R. W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *The Journal of the Acoustical Society of America*, *80*(2), 479–483.

Moore, B. C., Peters, R. W., & Glasberg, B. R. (1985). Thresholds for the detection of inharmonicity in complex tones. *The Journal of the Acoustical Society of America*, *77*(5), 1861–1867.

Ngo, K. B., Mahony, R., & Jiang, Z.-P. (2005). Integrator backstepping using barrier functions for systems with multiple state constraints. *Proceedings of the 44th IEEE Conference on Decision and Control*, 8306–8312.

Noreña, A. J., Gourévitch, B., Pienkowski, M., Shaw, G., & Eggermont, J. J. (2008). Increasing spectrotemporal sound density reveals an octave-based organization in cat primary auditory cortex. *Journal of Neuroscience*, *28*(36), 8885–8896.

Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *Journal of Neuroscience*, *33*(50), 19451–19469.

Oertel, D., Wu, S. H., Garb, M. W., & Dizack, C. (1990). Morphology and physiology of cells in slice preparations of the posteroventral cochlear nucleus of mice. *Journal of Comparative Neurology*, *295*(1), 136–154.

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*, *5*(6), 927–935.

Ojima, H., Honda, C. N., & Jones, E. (1991). Patterns of axon collateralization of identified supragranular pyramidal neurons in the cat auditory cortex. *Cerebral Cortex*, *1*(1), 80–94.

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

Osmanski, M. S., Song, X., & Wang, X. (2013). The role of harmonic resolvability in pitch perception in a vocal nonhuman primate, the common marmoset (Callithrix jacchus). *Journal of Neuroscience*, *33*(21), 9161–9168.

Osmanski, M. S., & Wang, X. (2011). Measurement of absolute auditory thresholds in the common marmoset (Callithrix jacchus). *Hearing Research*, *277*(1–2), 127–133.

Patterson, R. D., Allerhand, M. H., & Giguere, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, *98*(4), 1890–1894.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, *36*(4), 767–776.

Penagos, H., Melcher, J. R., & Oxenham, A. J. (2004). A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *Journal of Neuroscience*, *24*(30), 6810–6815.

Peretz, I., & Zatorre, R. J. (2005). Brain organization for music processing. *Annu. Rev. Psychol.*, *56*, 89–114.

Phillips, D., & Irvine, D. (1981). Responses of single neurons in physiologically defined primary auditory cortex (AI) of the cat: Frequency tuning and responses to intensity. *Journal of Neurophysiology*, *45*(1), 48–58.

Pickles, J. (2013). *An introduction to the physiology of hearing*. Brill.

Pickles, J. O. (2015). Auditory pathways: Anatomy and physiology. *Handbook of Clinical Neurology*, *129*, 3–25.

Pistorio, A. L., Vintch, B., & Wang, X. (2006). Acoustic analysis of vocal development in a New World primate, the common marmoset (Callithrix jacchus). *The Journal of the Acoustical Society of America*, *120*(3), 1655–1670.

Popham, S., Boebinger, D., Ellis, D. P., Kawahara, H., & McDermott, J. H. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature Communications*, *9*(1), 1–13.

Portfors, C. V., & Wenstrup, J. J. (2002). Excitatory and facilitatory frequency response areas in the inferior colliculus of the mustached bat. *Hearing Research*, *168*(1–2), 131–138.

Pressnitzer, D., De Cheveigné, A., & Winter, I. M. (2002). Perceptual pitch shift for sounds with similar waveform autocorrelation. *Acoustics Research Letters Online*, *3*(1), 1–6.

Priebe, N. J., Mechler, F., Carandini, M., & Ferster, D. (2004). The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature Neuroscience*, *7*(10), 1113–1122.

Qin, L., Sakai, M., Chimoto, S., & Sato, Y. (2005). Interaction of excitatory and inhibitory frequency-receptive fields in determining fundamental frequency sensitivity of primary auditory cortex neurons in awake cats. *Cerebral Cortex*, *15*(9), 1371–1383.

Rauschecker, J. P., Tian, B., Pons, T., & Mishkin, M. (1997). Serial and parallel processing in rhesus monkey auditory cortex. *Journal of Comparative Neurology*, *382*(1), 89–103.

Reale, R. A., Brugge, J. F., & Feng, J. Z. (1983). Geometry and orientation of neuronal processes in cat primary auditory cortex (AI) related to characteristic-frequency maps. *Proceedings of the National Academy of Sciences*, *80*(17), 5449–5453.

Rhode, W. S. (1995). Interspike intervals as a correlate of periodicity pitch in cat cochlear nucleus. *The Journal of the Acoustical Society of America*, *97*(4), 2414–2429.

Rhode, W. S., Roth, G. L., & Recio-Spinoso, A. (2010). Response properties of cochlear nucleus neurons in monkeys. *Hearing Research*, *259*(1–2), 1–15.

Robles, L., Ruggero, M. A., & Rich, N. C. (1991). Two-tone distortion in the basilar membrane of the cochlea. *Nature*, *349*(6308), 413.

Rogers, E. C., Clarkson, M. G., & Miciek, S. G. (1996). Infants' pitch perception: Masking by low-and high-frequency noises. *Infant Behavior and Development*, *19*, 708.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *336*(1278), 367–373.

Sadagopan, S., & Wang, X. (2008). Level invariant representation of sounds by populations of neurons in primary auditory cortex. *Journal of Neuroscience*, *28*(13), 3415–3426.

Sadagopan, S., & Wang, X. (2009). Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. *Journal of Neuroscience*, *29*(36), 11192–11202.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, *2*(6), 459–473.

Schnupp, J., Nelken, I., & King, A. (2011). *Auditory neuroscience: Making sense of sound*. MIT press.

Schonfeld, D. (1993). On the hysteresis and robustness of Hopfield neural networks. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, *40*(11), 745–748.

Schouten, J. (1968). The perception of pitch. *IPO Annual Progress Report*, *3*, 32–34.

Schouten, J. F., Ritsma, R., & Cardozo, B. L. (1962). Pitch of the residue. *The Journal of the Acoustical Society of America*, *34*(9B), 1418–1424.

Schreiner, C. E., & Winer, J. A. (2005). *The inferior colliculus*. Springer.

Schwartz, A., McDermott, J. H., & Shinn-Cunningham, B. (2012). Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *The Journal of the Acoustical Society of America*, *132*(1), 357–368.

Schwarz, D. W., & Tomlinson, R. W. (1990). Spectral response patterns of auditory cortex neurons to harmonic complex tones in alert monkey (Macaca mulatta). *Journal of Neurophysiology*, *64*(1), 282–298.

Shamma, S. A., Chadwick, R. S., Wilbur, W. J., Morrish, K. A., & Rinzel, J. (1986). A biophysical model of cochlear processing: Intensity dependence of pure tone responses. *The Journal of the Acoustical Society of America*, *80*(1), 133–145.

Shamma, S. A., & Morrish, K. A. (1987). Synchrony suppression in complex stimulus responses of a biophysical model of the cochlea. *The Journal of the Acoustical Society of America*, *81*(5), 1486–1498.

Shamma, S., & Klein, D. (2000). The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *The Journal of the Acoustical Society of America*, *107*(5), 2631–2644.

Shannon, R. V. (2016). Is birdsong more like speech or music? *Trends in Cognitive Sciences*, *20*(4), 245–247.

Sheikh, A.-S., Harper, N. S., Drefs, J., Singer, Y., Dai, Z., Turner, R. E., & Lücke, J. (2019). STRFs in primary auditory cortex emerge from masking-based statistics of natural sounds. *PLoS Computational Biology*, *15*(1), e1006595.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.

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

Slee, S. J., & Young, E. D. (2010). Sound localization cues in the marmoset monkey. *Hearing Research*, *260*(1–2), 96–108.

Slee, S. J., & Young, E. D. (2013). Linear processing of interaural level difference underlies spatial tuning in the nucleus of the brachium of the inferior colliculus. *Journal of Neuroscience*, *33*(9), 3891–3904.

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, *439*(7079), 978–982.

Song, X., Osmanski, M. S., Guo, Y., & Wang, X. (2016). Complex pitch perception mechanisms are shared by humans and a New World monkey. *Proceedings of the National Academy of Sciences*, *113*(3), 781–786.

Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. (2020). Fully convolutional recurrent networks for speech enhancement. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6674–6678.

Strata, P., & Harvey, R. (1999). Dale’s principle. *Brain Research Bulletin*, *50*(5–6), 349–350.

Suga, N., O’NeilL, W. E., Kujirai, K., & Manabe, T. (1983). Specialization of “combination-sensitive” neurons for processing of complex biosonar signals in the auditory cortex of the mustached bat. *Journal of Neurophysiology*, *49*(1), 1573–1626.

Suga, N., O’Neill, W. E., & Manabe, T. (1979). Harmonic-sensitive neurons in the auditory cortex of the mustache bat. *Science*, *203*(4377), 270–274.

Sutter, M. L., & Schreiner, C. E. (1991). Physiology and topography of neurons with multipeaked tuning curves in cat primary auditory cortex. *Journal of Neurophysiology*, *65*(5), 1207–1226.

Sutter, M., Schreiner, C., McLean, M., O’connor, K., & Loftus, W. (1999). Organization of inhibitory frequency receptive fields in cat primary auditory cortex. *Journal of Neurophysiology*, *82*(5), 2358–2371.

Tee, K. P., Ge, S. S., & Tay, E. H. (2009). Barrier Lyapunov functions for the control of output-constrained nonlinear systems. *Automatica*, *45*(4), 918–927.

Terashima, H., & Hosoya, H. (2009). Sparse codes of harmonic natural sounds and their modulatory interactions. *Network: Computation in Neural Systems*, *20*(4), 253–267.

Terhardt, E. (1974). Pitch, consonance, and harmony. *The Journal of the Acoustical Society of America*, *55*(5), 1061–1069.

Tomlinson, R. W., & Schwarz, D. W. (1988). Perception of the missing fundamental in nonhuman primates. *The Journal of the Acoustical Society of America*, *84*(2), 560–565.

Van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*(1394), 359–366.

Villa, A. E. (1990). Physiological differentiation within the auditory part of the thalamic reticular nucleus of the cat. *Brain Research Reviews*, *15*(1), 25–40.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, *11*(12).

Von Helmholtz, H. (1912). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Longmans, Green.

Wallace, M., Kitzes, L., & Jones, E. (1991). Intrinsic inter-and intralaminar connections and their relationship to the tonotopic map in cat primary auditory cortex. *Experimental Brain Research*, *86*(3), 527–544.

Wallace, M. N., Rutkowski, R. G., & Palmer, A. R. (2002). Interconnections of auditory areas in the guinea pig neocortex. *Experimental Brain Research*, *143*(1), 106–119.

Wang, X. (2013). The harmonic organization of auditory cortex. *Frontiers in Systems Neuroscience*, *7*, 114.

Wang, X., Lu, T., Snider, R. K., & Liang, L. (2005). Sustained firing in auditory cortex evoked by preferred stimuli. *Nature*, *435*(7040), 341–346.

Wang, X., Merzenich, M. M., Beitel, R., & Schreiner, C. E. (1995). Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: Temporal and spectral characteristics. *Journal of Neurophysiology*, *74*(6), 2685–2706.

Wehr, M., & Zador, A. M. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature*, *426*(6965), 442–446.

Whitfield, I. (1980). Auditory cortex and the pitch of complex tones. *The Journal of the Acoustical Society of America*, *67*(2), 644–647.

Wightman, F. L. (1973). The pattern-transformation model of pitch. *The Journal of the Acoustical Society of America*, *54*(2), 407–416.

Winer, J. A. (1992). The functional architecture of the medial geniculate body and the primary auditory cortex. In *The mammalian auditory pathway: Neuroanatomy* (pp. 222–409). Springer.

Winter, I., & Palmer, A. R. (1995). Level dependence of cochlear nucleus onset unit responses and facilitation by second tones or broadband noise. *Journal of Neurophysiology*, *73*(1), 141–159.

Yang, X., Wang, K., & Shamma, S. A. (1992). Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, *38*(2), 824–839.

Yarden, T. S., & Nelken, I. (2017). Stimulus-specific adaptation in a recurrent network model of primary auditory cortex. *PLoS Computational Biology*, *13*(3), e1005437.

Zatorre, R. J. (1988). Pitch perception of complex tones and human temporal-lobe function. *The Journal of the Acoustical Society of America*, *84*(2), 566–572.

Zhang, X., Heinz, M. G., Bruce, I. C., & Carney, L. H. (2001). A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *The Journal of the Acoustical Society of America*, *109*(2), 648–670.

Zhao, L., & Zhaoping, L. (2011). Understanding auditory spectro-temporal receptive fields and their changes with input statistics by efficient coding principles. *PLoS Comput Biol*, *7*(8), e1002123.

Zilany, M. S., Bruce, I. C., Ibrahim, R. A., & Carney, L. H. (2013). Improved parameters and expanded simulation options for a model of the auditory periphery. *Th ARO Midwinter Research Meeting. Association for Research in Otolaryngology. Baltimore, MD, Pgs*, 440–441.

Zoli, M., Jansson, A., Syková, E., Agnati, L. F., & Fuxe, K. (1999). Volume transmission in the CNS and its relevance for neuropsychopharmacology. *Trends in Pharmacological Sciences*, *20*(4), 142–150.

Curriculum Vitae

Kevin Kostlan

kkostla1@jh.edu

Education/Training*:*

Institution	Degree	Start Date	End Date	Field of Study
University of California, Davis	Bachelors	09/2008	06/2012	Bio Engineering
Johns Hopkins University	Masters	09/2012	09/2014	Biomedical Engineering
Johns Hopkins University	PhD	09/2014	In progress	Biomedical Engineering

Personal Statement

No, a triple major in computer science, biology, and physics wouldn't work. But bioengineering is the closest thing, given that computers are ubiquitous and engineering means physics. And it will solve huge practical problems! However, very little of "bioengineering" is actually biological engineering. It is mostly split between basic experimental biology and basic computational biology.

Always seeking the balance, my masters in Hopkins was a mixture of experiments and computation. The computer work itself gave me my fill of engineering and the experiments (single-unit recordings in the marmoset inferior colliculus) brought in the biology. I saw the art behind electrode-placement, subtleties of how the units adapt, the various sources of noise. Indeed, reality has countless twists and turns that are sorely missing from the land of code. However, the work was too model-free and there wasn't much room to design better systems. Dismayed by the lack of tinkering, I moved to the computational side for my PhD.

My current focus is combining powerful theoretical tools with pre-existing experimental data to model the auditory system's response to harmonic sounds. My past "gloves-on time" helps me identify and correct for artifacts in the data analysis that may be missed given a purely computational background. Working with Hopkins people who have a strong theoretical background provides insurance against getting stuck with the equations. This combination is needed for a strong PhD.

I plan on eventually having a mixture of experimental work and computational work. Automation, such as Transcriptic's cloud lab, is an emerging force amplifier that will vastly increase the amount of experimental design as well as data analysis. This will allow people to solve wonderfully difficult problems in both the *in-silico* and *in-vitro* worlds.

A. **Positions and Honors**

Received the Robert Roy Owen Scholarship upon entrance to UC Davis.
Currently a PhD candidate funded by a NIH grant.

B. **Contributions to Science**

Coded most of the data analysis portion including automatic image normalization and particle detection of rice extracted from pig stomachs. The algorithm used MATLAB's image analysis tools such as *imdilate()* and *imopen()*:

Bornhorst, G. M., Kostlan, K., & Singh, R. P. (2013). Particle size distribution of brown and white rice during gastric digestion measured by image analysis. Journal of food science, 78(9), E1383-E1391.

Single unit recording in the inferior colliculus's central nucleus of the awake marmoset in response to harmonics/mistuned harmonics. Cells responded to energy in a single band, no multi-band integration was found. Developed a pseudo-population model that demonstrated harmonic-sensitive units could be generated by combining the units that were recorded:

Kostlan, K. J. (2014). Responses to Harmonic and Mistuned Complexes in the Awake Marmoset Inferior Colliculus (Masters thesis).

Built a single-input integrate-and-fire model that adds saturation to inhibitory and excitatory channels. This model accounted for the "negative monotonic" behavior of cortical neurons.

Gao, L., Kostlan, K., Wang, Y., & Wang, X. (2016). Distinct subthreshold mechanisms underlying rate-coding principles in primate auditory cortex. Neuron, 91(4), 905-919.

Optimized sound-from-texture generation algorithm that uses the conjugate-gradient method to minimize how much the waveform's texture disagrees with the target texture it is supposed to represent. An 8-fold speedup was produced by applying gradient descent to the (much smaller) envelope of the sound-wave and then generating the wave from an envelope. This speedup allowed online *in vivo* experimentation.

Paper in preparation

C. **Graduate Courses**

Year	Graduate Course Title (all science related and taken at Johns Hopkins University)	Grade
2012	Models of the Neuron	A
2012	Systems Bioengineering III	A-
2012	Principles of Complex Networked Sys	P
2013	Introduction to Stochastic Processes	A
2013	Systems Bioengineering II	A
2013	Theoretical Neuroscience	A
2013	Structure & Function of Auditory & Vestibular	A+
2014	Genes to Society - Nervous system	P
2014	Genes to Society - Brain mind & Behavior	P
2015	Statistical Theory	A