Tree-Structured Neural Decoding

Christian d'Avignon
Department of Biomedical Engineering
Johns Hopkins University
Baltimore, MD, USA

Donald Geman[*]
Whitaker Biomedical Institute and
Department of Mathematical Sciences
Johns Hopkins University
Baltimore, MD, USA

Abstract

We propose adaptive testing as a general mechanism for extracting information about stimuli from spike trains. Each test or question corresponds to choosing a neuron and a time interval and checking for a given number of spikes. No assumptions are made about the distribution of spikes or any other aspect of neural encoding. The chosen questions are those which most reduce the uncertainty about the stimulus, as measured by entropy and estimated from stimulus-response data. Our experiments are based on accurate simulations of responses to pure tones in the auditory nerve and are meant to illustrate the ideas rather than investigate the auditory system. The results cohere nicely with well-understood encoding of amplitude and frequency in the auditory nerve, suggesting that adaptive testing might provide a powerful tool for investigating complex and poorly understood neural structures.

1 Introduction

Sensory stimuli are transmitted to higher processing centers via neural pathways from a variety of specific receptors. Characteristics of the stimuli are encoded into the sequence of times at which a cellular event called an *action potential* (or "*spike*") occurs. An important goal in neuroscience is to better understand the nature of this encoding process at various levels of the neural hierarchy. We shall adopt the "point of view of the organism" [1]: How might a figurative being, situated within the brain and observing only neural spikes, interpret incoming but unseen stimuli, in fact provide a "running commentary" about the stimuli based only on simple functions of the spike sequence?

We consider the possibility of decoding instructions in the form of a tree of binary questions somehow stored in the brain. Whatever its specific biological support, this mechanism would presumably develop rapidly in the early stages of the organism's lifespan; it might be continuously refined, even simplified, due to perpetual learning from new presentations, thereby enhancing reliability and reducing mistakes on unseen examples ("generalization error"). Minimizing over-dedication to specific examples ("overfitting") is evidently crucial to emulating a stable biological decision-making process (or for using this technique to analyze poorly understood neural structures). The rate of refinements would clearly diminish as the organism matures and new data becomes increasingly less informative; in the

---

[*] Address correspondence to: geman@cis.jhu.edu

mature organism, this computational procedure would presumably remain the same for a given class of stimuli.

The particular questions we entertain are based on the presence or absence of spikes in selected time intervals of the recent past. No assumptions are made about the encoding process. The questions are posed sequentially; the one posed at stage k may depend on the answers to the previous $k-1$ questions. Our trees are constructed by standard inductive learning [2]: At each stage choose the question which most reduces the conditional entropy of the "class variable" - in our case the stimulus - given the new question and the previous answers. Mean entropies and answer statistics are estimated from spike data generated by simulating the conduction of auditory nerve responses to pure tones. Although we do not consider transient stimuli, we attempt to estimate the stimulus based only on observing spikes in the recent past. In order to better appreciate the experiments, the interested reader might consult a standard neuroscience source [3] or an introductory text on the auditory system [4].

2 Scenario

The stimuli are pure tones and hence characterized by an amplitude ξ, a frequency ω and a phase ϕ, restricted to intervals Ξ, Ω and Φ, respectively. Amplitude represents loudness and can be expressed in decibels (dB); frequency represents the pitch and is expressed in Hertz (Hz); phase could be expressed in radians. For human hearing, we might take $\Xi = [-10\,\text{dB}, 60\,\text{dB}]$,$\Omega = [16\,\text{Hz}, 16\,\text{kHz}]$ and $\Phi = [0, 2\pi]$. Phase is an important neurophysiological cue, and particularly crucial for sound localization; however, since it is processed at stages ulterior to the auditory nerve, we set it to zero.

Consequently, we consider zero-phase sinusoidal acoustic waves with time-independent amplitude and frequency. Let $\mathcal{S} = \{\xi \sin(\omega t) : t \in I_0, \xi \in \Xi, \omega \in \Omega\}$ be the set of possible stimuli and let $\mathcal{R} = \{\{t_i\}_{i \geq 1}\}$ be the set of all possible single-neuron responses, namely finite sequences of spike times. In order to emphasize the "running commentary," the spikes are investigated on the subinterval $I_c = [\tau - \tau_c, \tau] \subset I_0$ where I_c represents the "recent past"; hence $\tau_c > 0$ is the length of the corresponding time buffer and τ shall be understood as the present - the instant at which the figurative being generates the current commentary.

Let $(\mathbf{S}, \mathbf{R}) \in \mathcal{S} \times \mathcal{R}^N$ be a stimulus-response pair. Note that $\mathbf{R}$ is a collection of N spike trains and that $\mathbf{S}$ and $\mathbf{R}$ are regarded as random variables whose joint distribution is of course unknown. Let $\mathcal{L} = \{(s^{(1)}, r^{(1)}), ..., (s^{(C)}, r^{(C)})\}$ be the training set, assumed, as usual, to be i.i.d. realizations of $(\mathbf{S}, \mathbf{R})$. In our case the stimulus is determined by ξ and ω. Let this mapping be denoted by $\theta : \mathcal{S} \to \Xi \times \Omega$, i.e., $\theta(S) = (\xi(S), \omega(S))$. The training set can be equivalently represented as

$$\mathcal{L} = \{((\xi^{(1)}, \omega^{(1)}), r^{(1)}), ..., ((\xi^{(C)}, \omega^{(C)}), r^{(C)})\}.$$

3 Constructing the Decoder

Suppose the continuous parameter space $\Theta = \Xi \times \Omega$ is quantized into rectangular bins corresponding to partitioning amplitude and frequency: $\Xi = \bigcup_{u=1}^{K_\xi} \Xi_u$ and $\Omega = \bigcup_{v=1}^{K_\omega} \Omega_v$, where all components are intervals. In effect, we discretize Θ to $K_\xi \times K_\omega$ values, which can be thought of as the centers of rectangular bins $\Theta_{uv}, 1 \leq u \leq K_\xi, 1 \leq v \leq K_\omega$. In practice, stimuli are selected by randomly choosing samples from each bin. In our experiments we take either $\Theta \subset \{\xi_o\} \times \Omega$ or $\Theta \subset \Xi \times \{\omega_o\}$.

The questions are defined as follows. Let $\mathbf{R} = (\mathbf{R}(1), ..., \mathbf{R}(N))$ denote the responses of the N neurons. There is a binary question, $X_{nAm}(\mathbf{R})$, for each neuron n, each subinterval

$A \subset [\tau - \tau_c, \tau]$ and each non-negative integer m representing a spike count; $X_{nAm}(\mathbf{R}) = 1$ if $|\mathbf{R}(n) \cap A| = m$ and $X_{nAm}(\mathbf{R}) = 0$ otherwise. Of course we cannot consider *all* such questions as actual candidates and if the number we do consider is very large there is sure to be overfitting. Hence we restrict the intervals to

$$\mathcal{A} = \bigcup_{l=1}^{L} \{[\tau - \frac{(j+1)}{l}\tau_c, \tau - \frac{j}{l}\tau_c] : j = 0, ..., l-1\}$$

for certain L and write $\mathbf{X}_{\mathcal{A}}$ for the resulting set of questions:

$$\mathbf{X}_{\mathcal{A}} = \{X_{nAm}(\mathbf{R}) : A \in \mathcal{A}, n = 1, ..., N, m = 0, ..., M\}.$$

We could imagine other families, such as only intervals of the form $[t, \tau]$ or $[\tau - \tau_c, t]$ for certain $\tau - \tau_c < t < \tau$, but for simplicity we restrict our discussion to $\mathbf{X}_{\mathcal{A}}$.

The tree $\mathbf{T}$ is constructed top-down by the standard recursive method, now summarized. Let $\mathbf{T}$ have maximal depth D and let $\eta = (d, p)$ denote the p'th node (left to right) at depth d, $p = 1, ..., 2^{d-1}$, $d = 1, ..., D$; the root is $\eta = (1, 1)$. Every internal node η is assigned a test $X_\eta \in \mathbf{X}_{\mathcal{A}}$. Since each test is a function of $\mathbf{R}$, the tree $\mathbf{T}$ is also regarded as a random variable taking values in the set of leaves. In particular every training example $((\xi^{(i)}, \omega^{(i)}), r^{(i)})$ traverses a unique path down the tree determined only by $r^{(i)}$. Let B_η denote the event that node η is reached (the history of the $d-1$ questions along the path from $(1, 1)$ to $\eta = (d, p)$) and let $\mathcal{L}_\eta \subset \mathcal{L}$ be the set of training examples arriving at η. The estimated distribution of the (quantized) stimulus at node η is

$$\hat{p}_\eta(u, v) = \frac{1}{|\mathcal{L}_\eta|} \sum_{i=1}^{C} \mathbf{1}_{\{(s^{(i)}, r^{(i)}) \in \mathcal{L}_\eta\}} \mathbf{1}_{\{\theta(s^{(i)}) \in \Theta_{uv}\}}, \quad u = 1, ..., K_\xi, v = 1, ..., K_\omega.$$

The uncertainty in the stimulus at η is characterized by the conditional entropy $H(\theta(\mathbf{S})|B_\eta)$, estimated by $H(\hat{p}_\eta)$. Let $\Psi(X) = H(\theta(\mathbf{S})|B_\eta, X) = P(X = 0|B_\eta)H(\theta(\mathbf{S})|B_\eta, X = 0) + P(X = 1|B_\eta))H(\theta(\mathbf{S})|B_\eta, X = 1)$. The test X_η at node η is the one minimizing the empirical estimate of $\Psi(X)$:

$$\text{Score}(X) = \frac{|\mathcal{L}_{(d+1,2p-1)}|}{|\mathcal{L}_{(d,p)}|} H(\hat{p}_{(d+1,2p-1)}) + \frac{|\mathcal{L}_{(d+1,2p)}|}{|\mathcal{L}_{(d,p)}|} H(\hat{p}_{(d+1,2p)}).$$

Finally, the left (resp. right) child is determined to be a leaf when $|\mathcal{L}_{(d+1,2p-1)}| < 10$ (resp. $|\mathcal{L}_{(d+1,2p)}| < 10$) or when the entropy falls below a threshold, taken as 1.25 in our experiments.

The algorithm is then the following. Set $\mathcal{L}_{(1,1)} = \mathcal{L}$. For $d = 1$ to $(D-1)$ and $p = 1$ to 2^{d-1}: If $\eta = (d, p)$ is not a leaf, then (i) Compute $X_\eta = \arg\min_{X \in \mathbf{X}_{\mathcal{A}}} \text{Score}(X)$; (ii) Determine $\mathcal{L}_{(d+1,2p-1)}$ based on $X_\eta = 0$ and $\mathcal{L}_{(d+1,2p)}$ based on $X_\eta = 1$; (iii) Set $(d+1, 2p-1)$ or $(d+1, 2p)$ as leaves if appropriate.

4 Experimental Results

We place ourselves in the position of a neurophysiologist who knows little about the auditory system and wishes to understand neural encoding of pure tones in the auditory nerve (a bundle of about 30,000 neurons in the inner ear which originate from the basilar membrane). He chooses a set of pure tones varying in either amplitude or frequency (but not both) and selects neurons into which he inserts microelectrodes to record the sequences of action potentials. Our experiments use simulated responses [5], yielding spike train data which can hardly be differentiated from authentic data by an expert.

Ideally, the trees constructed would have low entropy leaves - in other words, leaves η for which $\hat{p}_\eta$ is peaked around some parameter bin, which then serves as our estimate of the stimulus. Recall that a distribution with *two equal masses* has entropy $H = \log_2 2 = 1$, and hence our stopping criterion ($H \leq 1.25$) does not enforce strong peaking.

4.1 Amplitude Decoding

Case 1: Single-Neuron

Record the responses of a randomly chosen neuron to varying amplitudes at the frequency the neuron is most sensitive to. Specifically, we take $\Omega = \{2\,\text{kHz}\}$ and $\Xi = \bigcup_{u=-1}^{5} \Xi_u = \bigcup_{u=-1}^{5} (10u, 10u + 10]$; thus there are seven classes. The stimuli in the training set are 100 amplitudes randomly selected from each of the seven bins Θ_{u1}, $1 \leq u \leq 7$; thus $C = |\mathcal{L}| = 700$; the corresponding responses are then simulated for the given neuron. Figure 1 shows one spike train per amplitude.

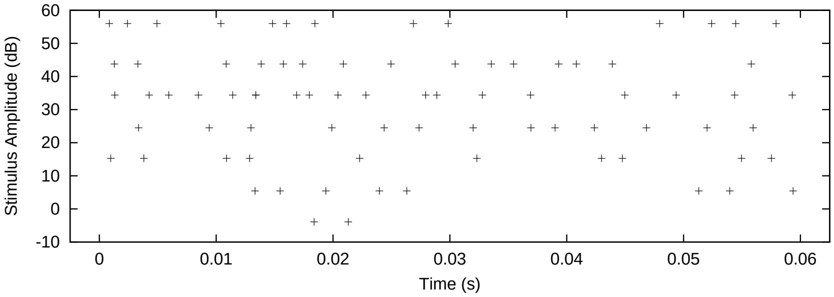

Figure 1: Spike trains for seven different amplitudes at frequency $\omega = 2$ kHz.

A tree was constructed with $I_c = I_0 = [0, 0.06]$, $D = 10$, and questions based on the parameters $M = 5$ and $L = 20$, i.e., checking for exactly $0 - 5$ spikes in intervals of 20 different sizes. In fact, nine different trees were constructed from nine training sets and statistics for each tree were collected based on independent test sets. Aggregate statistics refer to averaging over the nine experiments.The mean tree depth is 7.68 and mean leaf entropy is 1.63 (as compared with 1.13 with the training data), computed from empirical estimates of the probability of reaching each leaf. As a classifier of the amplitude, the performance is about the same for training and test data, namely about 77% of the time the reported amplitude bin is within one of the true one.

Figure 2 shows the distribution of the questions chosen in terms of the interval length, l, and the number of spikes, m, where the averaging is with respect to the (normalized) probabilities of reaching internal nodes. Again, all results are averaged over repeated experiments. (The results for a single tree are approximately the same.) Clearly, a wide range of interval sizes are selected, ranging from 3-60 ms. In contrast, there is a marked preference for checking for either at least one spike or exactly one spike.

To appreciate what the trees are doing - what spike information is sought - consider the first few questions. At the root $(1, 1)$ the amplitude counts are of course $(100, 100, 100, 100, 100, 100, 100)$ and the corresponding distribution has entropy $\log_2 7 = 2.81$. The first question is *"Is there at least one spike during time interval [0,0.0033]?"*; the "yes" child has counts $(9, 16, 32, 72, 93, 99, 100)$ and the "no" child has counts $(91, 84, 68, 28, 7, 1, 0)$. The question at the "no"child is *"Is there at least one spike during the time interval [0,0.0086]?"*, yielding "yes" and "no" offspring with counts $(28, 40, 63, 28, 7, 1, 0)$ and $(63, 44, 5, 0, 0, 0, 0)$, respectively. This last one is a leaf under our stopping criterion. One can observe a tendency to check shorter intervals in the half-tree with leaves identified with higher amplitudes and longer intervals in the other half-tree.

Now reduce the number of questions by considering only $L = 10$ interval sizes and taking $M = 0$, i.e., check only for the presence of at least one spike. Figure 3(a) provides an

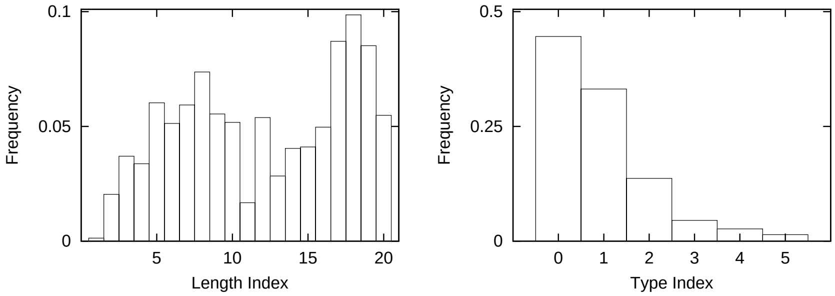

Figure 2: Frequency distributions for interval length (l) and numbers of spikes (m) for single-neuron amplitude decoding, averaged over over nine trees.

empirical justification for this reduction by comparing the mean node entropy by tree depth (for the first five levels) using both training data and test data for the two sets of parameters. (Setting $L = 20$ and $M = 5$ yields 1260 possible questions per spike train, which represents a considerably larger question pool than setting $L = 10$ and $M = 0$ which only yields 55 possible questions per spike train.) The entropy drops clearly demonstrate that there is substantial overfitting in the case of the larger set of questions, as training data entropy drops are much larger than with test data. This is not the case for the smaller set of questions, and *from here on all experiments use the smaller set.*

Figure 3(b) gives the resulting distribution of lengths for single-neuron amplitude decoding with the reduced set of questions (for which training estimates are less biased); again there is a more or less uniform length usage. The corresponding mean tree depth and mean terminal entropy are 7.48 and 1.56, respectively, and the "within-one-bin" classification rate is 77%.

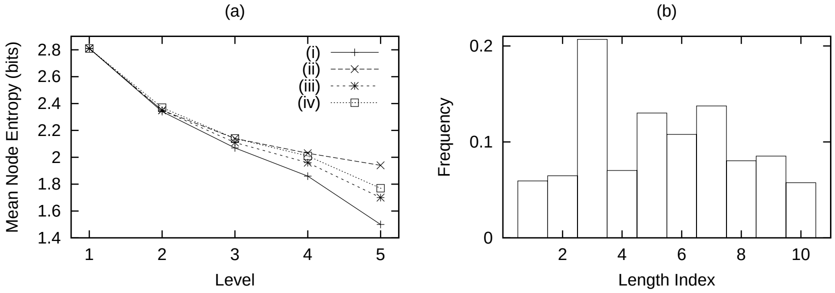

Figure 3: (a) Mean node entropies by depth for $I_c = [0, 0.06]$, $M = 5$ and $L = 20$ with (i) training data, and (ii) test data and for $I_c = [0, 0.01]$, $M = 0$ and $L = 10$ with (iii) training data and (iv) test data; (b) Frequency distributions for interval length (l) for single-neuron amplitude decoding, averaged over over nine trees.

Case 2: Multiple-Neuron

Now keep the same fixed frequency but randomly pick 15 neurons spread out along the basilar membrane. Thus, $C = 700$ but each $r^{(i)}$ is a vector of 15 spike trains. This time, however, we take $I_c = [0, 0.01]$. The reason is that in the previous experiment the chosen intervals usually started at or close to 0; this is likely due to information-rich transition from the simulator background state.

The distribution of interval size is now concentrated on shorter lengths; see Figure 4(a).

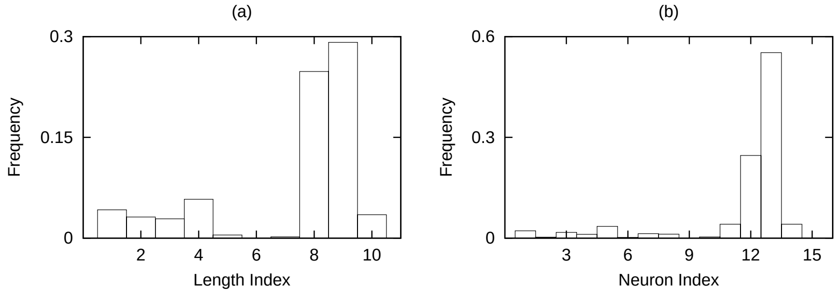

Figure 4: Frequency distributions for interval length (l) and chosen neuron (n) for multi-neuron amplitude decoding.

In addition, most of the questions are based on *only one* of the neurons (see Figure 4(b)), which in fact is the one physically the closest to the neuron in the preceding experiment - the one best-tuned to frequency $\omega_0 = 2$ kHz which is fixed. Moreover, there is a strong disparity between the statistics of the trees as estimated under training and test data: 6.43 vs. 4.73 for mean tree depth and 0.98 vs. 2.42 for mean leaf entropy, with only a 31% classification rate. It appears that amplitude is best decoded by looking at one neuron, and allowing others to be chosen leads to spurious entropy drops due to overfitting rather than decoding.

4.2 Frequency Decoding

Case 1: Single-Neuron

Pick a neuron in the auditory nerve and generate stimuli for a fixed amplitude (here $\xi_0 = 30$ dB) and varying frequencies. Frequency is quantized into 15 bins: $\Omega = \bigcup_{v=0}^{14} \Omega_v = \bigcup_{v=0}^{14} (10^{1.20+0.2v}, 10^{1.20+0.2(v+1)}]$, and $\Xi = \{\xi_0\}$. We randomly pick 100 frequencies in each bin Θ_{1v}, $1 \leq v \leq 15$, resulting in $C = 1500$ training stimuli and spike trains.

Whereas the mean tree depth is above 9 and the initial (root) entropy is $\log_2 15 = 3.91$, the mean terminal entropy is *still above* 3, meaning that the optimal questions yielded insignificant reductions in uncertainty about the frequency. However, this result is in fact consistent with tonotopic encoding in the auditory nerve (that is, a neuron is responsible for its "own frequency"). In particular, frequency decoding with a single neuron is not possible.

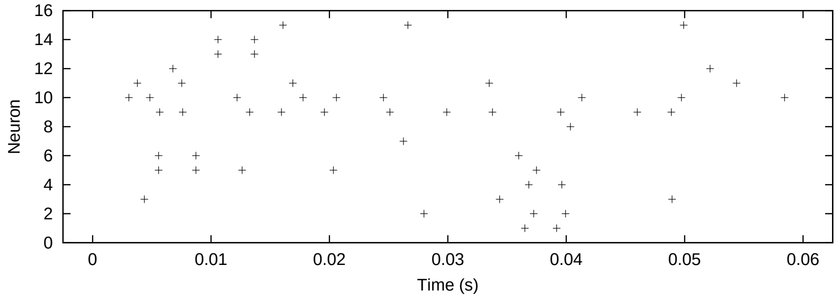

Figure 5: Spike trains for 15 neurons with $\xi = 30$ dB and $\omega = 1.066$ kHz.

Case 2: Multiple-Neuron

Sample 15 neurons in the auditory nerve uniformly along the basilar membrane. Otherwise, the protocol is the same as in Case 1, there again being 15 frequency bins. Figure 5 shows examples of spike trains for the fifteen neurons at one of the frequencies.

The mean tree depth is now 8.62 and the mean terminal entropy is 2.03 (1.6 with training data). Thus, in comparison with a single neuron, this tree is slightly less deep and considerably more powerful; the classification rate is 63% to within one bin. Figure 6 displays the distribution of question usage, which is diffuse, suggesting only a preference for larger time intervals.

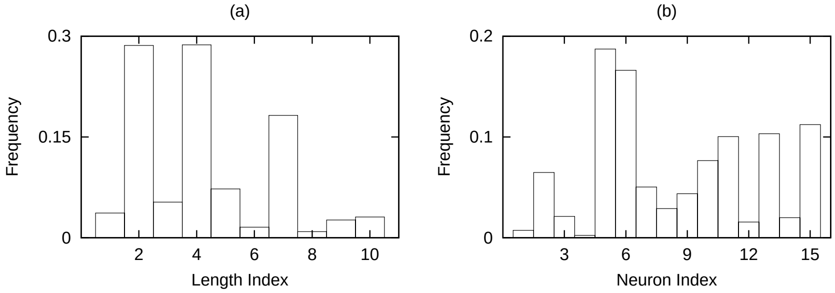

Figure 6: Frequency distributions for interval length (l) and neuron (n) for multi-neuron frequency decoding.

5 Discussion

Decision trees are a standard tool for inducing classifiers from data in machine learning. Typically the goal is to predict the label of an unseen example, often in cases which are trivial for human sensory processing, for example recognizing handwritten digits or detecting faces in digital images. Very high accuracy is desired. Our context is different: The measurements are spike trains, the "class labels" refer to properties of a time-varying stimulus, and, perhaps most importantly, our primary concern in this paper is to determine what the decoding process reveals about which aspects of the stimulus are encoded and how this is achieved.

We have examined tree-structured decoding in the special case of a sequence of binary tests which check for a given number of spikes for a given neuron in a given time interval. Our experiments are aimed at exploring the feasibility of this approach for obtaining information about pure tones based on simulated spike trains involving the auditory nerve. For amplitude decoding at a fixed frequency, it appears that, for all practical purposes, it is sufficient to query the single neuron most sensitive to this frequency; in fact, allowing multiple neurons to be examined degrades performance, probably due to overfitting the training data. Frequency decoding at a fixed amplitude is different: a single neuron is far from sufficient, whereas enough information to crudely predict the given frequency can be gathered from around eight questions from neurons distributed along the basilar membrane. In both cases, the nature of the questions seems to make sense biologically. On the other hand, how such mechanisms might arise in real brains is entirely unclear, as is the mechanism for *sequential learning* - how all this might be done efficiently as information accumulates, whether in natural or machine learning. In the case of trees, it remains to determine the updating or deepening process, for instance how inefficient questions might be replaced without starting over from scratch.

It would likely not be difficult to increase the decoding accuracy. First, the parameters have not been tuned. Second, more questions, and/or more powerful questions, could be considered without promoting overfitting by a variety of well-known methods, such as randomization [6]. In particular, it would be interesting to entertain questions which *simultaneously* query two or more neurons, for example checking for *conjunctions of events* in two spike trains, such as the joint occurrence of a spikes in intervals in two different spike trains. Finally, we considered only very small time buffers, whereas organisms greatly improve performance by integrating information over much longer time periods. For example, our procedure might be implemented every 10 ms and new decisions, and change detection, might be based on “recent” estimates as well as fresh explorations of the spike trains, especially when strong coherence is observed.

6 Conclusion

The encoding of the amplitude and frequency of pure tones in auditory nerve spike trains is a well-understood process. It was hoped that information provided entirely independently from tree-structured decoding would cohere with known results. This seems to have been the case: Whenever the stimulus was well-classified by adaptive testing, the questions made sense biologically, whereas in those cases in which performance was poor there was also a biological interpretation (for example attempting frequency decoding with a single neuron). The results are therefore encouraging.

As to future work, besides extending this framework to synchronized questions and generating a “running commentary”, as discussed earlier, we intend to examine more complex stimuli, such as jointly varying amplitude and frequency, and to explore less-understood neural structures of the auditory system.

Acknowledgments

We wish to thank Dr. Eric D. Young from the Neural Encoding Laboratory at the Johns Hopkins University School of Medicine for his valuable advice throughout this project.

References

[1] Rieke, F., Warland, D., van Steveninck, R.d.R. & Bialek, W. (1999) *Spikes: Exploring the Neural Code.* Cambridge, MA: MIT Press.

[2] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification and Regression Trees.* New York: CRC Press.

[3] Kandel, E.R., Schwartz, J.H. & Jessell, T.M. (2000) *Principles of Neural Science.* 4th Edition. New York: Appleton and Lange.

[4] Pickles, J.O. (1988) *An Introduction to the Physiology of Hearing.* 2nd Edition. London, UK: Academic Press.

[5] Heinz, M.G., Zhang, X., Bruce, I.C. & Carney, L.H. (2001) Auditory Nerve Model for Predicting Performance Limits of Normal and Impaired Listeners. *ARLO* **2**:91-96.

[6] Amit, Y. & Geman, D. (1997) Shape quantization and recognition with randomized trees. *Neural Computation* **9**:1545-1588.