What’s That Sound? Auditory Area CLM Encodes Stimulus Surprise, Not Intensity or Intensity Changes

Patrick Gill1, Sarah M.N. Woolley1, Thane Fremouw1, Frédéric E Theunissen1.

1Department of Psychology and Helen Wills Neuroscience Institute. University of California Berkeley, 3210 Tolman Hall, Berkeley, CA 94720-1650, USA

Running Head: Boo! CLM Encodes Surprise, Not Intensity Changes

Corresponding author:
Frédéric Theunissen
University of California, Berkeley
Department of Psychology
3210 Tolman Hall
Berkeley, CA 94720-1650
Theunissen@berkeley.edu
Abstract

High-level sensory neurons encoding natural stimuli are not well described by linear models operating on the time-varying stimulus intensity. Here we show that firing rates of neurons in a secondary sensory forebrain area can be better modelled by linear functions of how surprising the stimulus is. We modelled auditory neurons in the caudal lateral mesopallium (CLM), of adult male zebra finches under urethane anaesthesia with linear filters convolved not with stimulus intensity, but with stimulus surprise. Surprise was quantified as the logarithm of the probability of the stimulus given the local recent stimulus history and expectations based on conspecific song. Using our surprise method, the predictions of neural responses to conspecific song improved by 67% relative to those obtained using stimulus intensity. Similar prediction improvements cannot be replicated by assuming CLM performs derivative detection. The explanatory power of surprise increased from the midbrain through the primary forebrain and to CLM. When the stimulus presented was a random synthetic ripple noise, CLM neurons (but not neurons in lower auditory areas) were best described as if they were expecting conspecific song, finding the inconsistencies between birdsong and noise surprising. In summary, spikes in CLM neurons indicate stimulus surprise more than they indicate stimulus intensity features. The concept of stimulus surprise may be useful for modelling neural responses in other higher order sensory areas whose functions have been poorly understood.

Keywords: STRF, receptive field, nonlinear, information theory, birdsong, CLM
**Introduction**

The response properties of neurons in higher level sensory areas have been difficult to describe in functional terms, hampering our understanding of the computations performed in these brain areas. One systematic approach has been to describe the stimulus-response relation in terms of the spatio-temporal receptive field (STRF) in vision (DeAngelis et al. 1993) or the spectro-temporal receptive field (also STRF) in audition (Aertsen and Johannesma 1981). In higher sensory areas, this classical-STRF often fails to yield good estimates of the observed responses to natural stimuli (Cohen et al. 2007; David et al. 2004; Machens et al. 2004; Theunissen et al. 2000). The failure of classical-STRFs in higher sensory areas may be related to the consistent finding that the sensory cortex responds robustly to unexpected events (Goldstein et al. 2002; Kiehl et al. 2005; Ulanovsky et al. 2003). If the firing rates of forebrain sensory neurons are proportional to surprise, it should be possible to find linear filters relating firing rates to how surprising each stimulus element is, given an appropriate probabilistic model. To differentiate these filters from classical-STRFs, we call them surprise-STRFs. In as much as surprise-STRFs fit neural data well, we can say action potentials in the sensory forebrain indicate the degree of unexpectedness in the stimulus.

The hypothesis that higher-level sensory neurons are sensitive to unexpected features in the stimulus can also be framed in the context of redundancy reduction, which has been proposed as a general coding principle for sensory systems (Barlow 1961). Natural stimuli have prominent correlations, and thus have non-white power spectra, with low temporal, spatial and spectral modulation frequency components having more power than higher-frequency components (Dong and Atick 1995; Field
1987; Singh and Theunissen 2003). In order to maximize the information being carried by lower level sensory neurons, it has been argued that neuronal stimulus-response functions should account for expected stimulus correlations by attenuating the lower frequencies and accentuating the higher frequencies (Atick 1992; van Hateren 1992b). This operation would result in neural responses to natural stimuli that have equal power at all frequencies and, for this reason, such filters are called whitening filters. Indeed, this theoretical prediction appears to hold in both the visual and auditory systems (Attias and Schreiner 1998; Dan et al. 1996; Escabi et al. 2003; van Hateren 1992a). For natural stimuli that have 1/f amplitude spectra (or 1/f² power spectra), a filter that performs the derivative (spatial or temporal) of the stimulus will effectively whiten the stimulus, removing all second order correlations (Srinivasan et al. 1982).

However, the power spectra of natural stimuli do not capture all of the redundancy present in natural stimuli. We therefore hypothesized that higher order sensory areas represent natural stimuli using a neural code that performs redundancy reduction beyond whitening and thus, for natural stimuli with 1/f amplitude distributions, would have stimulus-response functions that decrease redundancy beyond what would be obtained from a derivative. If so, the neurons would be sensitive not only to differences in stimulus intensity, but also to how predictable these otherwise equal changes in intensity would be. For example, zebra finch song typically contains some harmonic stacks with a smoothly changing fundamental frequency. Modulating the fundamental of a harmonic stack produces widespread changes in the stimulus intensity at all frequencies. However, stimulus intensity changes within these harmonic stacks are more predictable than stimulus intensity changes associated with a song syllable onset, even though the magnitudes of these two intensity changes are similar.
In such cases, a surprise-STRF performing a full redundancy reduction would be more sensitive to the surprising changes at the syllable onset than to the predictable changes during frequency sweep of the harmonic stack.

We hypothesize that the surprise-STRF should be a good model of neural responses in brain regions where redundancy reduction extends beyond simply whitening the power spectrum of natural stimuli. To test this hypothesis, we compared the surprise-STRF’s predictive power to that obtained from both a classical-STRF and from a STRF based on the split absolute derivative of the stimulus intensity. Although the classical-STRF can perform a derivative, the underlying model assumes symmetric sensitivity to positive and negative intensity change. We designed the split absolute derivative STRF so that it could treat positive and negative changes separately. We call STRFs to the split absolute temporal derivative of stimulus intensity the derivative-STRF for short. To the extent the surprise-STRF outperforms the classical- and derivative-STRFs, we can say sensory spikes signify stimulus surprise, not stimulus intensity changes.

We estimated classical-, derivative- and surprise-STRFs for single neurons at three levels of the zebra finch auditory system: the midbrain nucleus mesencephalicus lateralis dorsalis (MLd; analogous to the inferior colliculus in mammals), the primary auditory forebrain area Field L (analogous to primary auditory cortex in mammals), and a secondary auditory region, the caudal lateral mesopallium (CLM). Caudal mesopallium has been implicated in the processing of learned and behaviourally relevant sounds (Gentner and Margoliash 2003). CLM may therefore have stimulus-response properties that are particularly efficient at representing birdsong. Songbirds are a model system studied to understand both vocal learning and processing of
complex vocalizations (Zeigler and Marler 2004). In the songbird auditory system, natural sounds are processed preferentially, eliciting higher spiking rates in the forebrain (Grace et al. 2003), higher information rates in the midbrain and forebrain (Hsu et al. 2004b) and higher neural synchronization in the midbrain (Woolley et al. 2006) than do synthetic sounds. Moreover, classical-STRFs of these neurons show tuning for the informative features of natural sounds (Woolley et al. 2005). Nonetheless, the classical-STRF cannot fully describe the observed spiking patterns, and the discrepancy between model prediction and actual response increases from MLd to Field L to CLM (Gill et al. 2006; Sen et al. 2001). We tested whether we could improve our model predictions by using surprise-STRFs or derivative-STRFs.
Materials and Methods

The Surprise Metric

The surprise of any specific sound element depends on what is expected. To determine how surprising each stimulus element is, we based our expectations on an ensemble of conspecific birdsongs because of the relevance of this class of sounds for the survival of the animal (Gil and Gahr 2002) and because higher auditory and vocal areas in songbirds might be specialized for processing behaviourally-relevant sounds (Theunissen et al. 2004). We then calculated stimulus probability based on the typical behaviour of birdsong and on a window of stimulus history we call the domain $D$. To find the best $D$, we first ran an analysis that showed that neurons in MLd, Field L and CLM are largely insensitive to the context of zebra finch song on long time scales; specifically their response does not appear to be sensitive to expectations based on song motif structure (see also Results). The domain $D$ could therefore be limited to the past few hundreds of ms (the length of one motif). The $D$ that lead to the best predictions was a rectangle of the spectrogram including the recent history of the stimulus between 4 and 7 ms prior to the stimulus whose surprise is predicted and nearby frequency information within .625 kHz of the stimulus frequency being predicted (see Figure 1B for an illustration of $D$ and the Results section for an account of how the optimal $D$ was found). We quantified surprise in the following way:

$$\text{Surprise} = \begin{cases} -\log(P(S | D)) + \log(P(S_{ML} | D)) & \text{if } S > S_{ML}, \ 0 & \text{otherwise} \\ -\log(P(S | D)) + \log(P(S_{ML} | D)) & \text{if } S < S_{ML}, \ 0 & \text{otherwise} \end{cases}$$

Here $D$ is the domain, $S$ is the stimulus intensity whose surprise is being considered, $S_{ML}$ is the most likely stimulus given $D$, and $P(S|D)$ stands for the conditional probability of
$S$ given $D$ and depends on the knowledge of statistical dependencies in the sounds from a representative corpus of unfamiliar conspecific songs. The $S_{ML}$ term ensures that $\text{Surprise}$ is continuous at $S = S_{ML}$. $\text{Surprise}$, $S_{ML}$, $S$, and $D$ are all functions of frequency and time. $\text{Surprise}$ is an array with twice as many spatial dimensions as the original spectrogram; its top half contains louder-than-expected features and its bottom half contains quieter-than-expected features. Treating louder-than-expected and quieter-than-expected features identically would destroy the sign of the surprise in the sense that an event so loud that it was only 2% as likely as the most probable event would have the same representation as a quieter-than-expected event with equally low probability. Therefore, as long as $P(S|D)$ is unimodal (with mode $S_{ML}$) the surprise representation is invertible, thus surprise is a complete stimulus representation: any neural code which efficiently relates stimulus surprise also relates all details of the stimulus intensity (and not just the surprising parts of the stimulus). It should also be noted that for certain domains, a stimulus change is more probable than a constant stimulus. In those cases, an observed constant stimulus would be more surprising than the expected change in the stimulus.

The surprise metric is similar to that used in another study (Itti and Baldi 2005) where the surprise in natural video was defined as the Kullback–Leibler (K – L) divergence between expectations about the stimulus (the Bayesian prior) and the stimulus representation once the stimulus had been perceived (the Bayesian posterior). The above K – L divergence is proportional to our surprise metric if there is a small degree of uncertainty in the Bayesian posterior and if this uncertainty is constant.

The surprise-STRF is mathematically similar to the classical-STRF in the sense that both models assume that firing rates are linearly related to a nonlinear
transformation of the sound pressure waveform. However, the nature of the nonlinear
stimulus transformations is very different. Spikes represent specific patterns of surprise
when modelled by surprise-STRFs, and specific patterns of stimulus intensity when
modelled by classical-STRFs, but mathematically the former is not more nonlinear than
the latter.

Our surprise-STRF is also formally similar to the Linear Probabilistic Receptive
Field (LPRF) (Averbeck and Romanski 2006) in that both assume stimulus probability
is encoded, but differs in that the LPRFs in that study made firing rates proportional to
the probability of the stimulus belonging to a given class of macaque vocalizations,
while our surprise-STRFs make firing rates proportional to surprise in local sound
intensity features.

**Why Log(P)?**

There are several reasons to expect the log of stimulus probability to be linearly
related to firing rates. Using the log function, the summed surprise of two independent
events $a$ and $b$ is the same as the surprise of the conjunction of $a$ and $b$, since

$$\log(P(a)*P(b)) = \log(P(a)) + \log(P(b))$$

and the probability of both $a$ and $b$ occurring is $P(a)*P(b)$. The log function also has special relevance to information theory (Shannon
and Weaver 1963). Equation 1 implies that surprise will be linearly related to the
number of bits needed to encode the signal given knowledge of $D$ and $P(S|D)$. Thus, if
bit rates and spike rates are roughly proportional as reported (reviewed in (Borst and
Theunissen 1999)) then we would expect - $\log(P(S|D))$ to be proportional to the number
of spikes a neuron fires.
Calculating $P(S|D)$

To sample our domain we used a time bin of 1 ms and a frequency bin of 125 Hz (the same sampling as for our spectrograms – see Figure 1A). Since the domain $D$ has 3 different latencies and 9 different frequencies, it has a total of 27 free parameters. For the estimation of surprise using Equation 1, we need to compute the conditional probability $P(S|D)$, which without modification would be a 28-dimensional object (one for the stimulus intensity $S$ in addition to the domain parameters). Since it is more difficult to estimate the conditional probabilities of high-dimensional objects, and since there is redundancy in the 27 adjacent spectrogram bins of any zebra finch song, we reduced $D$ to a 9-dimensional object using principal component analysis (PCA). Nine PCs were the maximum computationally-practical number of PCs to consider, and they captured over 97% of the variance of $D$. We then estimated the 10-dimensional $P(S|D)$ using a corpus of 59 adult male zebra finch songs chosen for their diversity and high recording quality. We made an independent estimate of $P(S|D)$ for each frequency band, since the statistics of spectrograms of zebra finch song depend on frequency. Frequency bands for which $D$ overlapped the edge of the spectrogram were discarded.

We used a discretized Gaussian kernel method for our estimate of $P(S|D)$ as follows. We created a 10 dimensional lattice array to keep track of the joint probability $P(S,D)$ of the domain and its stimulus. For each $(S,D)$ pair in the corpus, we incremented the probability of each of the vertices of the 10 dimensional hypercube containing $(S,D)$ (i.e., the closest lattice points in all directions) with an amount proportional to $e^{-d^2}$, where $d$ is the Euclidian distance from $(S,D)$ to the hypercube vertex. The $d$ between lattice points was normalized to 1. To make the computation feasible, our method differed from traditional kernel density estimation in two ways: the
discretization we used was relatively coarse, and we incremented only the closest lattice points in all directions. We used more lattice bins to describe $S$ and the stronger PCs of $D$ than for the weakest PCs of $D$. We divided $P(S,D)$ by $P(D)$ to find $P(S|D)$ (by Bayes’ Theorem). After computing $P(S|D)$, for every point in each stimulus we used for presentation, we quantified surprise by linearly interpolating $P(S|D)$ from the closest lattice points.

Our method for estimating surprise is not necessarily unique or definitive. For example, relatively large changes in the latency, time window and frequency range of $D$ lead to small differences in performance (see Results). We believe the key features for good stimulus representations for secondary forebrain areas are 1) to represent the stimulus not in terms of its intensity, but in terms of how unexpected it is, and 2) the use of the log function, which gives the representation a theoretical grounding in information theory, and compresses large discrepancies in different estimations of $P$ into small discrepancies in log($P$).

**Stimulus Intensity Derivatives**

Our surprise representation has two confounding potential advantages over traditional spectrograms: twice as many free parameters allow onsets and offsets to be modelled differently (i.e. onsets are not constrained to have the opposite effect of offsets), and temporal changes (which are known to drive higher auditory areas) are highlighted. To control for both of these effects we also estimated split absolute derivative STRFs using a split time-derivative of the spectrogram, where increasingly-loud features (crescendo) are separated from increasingly-quiet features (decrescendo) as in Equation 2:
\[ \text{Der} = \begin{cases} \frac{\delta S}{\delta t} & \text{if } \frac{\delta S}{\delta t} > 0, \ 0 & \text{otherwise} \\ \frac{\delta S}{\delta t} & \text{if } \frac{\delta S}{\delta t} < 0, \ 0 & \text{otherwise} \end{cases} \]  

(2)

Here \( S \) is the spectrogram intensity, \( t \) is time and \( \text{Der} \) is the split derivative to be used as a stimulus representation for prediction. If the role of neurons is to report all changes in stimulus intensity equally (as would be expected with a whitening code applied to stimuli with a \( 1/f^2 \) power spectrum), then this representation should be ideal.

**Receptive Field Estimation**

For all our stimulus representations, we calculated STRFs using STRFPAK version 5.2 (http://strfpak.berkeley.edu), which incorporates a general linear regression algorithm, regularization techniques, validation techniques and metrics described in detail before (Hsu et al. 2004a; Theunissen et al. 2001; Woolley et al. 2006). Version 5.2 also includes a double jack-knife, which ensures that data used to find the best regularization parameters is not also used to evaluate the final goodness of fit. The goodness of fit of the STRF was quantified by estimating the correlation coefficient and the mutual information between the predicted and actual PSTHs. The predicted information is a function of the integral of the coherence over all frequencies between the actual and predicted responses (Hsu et al 2004a). Predicted information has the advantage over correlation coefficients in that the experimenter is not forced to choose a time scale of relevant PSTH features. Correctly predicted temporally precise features are rewarded by predicted information, in contrast with correlation coefficient methods which usually smooth out fine PSTH features like those shown in Figure 1I. In this study we use predicted information to quantify most of our results, but also report correlation coefficients measures for comparison. The correlation coefficients were calculated after
smoothing the PSTH with an 11 ms Hanning window and correcting for noise in the
PSTH estimation (Hsu et al. 2004a).

**Neurophysiological Recordings.**

Neural data were obtained from 46 adult male zebra finches. All subjects were
reared in a colony in natural family groups, and were not exposed to any of the songs
used as a stimulus prior to the neurophysiological recordings session. Single-unit
responses were obtained with extracellular tungsten electrodes in urethane-anesthetized
birds. The location of the recordings was verified with standard histological techniques:
for CLM, n = 37; Field L, n = 189; and MLd, n = 143. Each subject underwent
simultaneous recording in either in both Field L and MLd (n = 29) or in both CLM and
Field L (n = 17). Sounds were played from a loudspeaker placed 15 cm in front of the
animal, and sound levels had peak intensity of 70 dB SPL. All neurons in CLM were in
the lateral subdivision. Neurons in Field L were sampled from all sub-regions (L1, L2a,
L2b, and L3). Data from 23 of these birds were also used in previously published work,
and additional information on stimulus design, neurophysiological recordings and
histological techniques can be found in previous studies (Hsu et al. 2004b; Woolley et
al. 2005; Woolley et al. 2006). All experimental procedures were approved by the
Animal Care and Use Committee of UC Berkeley.
Results

The quantified surprise of the stimulus was estimated by computing the log of the probability of the stimulus intensity at all frequencies and time points given the stimulus’ recent history, i.e., its domain $D$ (see Methods). These conditional probabilities were obtained from a corpus of 59 zebra finch songs (see Methods and Figure 1 A-E). To avoid throwing away stimulus information and to keep our representation invertible, surprisingly quiet and surprisingly loud stimulus elements were separated.

<<Please insert Figure 1 near here. It is a colour figure 18 by 14 cm.>>

We calculated and validated classical-, derivative- and surprise-STRFs for all the neurons in our dataset. Figure 1F shows the classical-STRF (the STRF based on stimulus intensity shown in the spectrogram) of a CLM neuron, while Figure 1G shows the surprise-STRF of the same neuron. The upper panel of Figure 1G shows the filter to be convolved with louder-than-expected events and the lower panel shows the filter for quieter-than-expected events.

From a purely linear standpoint, the classical-STRF model predicts that this neuron increases its firing at onsets and decreases its firing below mean levels at offsets. However, the surprise-STRF indicates that this neuron is not inhibited by quieter-than-expected activity; if anything, firing rates may increase slightly in response to unexpected quietness.

The prediction improvement of the surprise-STRF shown in Figure 1 is typical of area CLM. Figure 2 compares performances of classical-, derivative- and surprise-STRFs in CLM. As shown in Figure 2A, every CLM neuron that can be reasonably modelled with linear filters (those with a prediction score above 1 bit/s) is described
better by a surprise-STRF than by a derivative-STRF. Moreover, the preference for surprise-STRFs was evident in most subjects. For 16 of the 17 zebra finches that had CLM recordings, CLM neurons are described better by surprise-STRFs than by derivative-STRFs, and the one counterexample is a subject with only one CLM recording site. The surprise-STRF also outperforms the classical-STRF in CLM by a larger average margin, as shown in Figure 2B. Figure 2C shows that the derivative-STRF outperforms the classical-STRF by a smaller margin.

<< Please insert Figure 2 near here. It is a greyscale figure 4 by 10.5 cm.>>

In more peripheral areas, surprise-STRFs do not improve predictions to the same extent as in CLM. Figure 3A shows the mean prediction scores of classical-, derivative- and surprise-STRFs for CLM, Field L and MLd. The error bars show one standard error of the mean (SEM); however since the data are non-Gaussian (see Figure 2) the SEM is not a definitive indicator of variability. Surprise-STRFs consistently outperform classical-STRFs and derivative-STRFs in CLM (already shown in Figures 2B and 2A) and in Field L, as can be seen in Figure 3B, which shows the improvement of using surprise-STRFs over classical- and derivative-STRFs. Numbers above bars in Figure 3B are the p values using paired two-tailed Wilcoxon signed-rank tests. In CLM, the mean surprise-STRF prediction score is 67% larger (p = 2 * 10^{-6}) than the mean classical-STRF score. The improvement can also be seen when correlation coefficients (r) are used to quantify the goodness of fit. For classical-STRFs the mean r is .38, for derivative-STRFs the mean r is .39 and for surprise-STRFs the mean r is .48. The improvement of the surprise STRF over the classical STRF in terms of correlation coefficients is 26% (p = 4 * 10^{-6}). The correlation coefficient shows a smaller improvement because, in comparison to the information calculation, it is more insensitive to the gains obtained in predicting the faster transients. (In predicting
responses to natural stimuli, improvements on the order of 10% are considered noteworthy; e.g. see (Prenger et al. 2004)).

The prediction improvement of surprise-STRFs in CLM cannot be explained by the ability of surprise-STRFs to react to intensity derivatives. Although the derivative-STRF performs better than the classical-STRF (by 19%, \( p = 8 \times 10^{-4} \), see Figures 2C and 3A), the surprise-STRF outperforms the derivative-STRF by 40% (\( p = 5 \times 10^{-7} \), see Figures 2A and 3B). Although, as illustrated in Figure 1 there is room for prediction improvement, our current best model for CLM is that it encodes stimulus surprise, not intensity changes.

Field L, but not MLd, showed some evidence of surprise coding. On average, PSTHs in Field L are described only 24% better by surprise-STRFs than classical-STRFs (\( p = 8 \times 10^{-5} \)). The surprise-STRF also out-performs the derivative-STRF in Field L by 14% (\( p = 2 \times 10^{-7} \)), and the derivative-STRF outperforms the classical-STRF by 9% (\( p = .01 \)). The effect size of the surprise improvement over the derivative-STRF is smaller in Field L than in CLM (\( p = 7 \times 10^{-5} \), Wilcoxon rank sum test on the differences between surprise-STRF performance and derivative-STRF performance in CLM and Field L), and MLd shows no significant improvement of surprise-STRFs over either derivative-STRFs or classical-STRFs. Field L, physically between MLd and CLM along the ascending auditory pathway, also shows a functionally intermediate degree of tuning for surprise.

The advantages of using the surprise-STRF were not limited to better prediction of the neural response to onsets and offsets. In fact, surprise-STRFs offer the least improvement in MLd (see Figure 3), the area with the most onset detection (see Figure
Moreover, surprising syllable features other than onsets, like the spectral change within a syllable pointed out by the red braces above panels A, B, D, and E, in Figure 1, could elicit neural activity that was well-captured by a surprise-STRF and not as well captured by the other STRF models, as shown in panels H and I. Conversely, non-surprising acoustic changes often led to predicted spikes in classical- and derivative-STRFs that were not observed in the actual data.

Finding the Domain D

In the search for the correct domain (see Methods), only domains with a relatively short history were examined because the auditory areas studied did not appear to exhibit sensitivity to long range context (> 500 ms) and because computational and data limitation reasons prevented us from obtaining good estimates of longer probabilistic relationships in zebra finch song. Zebra finch song typically has repeated motifs consisting of the same syllables in the same order (Figures 4A and 4B). If neurons encode surprise and are sensitive to long term context such as syllable order within a recently-heard motif, they should be less surprised by the second repetition of a motif, and firing rates in response to the second repetition should be systematically lower. Figures 4C, 4D and 4E show the mean PSTHs for MLd, Field L and CLM for two repetitions of the same motif in one stimulus. These mean PSTHs are not substantially different, suggesting that, on average, the neuron’s responses are not sensitive to motif-long context.

We therefore focused on investigating the optimal domain on shorter time scales. For computational reasons, we tested domains that were restricted to the immediate past: domain widths of 2 to 6 ms, minimum latencies of 3 to 6 ms and spectral half-
widths of 325 to 750 Hz. These time scales allowed us to capture expectations within a zebra finch syllable but failed to capture expectations that would result from reproducible sequences of syllables in motifs. The spatio-temporal extent of the domain $D$ was found through gradient ascent of PSTH predictability by changing the latency range and the frequency width parameters of $D$. The optimal $D$ had latency from 4 to 7 ms and spectral width of 625 Hz. Our initial guess of $D$ was half the size of the optimal $D$ and yet predicted less than 1% worse than our final choice, suggesting that the details of how surprise is quantified do not affect model performance significantly.

Correlations in conspecific song on these spectral and temporal scales are so strong that all of the domains we tried contained nearly identical information, and thus our estimate of surprise does not depend on the exact extent of $D$. Conversely, since all $D$ options we chose performed similarly we do not claim that the shape of the $D$ has a biological analogue. Moreover, it is possible that a larger domain that captures relationships in syllable sequence could further improve predictions, but a mathematically more complicated $D$ would be needed to make searching for these dependencies computationally feasible.

**No Stimulus Adaptation Found**

Aside from characterizing the stimulus-response function of CLM neurons, we asked whether these neurons exhibited plasticity during the course of the recording sessions. If CLM neurons change their expectations on a timescale ranging from minutes to hours, then systematic reductions in firing rates to repetitions of the same stimuli should be observed. We did not find any evidence of habituation in CLM, Field L or MLd neurons, as is presented in the remainder of this section.

To assess habituation in firing rates, we ran two statistical tests on PSTHs. First, we performed a linear regression between the total number of spikes elicited in a given trial and the trial number (most stimuli were repeated 10 times): the “same neuron, later
trial” test, pooling all data from all CLM neurons. The slope of this regression was not
significantly different from 0, but showed an insignificant increase in mean firing rate of
0.4 spikes per stimulus per trial (two-sided F-test, p value = 0.3). Second, we
performed a linear regression between the mean number of spikes a CLM neuron fired
per stimulus presentation and the number of times the (approximately one hour long)
stimulus protocol had been played to the animal being tested: the “same subject, later
neuron” test. This regression yielded an insignificant decrease in mean firing rates of
0.2 spikes per stimulus per previous presentation (p = 0.9). Thus the CLM neurons we
recorded from did not systematically decrease their firing rates in response to repetitions
of initially-unfamiliar stimuli. Neither Field L nor MLd showed significant systematic
adaptation either (uncorrected p values for Field L are 0.3 and 0.8, and for MLd are 0.8
and 0.1, for the “same neuron, later trial” and the “same subject, later neuron” tests,
respectively; the smallest Bonferroni-corrected p value is 0.7). Therefore systematic
stimulus specific adaptation is not relevant to understanding MLd, Field L and CLM
under these experimental conditions, although firing rates might be plastic over the
course of longer habituation, under different anaesthesia, or by pairing a reward with a
particular stimulus. It should be noted however that during our recording sessions we
stopped acquiring data from specific sites if firing rates dropped to zero long before the
presentation protocol was finished. These neurons were not analyzed and it is possible
that they constitute a separate population of highly adapting neurons. Alternatively,
these are neurons that could have been damaged by the electrode, or the neural signal
was lost because of other experimental artefacts such as brain movement.

**Offset Cells and Auditory Complex Cells**

The use of the surprise-STRF revealed two functional classes of forebrain
auditory neurons that could not be well-characterized by classical-STRFs: offset-only
neurons and auditory complex cells. The former are modelled as firing in response to
offsets, while the latter are modelled as firing whenever there is any surprising change in intensity (be it louder or quieter) within the latency and frequency range of the neuron. The classical- and surprise-STRFs of an offset-detecting neuron from Field L are shown in Figure 5A.

<< Please insert Figure 5 near here. It is a colour figure 12.7 by 13.1 cm.>>

From its classical-STRF alone the neuron in Figure 5A appears to be both an onset detector at high latency (15 ms) and an offset detector at low latency (5 ms). However, using the surprise-STRF, it becomes clear that the best model for this neuron is of one that largely ignores louder-than-expected features (Figure 5A, surprise-STRF, louder) but is excited by quieter-than-expected features between about 0.5 and 5 kHz with a latency of about 8 ms (Figure 5A, surprise-STRF, quieter).

Auditory complex neurons, found primarily in CLM, are general change detectors, firing to any unexpected stimulus pattern happening within a window of frequency and latency, regardless of the direction of that change (i.e. louder- or quieter-than-expected features). Figure 5, parts B-D show three CLM complex neurons with different tuning characteristics. Figure 5B shows a neuron which fires when any surprising change happened between about 5 and 25 ms ago, in the frequency range of about 3 to 5 kHz. More broadband complex neurons are shown in Figures 5C and 5D.

Nine of the 37 CLM neurons we examined seemed to be auditory complex cells. Auditory complex cells and offset cells can be fit to some degree by derivative-STRFs, which is why in CLM and Field L derivative-STRFs out-performed classical-STRFs there (see Figure 3). However, since surprise-STRFs outperform derivative-STRFs by 40% (p = 5 * 10^{-7}), the probability-related aspect of surprise is still crucial for understanding CLM.
Random Stimuli and Unmet Expectations

Next, we checked if the surprise-STRF model yields similar increases in performance when we predict responses to modulation-limited noise (ML noise), a type of random stimulus that is frequency limited in its temporal (0 – 50 Hz) and spectral (0 – 2 cycles per kHz) modulations (Hsu et al. 2004b). Contrary to white noise, ML noise drives high-level auditory neurons with firing rates similar to those driven by behaviourally relevant complex sounds and with spike patterns that are reliably in phase with acoustical structure in the sound. ML noise has therefore been used to extract basic response parameters including STRFs from cortical auditory neurons (Escabi and Schreiner 2002; Klein et al. 2000). Figure 6A shows a spectrogram of a sample of ML noise.

<<Please insert Figure 6 here. It is a colour figure 8.9 by 15 cm. >>

As with song, it is possible to represent ML noise in terms of surprise rather than intensity. We quantified the surprise associated with ML noise in two ways: first by assuming knowledge of the statistics of ML noise (see Figure 6B), and second by assuming that song was expected (see Figure 6C). Expectations were controlled by estimating $P(S|D)$ (see Equation 1) using a corpus of zebra finch song (for the song-expecting surprise as in Figure 6C) or a corpus of ML noise (for the ML noise-expecting surprise as in Figure 6B). As can be seen in the colour scale bars in Figures 6B and 6C, when song is expected, ML noise has more surprising features than when ML noise is expected. (The maximum surprise value in Figure 6C is 7.5, meaning this stimulus feature is only $e^{-7.5}$, or 0.06%, as likely as the most likely stimulus intensity. The maximum surprise value in Figure 6B is 3.9, meaning this stimulus feature is 2% as likely as the most likely stimulus intensity.) Also note that when $P(S|D)$ is based on song, different features are unexpected. Since the stimulus model $P(S|D)$ expects slow broadband features common in zebra finch song, the rapid offsets occurring
immediately after onsets in ML noise are quite surprising, as are ML noise’s sharp spectral edges.

We calculated three types of STRF to ML noise: a classical-STRF, a surprise-STRF expecting ML noise and a surprise-STRF expecting song. Figure 7 gives all pairwise performance comparisons of these three STRF types in CLM. Figures 7A and 7B show that surprise-STRFs based on song expectations in general outperform both classical-STRFs and surprise-STRFs based on expectations of ML noise.

Once again, surprise-STRFs performed differently in the three brain regions examined here. Mean and relative prediction scores are plotted in Figure 8.

Figure 8A shows mean prediction qualities of the three ML noise STRF types for the three brain areas investigated. The most striking result from ML noise STRFs is the improved performance of the surprise-STRF in CLM when song is expected, as was shown in Figure 7A. Not only does the song-expecting surprise-STRF outperform the classical-STRF (by 62%, p = 3 * 10^{-5}, see Figures 7B and 8B), but estimating surprise using expectations of song predicts 58% better (p = 6 * 10^{-5}, see Figures 7A and 8B) than by expecting ML noise. In other words, responses to ML noise are best described if we assume that CLM uses a relatively fixed coding strategy: its neurons spike when they encounter an unexpected acoustical feature given they are expecting song even when the stimulus is not song. This song-expectation is not found in Field L or in MLd, as shown in Figure 8B. However, neurons in MLd tended to be described better by either surprise-STRF than by a classical-STRF.
Discussion

Using the surprise-STRF model, we were better able to predict the responses of forebrain auditory neurons to both natural and synthetic sounds, especially in the secondary auditory forebrain area, CLM. Two main points underscore the importance of surprise in the zebra finch higher auditory forebrain. First, surprise-STRFs in CLM outperform classical-STRFs by 67% and outperform derivative-STRFs by 40%. The evidence that CLM performs an operation more sophisticated than derivative detection is conclusive: the size of the effect is large (40%) and the difference is highly significant (p = 5 * 10^{-7}, see Figure 2A). In comparison, in CLM, the derivative-STRF only outperforms the classical-STRF by 19% (p = 8 * 10^{-4}). Therefore, the explanatory power of surprise-STRFs arises mostly from capturing the higher order structure of the stimulus correlations and not just from the emphasis on temporal stimulus changes or from an increased number of model parameters. Second, when the stimulus is not song, the best description for CLM behaviour is still a surprise-STRF. When processing ML noise, the improvement over other models of CLM activity is clear: the song based surprise-STRF performs 62% better (p = 3 * 10^{-5}) than the classical-STRF and 58% better (p = 6 * 10^{-5}) than a surprise-STRF expecting ML noise. Although the surprise-STRF is not capable of capturing all of the deterministic behaviour of CLM neurons, we propose that it is a better foundation than stimulus intensity-based models on which to build more sophisticated nonlinear models, including surprise-STRF models with longer memory.

An alternative interpretation to our main result (that coding in CLM is well-characterized by surprise-STRFs) is that CLM neurons do not so much encode surprise as they implement some nonlinearities which are similar to our surprise formulation...
Our theoretical methodology is still useful for two reasons. First, surprise has lead us to a model for CLM neurons with unprecedented predictive power, which may prove to be an important step towards even better models of secondary forebrain neural functions. Second, the concept that higher sensory neurons encode surprise is theoretically rich, providing not just a functional description of CLM but a theoretical grounding for the type of stimulus representation found there.

The fact that surprise has more predictive power in CLM than in Field L suggests two possibilities. First, Field L might not complete the computation of what is surprising. Some computation is required to silence unsurprising features, and in Field L auditory information may not have passed through enough synapses to perform the needed surprise calculation. Second, the neural representation of surprising features in Field L might depend on expectations from any sound stimulus and not just conspecific song. This second hypothesis would be verifiable in a study where surprise is estimated with the expectation of a variety of different natural sounds, not just zebra finch song.

When determining the optimal domain for the surprise estimation, we selected a relatively small domain (with latency from 4 to 7 ms and spectral width of 625 Hz) that did not capture the expectations in song found at the level of syllable sequence and repeated motifs. What advantage could there be for the zebra finch auditory code in MLD, Field L and CLM to effectively ignore the large redundancies of motif repetition, as shown in Figure 4? One possible explanation is that having the same neural representation for the two motifs makes it easier for a song grammar detector downstream of CLM to notice if the repetition of motifs was exact. Note the remarkable similarity between the first and second motif repetitions in Figures 4A and 4B. The ability to produce exactly the same motif might be used in female zebra
finches as a proxy for general fitness during mate selection. Detecting precise motif repetition may be easier if the neural representation of each motif is identical. Another explanation for the absence of grammar-sensitive coding is that the neural circuitry needed to store uncompressed detailed auditory information may be unwieldy. If one of the listener’s tasks is to learn a song efficiently or to notice repeated motifs, it is most efficient to first remove redundancy from the song’s acoustic representation so that fewer spikes need to be recalled. The surprise-detecting mechanism found in CLM is ideal to perform this pre-processing step, since it provides an efficient and complete representation of zebra finch song.

Because of the short timescale we use, our use of the term “surprise” can appear to be significantly different than the use of “surprise” in the cognitive neuroscience literature, in which the domain is longer in time and stimulus semantics are more relevant. However, these differences are differences of scale and level, and surprise might be represented at all levels of forebrain processing. Surprise is used here to mean sensory-model mismatch left over after extensive redundancy reduction (Barlow 2001). The type of redundancy reduction at the level of single neuronal responses that we observed in CLM (also referred to as life-time sparseness) might work in conjunction with redundancy reduction across neurons (population sparseness). In the mammalian auditory system, population redundancy has been shown to decrease in the ascending auditory pathway (Chechik et al. 2006).

Another interpretation consistent with our results is that CLM acts less as a specialized auditory encoder than as a mediator of bottom-up attention. In a study estimating surprise in natural movies (Itti and Baldi 2005), it was noted that the observer’s gaze consistently shifted towards areas in the movie which had more surprise
quantified using a mechanism similar to that of Equation 1, see Methods). A correlate to our central hypothesis (that CLM encodes birdsong well by using a surprise-based coding strategy) is that CLM can signal to other neurons when parts of a song have changed enough to warrant a resampling of recent auditory history.

Neural codes that use few spikes to represent surprising features of stimuli are efficient for both metabolic and computational reasons (Olshausen and Field 2004). Although estimations of metabolic cost are somewhat controversial given the physiological complexity of the problem, there is a consensus that spiking and synaptic transmission account for most of the energy expenditure of the brain (Attwell and Laughlin 2001; Lennie 2003). In some calculations, the cost of spiking is estimated to be so high that only a very small fraction of neurons could be active at the same time, necessitating a very efficient representation (Lennie 2003). Among sparse representations, encoding surprising features is desirable since the surprise transformation is invertible (see Methods), meaning the entire stimulus has been encoded.

Independent of metabolic constraints, representing the stimulus in terms of surprise might be important for computational reasons. For example, sparse codes have also been shown to be beneficial in memory systems: it is easier to make new associations between features when their representations are sparse (Schweighofer et al. 2001). Moreover, the complicated task of visual object recognition is facilitated by combining prior knowledge of the images while processing current image features (Kersten et al. 2004). In a Bayesian framework, prior knowledge and sensory information are combined by multiplying probability distributions. If the neural representation of the stimulus is in a form where spikes correspond to log probabilities,
posteriors could be calculated through addition (not multiplication) of spikes, since adding log probabilities is equivalent to multiplying raw probabilities. In most dendrites, addition is a simpler operation than multiplication, so representing the stimulus in terms of its log probability is an ideal pre-processing step for Bayesian object recognition.

In the context of memory storage, it should be noted that CM has been implicated in storing songs used in learning tasks (reviewed (Bolhuis and Gahr 2006)). Firing rates in CMM (the medial portion of CM) in response to familiar conspecific song motifs used in two-alternative choice tasks and go/no-go tasks are higher than in response to unfamiliar conspecific song motifs (Gentner and Margoliash 2003). In the song system, responses to the bird’s own song (which is very familiar) are maximal (reviewed in Theunissen et al. 2004). Higher firing rates to familiar songs apparently contradict the idea that firing rates are proportional to stimulus surprise. However, it is possible that surprise is used to represent songs efficiently in the naïve state. This representation then gets modified in behavioural tasks that engage motivational systems and learning. Task-relevance (not reflected in Equation 1) should also influence the neural code; task-relevant features should have a stronger representation than irrelevant features, although our surprise model does not yet reflect this principle. Training or vocal learning would enhance relevance of certain stimuli and thus the corresponding spike rates. It is also likely that task-relevance explains why random (and thus intrinsically surprising) artificial stimuli like white noise do not drive CLM as well as familiar but relevant stimuli like conspecific song. Furthermore, higher responses to a familiar song might still be to the surprising features of that song, given the statistics of all recently heard songs. It should also be noted that there are anaesthetic (urethane vs. awake), species
(zebra finch vs. starling) and region (CLM vs. CMM) differences between this study and (Gentner and Margoliash 2003). Moreover, the surprise-STRF model did not capture all the nonlinear encoding observed in the response. What acoustical features CLM neurons represent in behavioural tasks and how this representation changes with learning remain open questions.

We also note that the surprise formulation is relevant to the stimulus specific adaptation observed in auditory cortical areas (Ulanovsky et al. 2003) as well as in another secondary auditory region NCM in the songbird (Chew et al. 1995; Phan et al. 2006; Stripling et al. 1997). In both the mammalian and avian systems, repeated presentations of the same stimulus lead to long-term adaptation and by consequence a priming of response to novel (or deviant) stimuli. These responses could be modelled as a surprise as we have done here but with an update on what is expected given recent experience. Our current formulation assumes expectations that we labelled as naïve, meaning that they are either innate or acquired through the course of normal development. To model areas such as NCM, recently-learned expectations could be added to the model to capture memory effects. More specifically, $P(S|D)$ could be made to be experience-dependent and in this way capture the effect that repeated songs become unsurprising relative to novel song. In other words, the stimulus specific adaptation that is observed in these higher auditory areas could reflect how the stimulus prior is incorporated in the neural circuitry.

Our analysis was also useful for characterizing functional properties at the single neuron level. Specifically, we found some neurons in CLM that were sensitive to surprising features irrespective of whether they were surprisingly soft or loud. We labelled such neurons auditory complex cells because they have properties that are
analogous to those of visual complex cells in V1 (Skottun et al. 1991). They are also reminiscent of the onset-offset or phasic neurons observed in the mammalian auditory cortex (Chimoto et al. 2002; Recanzone 2000; Wang et al. 2005). We also found offset-only neurons in both Field L and CLM. Similar response properties have been described in Field L in awake birds when using high-intensity temporally-shaped noise stimuli (Nagel and Doupe 2006).

In summary, our results suggest that the firing rates of high-level sensory neurons depend more on the probability of natural stimulus features than on intensity or intensity changes. Thus, expectations and natural statistics form a key part of the neural code. Moreover, using our approach, we were able to describe the responses of single neurons that were poorly described with the classical-STRF and the derivative-STRF. Our technique of using log probabilities is motivated by information theory, and not by any considerations unique to zebra finch CLM (or to audition in general). If neurons in other forebrain areas also encode stimulus surprise, our methodology will lead to a better understanding of sensory coding not only in terms of being better able to predict spike trains, but also by revealing a new operational principle: spikes in higher sensory areas indicate stimulus surprise, not intensity changes.
Acknowledgements

The authors would like to thank Yuka Minton for outstanding animal husbandry, our research assistant (BR) for invaluable histological assistance and Taffeta Elliott for thoughtful comments on the manuscript. Furthermore, we wish to thank our anonymous reviewers for their helpful comments and suggestions.

Sarah M. N. Wooley’s present address is Department of Psychology, Columbia University, 406 Schermerhorn Hall. 1190 Amsterdam Ave, NY, NY 10027

Thané Fremouw’s present address is Department of Psychology, University of Maine, 301 Little Hall, Orono, ME 04469-5742

Grants

This work was supported by NIH grants DC007293, MH66990 and MH59189 to FET.
References.


from a large scale (n = 100) fMRI study of an auditory oddball task. Neuroimage 25: 899-915, 2005.


**Figure Legends.**

**Figure 1: The Classical-STRF and The Surprise-STRF.** A) and B) Spectrogram segments of a zebra finch song. B) shows detail from A). Rectangles highlight a stimulus element $S$ and the domain $D$ used to calculate how surprising $S$ is. C) Illustration of Equation 1. The quantified surprise of $S$ is $-\log(P(S|D))$ (see Methods and Equation 1). D) and E) Surprisingly loud (top panel) and quiet (bottom panel) elements of the zebra finch song shown in A) and B). E) shows detail from D). F) The classical-STRF obtained from reverse correlation between the PSTH and a spectrogram of the stimulus. G) The surprise-STRF, obtained from reverse correlation using the quantified surprise of the stimulus represented as in D). H) and I) Comparison of the predicted PSTHs to the actual PSTHs using: the surprise-STRF, the classical-STRF and the derivative-STRF (see Equation 2 and Methods). I) shows detail from H). Legend: PSTH is the raw PSTH smoothed with a 3 ms Hanning window, and S-STRF, D-STRF and C-STRF are jackknife predictions made by the surprise-STRF, derivative-STRF and classical-STRF, respectively. Red braces with arrows highlight areas in A) and D) where there is a surprising change within a syllable. This intra-syllable change, like the syllable onsets, elicits a response which is predictable using the surprise-STRF, as seen in H) and I).

**Figure 2: Pairwise STRF Comparisons in CLM.** A) Pairwise comparison between the performance of surprise-STRFs and derivative-STRFs for all neurons in CLM. B) and C) show scatter plots of surprise- and classical-STRFs, and of derivative- and classical-STRFs respectively. All units are in predicted information (see Methods). One neuron (with a surprise-STRF performance of 84 bits/s, a derivative-STRF performance of 70 bits/s and a classical-STRF performance of 59 bits/s) was omitted from this figure to keep the x and y axes on an appropriate scale for viewing the other data, but this neuron is included in all other analyses.
**Figure 3: Evaluating Stimulus Representations.** A) Bar graph of mean performance of different stimulus representations and different areas. Key: “Specs” = spectrograms, “Derivative” = split temporal derivative of the spectrogram (see Equation 2), “Surprise” = quantified surprise (see Equation 1). Error bars show one standard error of the mean, although the spread of the data was not Gaussian, so these error bars are not sufficient to characterize performance distributions. B) Pairwise improvements from using surprise-STRFs instead of classical-STRFs and derivative-STRFs. Error bars show one standard error, although the spread of the data was not Gaussian. P values, measured with a two-tailed Wilcoxon signed-rank test, are shown above each comparison.

**Figure 4: Comparisons in MLd, Field L and CLM of Ensemble PSTHs to Repetitions of a Motif.** A) Spectrogram of the first instance of the motif in a song. The first motif lasts from ms 818 to ms 1529 (relative to the song’s beginning), the second lasts from ms 1724 to ms 2438. B) Spectrogram of the second instance of the motif in the song. C), D) and E) Ensemble PSTHs (units are spikes per second per trial per neuron) for MLd, Field L and CLM respectively. We claim there is no large difference between PSTHs to the first and second motif repeats, suggesting these areas are not less surprised by the second repetition than by the first.

**Figure 5: Classical- and Surprise-STRFs for Four Neurons.** A) The first column shows the classical-STRF of a Field L neuron as estimated by normalized reverse correlation to a spectrogram. The middle and the right columns together show the results of the surprise-STRF model; the firing rate is modelled by the middle filter convolved with surprisingly loud events plus the filter on the right convolved with surprisingly quiet events. B), C) and D) are analogous to A), but show surprise-STRFs for CLM neurons which have an auditory complex character, in that they fire equally to
surprisingly-loud and quiet stimulus changes within the same window of latencies and frequencies.

**Figure 6: Three Representations of Modulation-Limited Noise.** A) Spectrogram of a segment of modulation-limited (ML) noise. B) Quantified surprise of this segment of ML noise using a probability model based on ML noise. C) Quantified surprise of this segment of ML noise using a probability model based on the same corpus of zebra finch song used to estimate song surprise. Since song has greater spectral and temporal correlations, the offsets quickly following onsets of ML noise chunks are particularly surprising, as are the sharp spectral edges in ML noise. Also, due to a generally higher degree of mismatch between stimulus and expectations, ML noise is more surprising when song is expected than when ML noise is expected, as can be seen comparing the scale of the colourbars in B) and C).

**Figure 7: Pairwise STRF Comparisons in CLM with ML Noise as the Stimulus.** A) Pairwise comparison between the performance of surprise-STRFs expecting song and expecting ML noise for all neurons in CLM. Key: “Classical Fit” = fit using the classical-STRF operating on the spectrogram (see Figure 6A), “S, MLN Fit” = fit using surprise-STRF expecting ML noise (see Figure 6B), “S, Song Fit” = fit using the surprise-STRF expecting song (see Figure 6C). Most points are above the line, indicating that in CLM surprise-STRFs based on the expectation of song outperform those based on the expectation of ML noise even when ML noise is the stimulus. B) and C) show scatter plots of the performance of these two surprise-STRFs compared to the performance of a classical-STRF. All units are in predicted information (see Methods). One neuron (with an S, Song Fit of 36 bits/s, an S, MLN Fit of 34 bits/s and a Classical Fit of 20 bits/s) was omitted from this figure to keep the x and y axes on an
appropriate scale for viewing the other data, but this neuron is included in all other analyses.

**Figure 8: Evaluating ML Noise Representations.** A) Bar graph of mean performance of different stimulus representations and different areas. Key: “Specs” = spectrograms, “Surp, MLN” = surprise expecting ML noise (see Figure 6B), “Surp, Song” = surprise expecting song (see Figure 6C). Error bars show one standard error of the mean, although the spread of the data was not Gaussian, so these error bars are not sufficient to characterize performance distributions. B) Pairwise improvements from using surprise-STRFs expecting song instead of classical-STRFs and surprise-STRFs expecting ML noise. Error bars show one standard error, although the spread of the data was not Gaussian. P values, measured with a two-tailed Wilcoxon signed-rank test, are shown above each comparison. The MLd comparison between the two surprise-STRFs is within one standard error of being zero, but for 88 of the 142 MLd neurons the surprise-STRF expecting ML noise outperforms the surprise-STRF expecting song, thus the Wilcoxon test is significant.
Figure 1: The Classical-STRF and The Surprise-STRF. A) and B) Spectrogram segments of a zebra finch song. B) shows detail from A). Rectangles highlight a stimulus element S and the domain D used to calculate how surprising S is. C) Illustration of Equation 1. The quantified surprise of S is $-\log(P(S|D))$ (see Methods and Equation 1). D) and E) Surprisingly loud (top panel) and quiet (bottom panel) elements of the zebra finch song shown in A) and B). E) shows detail from D). F) The classical-STRF obtained from reverse correlation between the PSTH and a spectrogram of the stimulus. G) The surprise-STRF, obtained from reverse correlation using the quantified surprise of the stimulus represented as in D). H) and I) Comparison of the predicted PSTHs to the actual PSTHs using: the surprise-STRF, the classical-STRF and the derivative-STRF (see Equation 2 and Methods). I) shows detail from H). Legend: PSTH is the raw PSTH smoothed with a 3 ms Hanning window, and S-STRF, D-STRF and C-STRF are jackknife predictions made by the surprise-STRF, derivative-STRF and classical-STRF, respectively. Red braces with arrows highlight areas in A) and D) where there is a surprising change within a syllable. This intra-syllable change, like the syllable onsets, elicits a response which is predictable using the surprise-STRF, as seen in H) and I).
Figure 2: Pairwise STRF Comparisons in CLM. A) Pairwise comparison between the performance of surprise-STRFs and derivative-STRFs for all neurons in CLM. B) and C) show scatter plots of surprise- and classical-STRFs, and of derivative- and classical-STRFs respectively. All units are in predicted information (see Methods). One neuron (with a surprise-STRF performance of 84 bits/s, a derivative-STRF performance of 70 bits/s and a classical-STRF performance of 59 bits/s) was omitted from this figure to keep the x and y axes on an appropriate scale for viewing the other data, but this neuron is included in all other analyses.
Figure 3: Evaluating Stimulus Representations. A) Bar graph of mean performance of different stimulus representations and different areas. Key: Specs = spectrograms, Derivative = split temporal derivative of the spectrogram (see Equation 2), Surprise = quantified surprise (see Equation 1). Error bars show one standard error of the mean, although the spread of the data was not Gaussian, so these error bars are not sufficient to characterize performance distributions. B) Pairwise improvements from using surprise-STRFs instead of classical-STRFs and derivative-STRFs. Error bars show one standard error, although the spread of the data was not Gaussian. P values, measured with a two-tailed Wilcoxon signed-rank test, are shown above each comparison.
Figure 4: Comparisons in MLd, Field L and CLM of Ensemble PSTHs to Repetitions of a Motif. A) Spectrogram of the first instance of the motif in a song. The first motif lasts from ms 818 to ms 1529 (relative to the song's beginning), the second lasts from ms 1724 to ms 2438. B) Spectrogram of the second instance of the motif in the song. C), D) and E) Ensemble PSTHs (units are spikes per second per trial per neuron) for MLd, Field L and CLM respectively. We claim there is no large difference between PSTHs to the first and second motif repeats, suggesting these areas are not less surprised by the second repetition than by the first.
Figure 5: Classical- and Surprise-STRFs for Four Neurons. A) The first column shows the classical-STRF of a Field L neuron as estimated by normalized reverse correlation to a spectrogram. The middle and the right columns together show the results of the surprise-STRF model; the firing rate is modelled by the middle filter convolved with surprisingly loud events plus the filter on the right convolved with surprisingly quiet events. B), C) and D) are analogous to A), but show surprise-STRFs for CLM neurons which have an auditory complex character, in that they fire equally to surprisingly-loud and quiet stimulus changes within the same window of latencies and frequencies.

127x130mm (600 x 600 DPI)
Figure 6: Three Representations of Modulation-Limited Noise. A) Spectrogram of a segment of modulation-limited (ML) noise. B) Quantified surprise of this segment of ML noise using a probability model based on ML noise. C) Quantified surprise of this segment of ML noise using a probability model based on the same corpus of zebra finch song used to estimate song surprise. Since song has greater spectral and temporal correlations, the offsets quickly following onsets of ML noise chunks are particularly surprising, as are the sharp spectral edges in ML noise. Also, due to a generally higher degree of mismatch between stimulus and expectations, ML noise is more surprising when song is expected than when ML noise is expected, as can be seen comparing the scale of the colourbars in B) and C).
Figure 7: Pairwise STRF Comparisons in CLM with ML Noise as the Stimulus. A) Pairwise comparison between the performance of surprise-STRFs expecting song and expecting ML noise for all neurons in CLM. Key: Classical Fit = fit using the classical-STRF operating on the spectrogram (see Figure 6A), S, MLN Fit = fit using surprise-STRF expecting ML noise (see Figure 6B), S, Song Fit = fit using the surprise-STRF expecting song (see Figure 6C). Most points are above the line, indicating that in CLM surprise-STRFs based on the expectation of song outperform those based on the expectation of ML noise even when ML noise is the stimulus. B) and C) show scatter plots of the performance of these two surprise-STRFs compared to the performance of a classical-STRF. All units are in predicted information (see Methods). One neuron (with an S, Song Fit of 36 bits/s, an S, MLN Fit of 34 bits/s and a Classical Fit of 20 bits/s) was omitted from this figure to keep the x and y axes on an appropriate scale for viewing the other data, but this neuron is
included in all other analyses.
40x104mm (600 x 600 DPI)
Figure 8: Evaluating ML Noise Representations. A) Bar graph of mean performance of different stimulus representations and different areas. Key: Specs = spectrograms, Surp, MLN = surprise expecting ML noise (see Figure 6B), Surp, Song = surprise expecting song (see Figure 6C). Error bars show one standard error of the mean, although the spread of the data was not Gaussian, so these error bars are not sufficient to characterize performance distributions. B) Pairwise improvements from using surprise-STRFs expecting song instead of classical-STRFs and surprise-STRFs expecting ML noise. Error bars show one standard error, although the spread of the data was not Gaussian. P values, measured with a two-tailed Wilcoxon signed-rank test, are shown above each comparison. The MLd comparison between the two surprise-STRFs is within one standard error of being zero, but for 88 of the 142 MLd neurons the surprise-STRF expecting ML noise outperforms the surprise-STRF expecting song, thus the Wilcoxon test is significant.