Representations of Conspecific Song by Starling Secondary Forebrain Auditory Neurons: Toward a Hierarchical Framework

C. Daniel Meliza,1 Zhiyi Chi,2 and Daniel Margoliash1

1Department of Organismal Biology and Anatomy, University of Chicago, Chicago Illinois; and 2Department of Statistics, University of Connecticut, Storrs, Connecticut

Submitted 28 May 2009; accepted in final form 22 December 2009

Meliza CD, Chi Z, Margoliash D. Representations of conspecific song by starling secondary forebrain auditory neurons: toward a hierarchical framework. J Neurophysiol 103: 1195–1208, 2010. First published December 23, 2009; doi:10.1152/jn.00464.2009. The functional organization giving rise to stimulus selectivity in higher-order auditory neurons remains under active study. We explored the selectivity for motifs, spectrotemporally distinct perceptual units in starling song, by recording the responses of 96 caudomedial mesopallium (CMM) neurons in European starlings (Sturnus vulgaris) under awake-restrained and urethane-anesthetized conditions. A subset of neurons was highly selective between motifs. Selectivity was correlated with low spontaneous firing rates and high spike timing precision, and all but one of the selective neurons had similar spike waveforms. Neurons were further tested with stimuli in which the notes comprising the motifs were manipulated. Responses to most of the isolated notes were similar in amplitude, duration, and temporal pattern to the responses elicited by those notes in the context of the motif. For these neurons, we could accurately predict the responses to motifs from the sum of the responses to notes. Some notes were suppressed by the motif context, such that removing other notes from motifs unmasked additional excitation. Models of linear summation of note responses consistently outperformed spectrotemporal receptive field models in predicting responses to song stimuli. Tests with randomized sequences of notes confirmed the predictive power of these models. Whole notes gave better predictions than did note fragments. Thus in CMM, auditory objects (motifs) can be represented by a linear combination of excitation and suppression elicited by the note components of the object. We hypothesize that the receptive fields arise from selective convergence by inputs responding to specific spectrotemporal features of starling notes.

INTRODUCTION

In many species, social interactions are supported by individuals exchanging information via vocalizations. Processing of vocalizations has long been thought to be mediated by a functional hierarchy (Leppelsack and Vogt 1976; Roeder 1966; Wilczynski and Capranica 1984), with higher-order auditory forebrain neurons responding selectively to complex, often conspecific stimuli (see Gentner and Margoliash 2002; Rauschecker 1998). It has been suggested that the identity of behaviorally relevant signals is represented by coordinated activity in small groups of secondary neurons (Barlow 1969; Olshausen and Field 2004; Theunissen 2003), but little is known about how these representations emerge from earlier stages.

For the songbirds, vocalizations include repertoires of relatively acoustically simple calls and one or more songs, which can be acoustically complex. European starlings, Sturnus vulgaris, are highly social songbirds with diverse vocal repertoires (Feare 1984; Hausberger 1997). Starlings recognize conspecific individuals by their songs, which are long bouts of temporally discrete “motifs” (Adret-Hausberger and Jenkins 1988; Eens et al. 1989). Motifs comprise sequences of “notes,” which appear in spectrograms as contiguous regions of high power and may represent basic motor gestures (Gardner et al. 2001). Starlings combine notes to form motifs that are almost always unique to a particular bird (Eens 1997). Field (Eens et al. 1991) and laboratory operant (Gentner and Hulse 2000b) behavioral studies indicate that starlings associate different motifs with particular individuals, suggesting that motifs are the basic perceptual objects of song recognition.

Here, we explore the neuronal mechanisms for representing motifs by examining the responses of neurons in the caudomedial mesopallium (CMM), a secondary forebrain auditory area that may be involved in the song recognition process. Auditory thalamic afferents synapse primarily onto field L2a and L2b neurons (Karten 1968; Wild et al. 1993), which in turn project only indirectly to CMM (Vates et al. 1996). Whereas many field L neurons respond broadly to natural auditory stimuli (Bonke et al. 1979; Grace et al. 2003; Leppelsack and Vogt 1976; Lewicki and Arthur 1996; Margoliash 1986), CMM neurons are substantially more selective, responding to fewer stimuli in a repertoire of natural sounds (Gentner and Margoliash 2003; Müller and Leppelsack 1985; Theunissen et al. 2004). Training starlings to discriminate between conspecific songs can bias this selectivity toward the motifs in the learned songs (Gentner and Margoliash 2003). The models developed for CMM neurons to date (Gentner and Margoliash 2003; Sen et al. 2001) indicate that the receptive fields are highly nonlinear and are unable to describe how the selectivity emerges from lower levels in the auditory processing hierarchy.

Little is known about auditory processing in the areas between field L and CMM, but receptive fields at higher areas are likely built up from the simpler receptive fields in field L, which are matched to the spectrotemporal features of conspecific song (Woolley et al. 2005). We tested the hypothesis that CMM neurons are tuned to complex elements commonly found in conspecific song, measuring their responses to starling motifs and the component notes. We found that CMM neurons responded robustly to motifs and notes. For most neurons, the responses to motifs could be predicted as a linear sum of the excitatory responses to the notes.

Address for reprint requests and other correspondence: C. D. Meliza, Dept. of Organismal Biology and Anatomy, Univ. of Chicago, 1027 E 57th St., Chicago, IL 60637 (E-mail: dmeliza@uchicago.edu).
METHODS

The subjects used in this study were seven adult European starlings (sex was not recorded), captured from farms in northeastern Illinois. Birds were housed in mixed-sex flight aviaries, where they received food and water ad libitum. The light schedule in the aviary was adjusted to match the local day length in Chicago. All animal procedures were performed according to protocols approved by the University of Chicago Institutional Animal Use and Care Committee and consistent with the guidelines of the National Institutes of Health.

Starling songs and motif stimuli

Starling song consists of a sequence of complex elements about a second in duration, called motifs (Fig. 1A). Although motif structure is highly variable, three major classes have been described: the variable motifs, the rattles, and the high-frequency motifs (Fig. 1, B–D) (Eens 1997), which consistently occur in that order in songs. The motifs used in this study consisted of 54 unique motifs that were visually identified and extracted from a library of starling song (see Supplemental Materials and Methods; for the complete set of motifs, see Supplemental Fig. S1).1 Stimuli derived from these songs were presented free-field in a sound isolation chamber.

Spectrotemporal isolation of note components of motifs

The most elementary components of birdsong are typically referred to as notes. In spectrographic representations of song, notes appear as regions of high power that are contiguous in time and frequency (Fig. 1, B–D). A tonal sound, for example, appears as a horizontal line, whereas a click or other sharp transient will appear as a vertical line (e.g., Fig. 1, notes e,f). Although birds often modulate the frequency of tones, they generally do so smoothly, so sweeps (Fig. 1, notes a,h) and trills (Fig. 1, note d) appear as continuous, curved lines in the spectrogram. Some notes are harmonic (Fig. 1, notes b,g). Starlings are also capable of producing screeches and other sounds that consist of broadband or band-limited noise; these appear as regions of diffuse power extended in both time and frequency (Fig. 1, note c).

We used a novel method to segment starling motifs into constituent notes and extract the corresponding sound pressure waveforms. Notes can overlap in time (Fig. 1B), but nevertheless, they occupy disjoint regions of the spectrogram (Fig. 2A). This permitted us to extract notes by spectrographic inversion techniques (Supplemental Materials and Methods). Extracted notes have spectrograms that optimally match (in least-squares distance) the notes in the original motif (Fig. 2B). In this report, we presented neurons with isolated notes. We also presented stimuli in which multiple notes were recombined at their original temporal offsets. These included motif reconstructions, in which all the notes of a motif were included (Fig. 2C), and note deletions, in which one of the notes was omitted from the reconstruction.

In a second set of experiments, we presented neurons with “note noise” and “fragment noise.” Note noise consisted of all the notes extracted from a 10-s segment of song recombined in random order with gaps of 10 ms between each note. Five different permutations of the notes were used, so that responses were recorded to each note in a variety of different contexts. Fragment noise was generated in the same way, but the motifs were divided into nonoverlapping note fragments (Supplemental Materials and Methods). The fragments contained the same spectrotemporal structures as the notes, but the spectral and temporal extent of each fragment was such that few fragments contained a single, whole note.

Electrophysiology

Birds were prepared for serial restrained recordings (Supplemental Materials and Methods). We made serial recordings from 27 units in five birds under restrained conditions over the course of three to eight sessions. The animals were restrained in a soft cloth jacket and attached to a stereotaxic device via the head pin. The sessions were timed to coincide with the birds’ sleep period, lasted no longer than 5 h each, and were terminated if the animal became excessively active. Trials with movement artifacts were excluded from analysis.

An additional 45 units were recorded from four of the birds in terminal sessions under urethane anesthesia (20% by volume, 5 ml/kg, im). We observed significant differences between behavioral states in spike time precision (see Supplemental Materials and Methods for definition) and some of the results, but basic response properties were

---

1 The online version of this article contains supplemental data.
Neuronal response analysis

The firing rate response (FR) of a neuron to each motif was defined as the number of spikes per second during the interval from the onset of the motif until 100 ms after its offset, averaged across trials. Neurons were considered responsive if, for at least one motif, FR was significantly greater than the spontaneous firing rate measured during a 1,000-ms baseline period as the contrasting condition. Neurons with low firing rates often do not satisfy the Gaussian assumptions of ANOVA, so we used the empirical quantiles of the FR distribution at 5% intervals.

Stimulus-driven spike precision was defined as the median (across stimuli) of the maximal frequency at which the neuron showed significant intertrial coherence (Supplemental Materials and Methods).

Linear summation of CMM responses to notes

The smoothed response to each note, \( r(t) \), was calculated by convolving the spike trains with a Gaussian smoothing kernel (\( \sigma = 10 \) ms), averaging across presentations, and subtracting the baseline firing rate. The response window was the interval from the start of the note to 250 ms after the end of the note. We used a discrete-time framework with a step size of 5 ms. To predict the response of the neuron to the full motif, we added the responses to each of the \( N \) notes in the motif, aligned with the time when each note occurred (\( t_i \)).

\[
    r(t) = \sum_{i=0}^{N} r_i(t - t_i) \quad (1)
\]

The predicted responses were compared with the neuron’s actual response by computing the correlation coefficient ratio (CCR), which incorporates a correction for the intrinsic variability of the spiking responses (Hsu et al. 2004). In brief, CCR is determined by first calculating the intertrial correlation, defined as the correlation between the spiking responses of the neuron and \( R(t) \), the mean response. Because \( R(t) \) is not known, the intertrial correlation is calculated from the correlation coefficient between the averages of the even and odd trials, using a finite sample size correction. Then the correlation coefficient between the predicted response and the neuronal response is calculated (also with a finite sample size correction) and divided by the intertrial correlation.

Modeling CMM responses with feature receptive fields

The summation described in the previous section can be generalized to any stimulus composed of discrete elements. Each element is represented as a time-varying binary variable that indicates when it is present in a stimulus. A motif or song of length \( T \), comprising \( N \) notes with onset times \( t_1, \ldots, t_N \), is represented as \( M(i, t) = 8(t_i - t) \) for \( 1 \leq i \leq N \). This is analogous to a spectrogram, except that notes are used as a basis set instead of frequencies. The responses of a neuron to an arbitrary combination of notes can be modeled as a sum of the responses to the components, or equivalently as the convolution of the stimulus matrix with the note responses

\[
    r(t) = \sum_{i=0}^{T} r_s(i, t - s)ds + \mu \quad (2)
\]

where \( r(t) \) is the response to note \( i \), \( \mu \) is the background firing rate, and \( S_i \) is the largest time lag associated with each note. This model was implemented on a uniform time grid with a step size of 5 ms.

It can be seen that Eq. 1 is a special case of Eq. 2, in which notes are presented in isolation to directly determine \( r(t) \), and \( S_i \) is chosen to include time points \( \leq 250 \) ms after the offset of each note. However, the model can also be used to estimate responses to notes when they are not presented in isolation. In the first set of experiments, neurons were also presented with note deletions, which are motif reconstructions from which single notes were omitted. The stimulus representations \( M(i, t) \) for these stimuli are the same as for the unmodified motif, except that they are zero for the omitted note. Fitting Eq. 2 to these responses gives estimates of the responses to the notes in the context of the motif, which we used to calculate the effects of context on note responses (Supplemental Materials and Methods).
In a second set of experiments, we presented neurons with random sequences of notes and note fragments (“note noise” and “fragment noise”). This allowed us to test a much larger number of components. Each note (or fragment) appears in a variety of contexts and may be affected by suppressive or excitatory effects from preceding notes. The ordinary least squares solution accounts for these interactions and provides an estimate of the mean effect of each note (or fragment) independent of its context. In light of the similarity to the spectrotemporal receptive field (STRF), we called the responses of a neuron to each of the notes, \( r_i(t) \) for all \( i \), the neuron’s feature receptive field (FRF). The FRF could be used to predict the response to the original song segment using Eq. 2.

All statistical tests are two-tailed unless otherwise noted. For linear regressions, where unequal variance or nonnormality of the data were a concern, we calculated robust standard errors for regression coefficients (using the sandwich package; Zeileis 2004) and report the associated \( t \) value.

### RESULTS

We recorded from single units in starling CMM and examined their responses to motifs and notes, which are the basic components of starling song (Fig. 1). Neurons were presented with a set of 39 to 54 unique motifs (median 10 repeats per motif). In the first experiment, out of a total of 72 well-isolated units, recorded from five adult starlings, 61 exhibited a significant excitatory response to at least one of the motifs in the test set (ANOVA on firing rates; \( \alpha = 0.01 \)). Another two neurons showed no significant differences in mean firing rate but did exhibit reliable time-varying responses to several motifs, so they were also included in further analysis.

#### CMM responses and motif selectivity

We first characterized the general response properties and selectivity of CMM neurons. Responsive neurons varied extensively in their spontaneous firing rate (range, 0.13–18.3 Hz), spike patterns, and selectivity. Some neurons responded to motifs with one or more transient peaks of activity at high temporal precision (Fig. 3A), whereas others responded with sustained increases in firing rate at low temporal precision (Fig. 3B). Some neurons were highly selective, responding strongly to only a few motifs. This can be seen in the cumulative distribution of firing rates across the motifs, with most of the motifs eliciting weak responses relative to the peak of the distribution (Fig. 3C, red). Other neurons were less selective, with responses more evenly distributed across the stimuli (Fig. 3C, blue). To summarize the firing rate distributions, we used a sparseness index (SI) (Olkhausen and Field 2004) (see Methods), which indicates how similar the neuron’s response is across the set of motifs. SI can range between 0, for a neuron that responds the same to all stimuli, to 1, for a neuron that responds only to one stimulus. In CMM, selectivity was broadly distributed (median SI = 0.21, interquartile range 0.10–0.44, \( n = 63 \) neurons) and...
inversely correlated with spontaneous firing rate (Fig. 3D). A group of 23 neurons (36.5% of responsive neurons) characterized by higher selectivity and lower spontaneous activity seemed to form a distinct cluster.

Selectivity and spontaneous firing rate were correlated with the spike waveforms of the neurons. Most of the units recorded in CMM (45/63) were characterized by spikes with an initial positive deflection, followed by a much broader, shallow repolarization (Fig. 4C, left). The other neurons (17/63) tended to have narrower positive deflections and deeper, faster repolarizations (Fig. 4C, right). Spike shapes for these neurons were more heterogeneous than the wide spikes and probably originated from more than one type of neuron. We classified the spikes based on trough width and the ratio of peak to trough amplitude (Fig. 4A). One neuron had an exceptionally strong signal that clipped during data acquisition and was not included in this analysis. Neurons with wide spikes were significantly more selective (median SI = 0.31) than neurons with narrow spikes (median SI = 0.095; Wilcoxon rank-sum test: P = 3.64 × 10^-5, n = 62) and had lower spontaneous firing rates (median 1.50 Hz vs. 5.95 Hz; Wilcoxon rank-sum test: P = 1.4 × 10^-6). All but 1 of the 23 highly selective neurons fell into the wide spike class (Fig. 4B). The remaining selective neuron, which was recorded on an electrode with particularly high impedance, had a complex spike waveform with a smaller positive and negative deflection before the main peak of the spike (Fig. 4C, green trace). Additional comparisons between neurons with wide and narrow spikes are given in Supplemental Table S2.

Neurons in CMM exhibited a broad range of temporal precision and reliability in their responses (e.g., Fig. 3, A vs. B). We used intertrial coherence to quantify the reliability of the spiking responses at multiple timescales (Fig. 5, A and B). The spike timing precision for each neuron was defined as the maximum frequency at which the neuron showed significant intertrial coherence (median across motifs; Supplemental Materials and Methods). The spike timing precision of the responsive CMM neurons had a median of 8.74 Hz (interquartile range, 6.20–15.3 Hz). Spike timing precision was weakly correlated with selectivity (Fig. 5C; Spearman rank correlation: ρ = 0.34, P = 0.007) and uncorrelated with spontaneous firing rate (Spearman rank correlation: ρ = −0.17, P = 0.18). Within the subset of neurons with wide spikes, the correlation with selectivity was even higher (Fig. 5C, red points; Spearman rank correlation: ρ = 0.65, P = 1.4 × 10^-6). We also observed that neurons in unanesthetized animals showed significantly higher spike timing precision (median 13.2 Hz in unanesthetized animals vs. 7.1 Hz under urethane; Wilcoxon rank-sum test: P = 0.004). Neurons recorded in unanesthetized animals tended to be more selective as well, although the difference was not significant (median SI 0.36 unanesthetized vs. 0.16 urethane; Wilcoxon rank-sum test: P = 0.19). Additional comparisons between the response properties of neurons recorded under restrained or anesthetized conditions are given in Supplemental Table S1.

Equivalent excitatory responses to isolated notes and notes in motifs

We asked how CMM neurons responded to more elementary components of starling song. We segmented motifs into notes, which are stereotyped vocal structures that appear in spectro-
grams as contiguous regions of high-intensity (Figs. 1, B–D, and 2A). The sound waveforms associated with these structures were extracted by masking out the rest of the motif spectrogram and computing the inverse of the spectrographic operation on the unmasked region (Supplemental Materials and Methods). The extracted notes had spectrograms that closely matched the notes in the full motif (Fig. 2B) and could be combined to reconstruct the motif (Fig. 2C) or to generate novel synthetic stimuli.

We tested the responses of 49 CMM neurons to isolated notes from 1 to 6 excitatory motifs, for a total of 39 unique motifs and 104 different motif-cell pairs. By comparing the responses to individual notes with those elicited by the original motif, we could determine which notes contributed to the response. We also recorded each neuron’s response to the reconstructed motif (e.g., Fig. 2C) as a control for the quality of the note extraction algorithm. CMM neurons responded robustly to the motif reconstructions and to many of the isolated notes. Figure 6A shows the responses of an example neuron to a motif, its reconstruction, and all the component notes. The motif and the reconstruction elicited responses with statistically indistinguishable spike counts (median firing rate 19.0 vs. 18.0 Hz; Wilcoxon rank-sum test: $P = 0.09$). The temporal patterns of the responses were highly similar, which we quantified using correlation coefficient ratio ($CCR = 0.98$; see METHODS). The neuron also

**FIG. 5.** Spike timing precision of CMM neurons. $A$: raster plots of 2 neurons, responses to the same motif (same as in Fig. 3, A and B). Spike precision was analyzed over the interval from the onset of the motif to 100 ms after the offset (black bars). $B$: intertrial coherence of the responses in $A$ for unit A (red) and unit B (blue). Vertical lines indicate the highest frequency in a band starting at 0 Hz where the coherence is significantly $>0$; this is the smallest timescale over which the response is reliable. $C$: scatter plot of selectivity between motifs (SI) and spike timing precision for 62 responsive CMM neurons. The spike timing precision of a neuron is the median of the maximum coherent frequency (as in $B$) for all the motifs presented to the neuron. Symbol colors indicate the spike waveform type (see Fig. 4), and symbol shapes indicate the behavioral state of the animal. Example units A and B are labeled with capital letters, and the open symbol corresponds to the neuron with the unusual spike waveform (Fig. 4C).

**FIG. 6.** Responses of CMM neuron to motif and isolated notes. $A$: top: spectrogram of a motif, with the notes outlined in different colors. Plotted below the spectrogram are the responses of the neuron to the original motif, the note-based reconstruction (recon) and to each of the component notes, which are identified by the numbers of the notes and the colors of the note outlines. Raster plots are aligned so they show the response to each note at its original temporal offset in the motif. $B$: comparison of the responses of the neuron to the notes and the motif. $Top$: the smoothed response to each note is plotted in the color corresponding to the note and at the time when the note occurs in the motif. $Bottom$: the note responses have been summed (red) for comparison with the response to the motif (black). The correlation coefficient ratio ($CCR = 0.98$; see METHODS). The neuron also
responded to six of the eight notes. When we aligned the responses to each note with the temporal position of the note in the motif, we found that three of the notes (Fig. 6A, notes 2, 4, and 7) elicited transient responses that aligned with peaks in the full response and were similar in amplitude and duration. These included the notes that elicited the strongest responses (Fig. 6A, notes 4 and 7). For neurons that gave sustained responses with no distinct peaks of activity, for example in Fig. 7A, several of the notes elicited responses that corresponded with the firing rate at that point in the response to the full motif. The combined responses to the sequence of individual notes seemed to fill the response to the entire motif.

These observations were consistent across the motifs and neurons we tested in CMM. Responses to note reconstructions and the original motifs were reliably similar (median CCR 0.92; interquartile range 0.84–0.97; n = 104 motif-neuron pairs). The similarity of the responses was higher in unanesthetized birds (median CCR 0.96 vs. 0.90; Wilcoxon rank-sum test: P = 0.01), perhaps because of greater spike precision in this behavioral state. The component notes of motifs consistently elicited robust responses that could be summed to predict the response to the original motif (Figs. 6B and 7B). The median CCR between responses to the motif and the summed responses to notes was 0.59 (interquartile range 0.28–0.72; n = 104 pairs). This linearity depended to some extent on how well the note extraction algorithm worked; that is, the CCR between the response to a motif and the sum of the responses to the component notes was correlated with the CCR between the responses to a motif and its reconstruction (Fig. 8).

Thus we have probably somewhat underestimated the extent of the linear summation of notes.

Suppressive interactions between notes

Although most notes elicited similar responses in isolation and in the context of the motif, we observed a number of instances where responses were different in these contexts. For instance, some notes elicited excitatory responses in isolation that had no correspondence in the response to the full motif (Figs. 6A, notes 5 and 6, and 9A, notes 7 and 11). This
suggested that some other part of the motif suppressed the responses to these notes. The low spontaneous rate of many CMM neurons prevented observing suppressive effects directly. Instead, we presented neurons with synthetic motifs comprising all but one of the notes (i.e., note deletions), and found that removing notes could unmask excitatory responses during or immediately following the note that was removed. As seen in Fig. 9A (notes 1, 4, 9, and 10), the unmasked excitation often coincided with the responses to the isolated notes (i.e., with times where the sum of note responses overestimates the response to the motif).

To further characterize this phenomenon, we used a linear model to estimate the responses to notes in the context of the motif, using the responses to the note deletions to fit the model (see METHODS). We compared the context-dependent response to the response elicited by the isolated note. As shown in Fig. 9B, the effects of motif context varied extensively. Most notes elicited slightly weaker responses in the motif, but with similar temporal profiles (Fig. 9B, notes 2, 5, 11, 12, and 13). Some notes were completely suppressed (Fig. 9B, note 7). Others were facilitated by the motif context (Fig. 9B, note 6), and a few notes that were weakly excitatory or suppressive in isolation had strong effects of the opposite sign on the response to the full motif (Fig. 9B, notes 3, 4, and 10).

These diverse contextual effects were observed across CMM. For each note, we computed a context dependence index \( \text{CD} \), which is the difference between the excitatory response to the isolated note and the excitatory response to the note in the context of the motif, relative to the larger of the two responses (Supplemental Materials and Methods). \( \text{CD} \) ranges between \(-1\) and 1 and is negative for suppressive effects and positive for facilitative effects. Nineteen percent of the notes showed significant facilitation by context, and 35% of the notes were suppressed by context.
affected the distribution of suppression and facilitation (Supplementary Fig. S3; \( \chi^2 \) test for both factors: \( P = 1.4 \times 10^{-5} \)). In unanesthetized animals, there was a significantly higher proportion of facilitated notes (26 vs. 16\% in anesthetized animals), and neurons with narrow spikes showed significantly lower facilitation (12\% of notes vs. 22\% for wide-spike neurons).

We also computed \( CD \) for each motif-neuron pair by comparing the mean of the excitatory responses to all the isolated notes against the mean of all the context-dependent responses (e.g., Fig. 9B, mean excitation). Most motifs had an overall suppressive effect on note responses (median \( CD \), \(-0.17\); interquartile range, \(-0.38\) to \(-0.06\); \( n = 104 \) pairs). Low absolute values of \( CD \) indicate that the responses to notes tend to sum linearly, and unsurprisingly, the responses to motifs with low \( CD \) magnitudes tended to be predicted well from the sum of the responses to notes (Fig. 10A; Spearman rank correlation of \( CD \) magnitude and \( CCR \), \( \rho = -0.29 \), \( P = 0.002 \)). There was also a tendency for less selective neurons to be dominated by suppressive interactions (negative \( CD \)) and more selective neurons to be dominated by facilitative interactions [Fig. 10B; Spearman rank correlation of \( CD \) (median across motifs) and SI, \( \rho = 0.24 \), \( P = 0.02 \)], although the relationship between context dependence and selectivity appears to be substantially nonlinear.

**Feature receptive fields**

The linear summation of note responses suggests a model for CMM responses similar to a STRF, but using notes as a basis set rather than frequencies. In a second experiment, we therefore recorded responses from an additional 24 CMM neurons (in 2 birds) presented with a novel stimulus comprising all the notes from a 10-s segment of song presented in random order, which we termed “note noise.” In this model, which we termed the FRF, each note was encoded by a binary variable, which was equal to 1 at the onset of the note and 0 elsewhere, an extremely sparse representation. We fit the model to the responses to note noise to get a linear estimate of the response to each note (the FRF; Fig. 11A). Each row in the FRF corresponds to a particular note. To predict the response to the original song, the FRF is convolved with the note-based representation of the song (Fig. 11D, 3rd panel), which amounts to a linear summation of the responses to the notes at their original temporal offsets. The FRF model consistently predicted the responses of neurons in CMM to startling song (median \( CCR \), 0.62; interquartile range, 0.49–0.67). FRFs for each of the neurons in this experiment are shown in Supplementary Fig. S5.

We also used the FRF model to test whether neurons are specifically tuned to notes. We made an alternative segmentation of the song into “note fragments,” which were nonoverlapping temporal slices of the song, with the endpoints of the fragments chosen in such a way to cut as many notes as possible into different fragments, while keeping the number of features approximately the same (see METHODS for details). The same neurons were presented with “fragment noise,” and the model was fit to these data. The resulting fragment FRFs (Fig. 11B) were nearly as good as the note FRFs in predicting the responses to the song (Fig. 11D, 4th panel). The median difference in \( CCR \) for the predictions of the note-based and fragment-based models was 0.07 (Wilcoxon signed-rank test: \( P = 0.002 \)).

To compare these results with previous findings in CMM and other auditory areas, we also computed STRFs for each of the 24 neurons in this second study. We used maximally informative dimensions (MIDs), a method that eliminates several of the statistical issues in estimating receptive fields from natural stimuli (Atencio et al. 2008; Sharpee et al. 2004), to estimate the spectrotemporal filter for each neuron and an associated static nonlinearity. Only the first MID was calculated. The data used to fit the model were the responses to the note and fragment noise. In general the STRFs for CMM neurons tended to be complex, with multiple subfields (Fig. 11C; see Supplemental Fig. S5 for the STRFs and FRFs of all the neurons). STRFs gave less accurate predictions of the response to songs, with a median \( CCR \) of 0.44 (interquartile range, 0.30–0.53). For all but two units, the note FRF gave a better prediction than the STRF (Fig. 12; median difference in \( CCR \) = 0.12; Wilcoxon signed-rank test: \( P = 0.0002 \)). Thus although MID estimates of the spectrotemporal response properties of CMM neurons give substantially better results than reported using other statistical methods (Gentner and Margoliash 2003; Sen et al. 2001), significant spectrotemporal nonlinearities remain that are captured by the feature-based models.

**DISCUSSION**

The structure of vocalizations, especially complex vocalizations, is suggestive of hierarchical organization, in which objects...
at one level are composed of simpler elements. Vocal elements at an intermediate level of the hierarchy, such as words in speech or motifs in starling song, may be perceived as units. An attractive assumption is that representations of such auditory objects may follow combinatoric principles, being defined by spectral and temporal combinations of smaller, often distinct vocal elements. In this encoding scheme, changing components or varying their spectrotemporal relationships can lead to sharp changes in the perceived identity of the object (Liberman et al. 1961; Nelson and Marler 1989; Searcy et al. 1999). There is some electrophysiological evidence in support of this hypothesis (Margoliash 1983; Prather et al. 2009; Suga 1978).

To address this question, we decomposed starling motifs into notes, using a method that was partially heuristic. We found that neurons in the starling higher-order auditory region CMM respond robustly to notes thus defined. For most neurons, the excitatory responses to notes were surprisingly independent of the motif context in which they were normally expressed. Suppressive effects of notes in the context of motifs were also observed. On average, notes were better at driving neurons than note fragments. Our results support a hierarchical model whereby selectivity for simple auditory features emerges prior to CMM in the auditory processing hierarchy, with CMM neurons combining these inputs to create note-level representations.

Selectivity for complex auditory objects in CMM

Motifs are thought to be the primary perceptual objects of starling song recognition (Gentner and Hulse 2000b), although note-level perception has yet to be examined. We observed that in CMM, selectivity between motifs varied broadly among neurons (Gentner and Margoliash 2003). A small number of
patterns of activity elicited by the notes when they were presented in isolation. The responses closely matched the to one or more of the component notes of motifs when they were operated train-

Linear summation of note responses

Both selective and nonselective neurons responded robustly to one or more of the component notes of motifs when they were presented in isolation. The responses closely matched the patterns of activity elicited by the notes when they were embedded in the motifs. We could predict the responses of CMM neurons to motifs with a high degree of accuracy by simply adding together the responses to the component notes. This suggested a linear model for CMM responses, which we termed the FRF. We estimated FRFs for 24 additional CMM neurons from their responses to note noise, which consisted of the component notes of a song presented in random order, and found that they could predict responses to the original song (Fig. 12). Taking the square of CCR as the percentage of explained variance, FRFs accounted for ~38% of the stimulus-dependent variance in the responses to the original song. In contrast, STRFs calculated from the same data only accounted for 19% of the variance.

STRF models represent neuronal responses as a linear (or statically nonlinear) function of the spectrotemporal content of the stimulus (Aertsen and Johannesma 1981; Eggermont et al. 1983; Theunissen et al. 2001). If there are nonlinear interactions between components of the stimulus, STRFs generally fail to make good predictions (Ahrens et al. 2008; Christianson et al. 2008; Sharpee et al. 2004, 2008). For instance, if a neuron is selective for a pair of tones, a linear model will predict a weak response to either of the tones by itself, and fail to predict the nonlinear dependence on the combination. A similar argument obtains for contiguous features with complex spectrotemporal modulations or contours. Higher-order auditory neurons often exhibit temporal or spectral combination sensitivity (Bar-Yosef et al. 2002; Margoliash 1983; Margoliash and Fortune 1992; O’Neill and Suga 1979; Peña and Konishi 2001). However, if a combination-sensitive neuron is presented with stimuli in which the pair of tones (or other complex structure) is manipulated as a unit, its response will be a linear function of whether that unit is present or not. Using the MID method to estimate STRFs gave more accurate predictions than previously obtained with other methods (Gentner and Margoliash 2003; Sen et al. 2001), but the much better performance of the FRF model indicates that substantial spectrotemporal nonlinearities remain. Dividing starling song into notes kept intact the complex spectrotemporal patterns driving the responses.

One implication of this interpretation is that the FRF can be used to test how well a particular basis set preserves the structures the neuron is tuned to. The responses to the bases of the model will only sum linearly if the excitatory components of the stimuli are not split across multiple bases. Our segmentation of starling song was based on visual inspection of the spectrograms and can only approximate the actual perceptual or motor structure of the song. However, the FRF based on notes outperformed a FRF model based on note fragments (Fig. 13), which suggests that CMM neurons are preferentially tuned to the spectrotemporal patterns contained in notes. This is consistent with the likely perceptual experience of starlings. They would hear the same notes in different motifs and under variable auditory conditions. Experience-dependent plasticity such as that observed in auditory cortex (Dahmen and King 2007) could lead to selectivity for the spectrotemporal patterns found in notes.

Although FRF models gave better performance than did STRF models, the limitations of FRFs compared with STRFs should be clearly noted. FRFs as we have described them here are limited to predicting responses to the repertoire of notes that were tested. We have yet to extend this approach so as to predict responses to arbitrary notes. In distinct contrast, a
STRF model that is a good prediction of a neuron’s response will, in principle, predict the response to any stimulus. There may be a basis set embedded with starling notes, but if so we have yet to delineate it.

Correspondingly, it is not yet clear which aspects of the complex tuning of CMM neurons emerge within CMM and which are inherited from its afferents. Little is known about receptive field properties of the neurons that project to CMM in the caudolateral mesopallium (CLM) and the caudomedial nidopallium (NCM). Gene expression studies indicate that different vocal gestures from canary song may form distinct representations in NCM (Ribeiro et al. 1998). In the zebra finch, neurons in the L1 and L3 subregions of field L tend to be tuned to simple spectrotemporal modulations (i.e., onsets and offsets, pure tones, and simple frequency sweeps) that are common in conspecific song (Sen et al. 2001; Woolley et al. 2005). In starlings, many units in these areas respond selectively to species-specific features of vocalizations (Bonke et al. 1979; Leppelsack and Vogt 1976). Convergence by afferents tuned to simpler, slightly different structures is thought to account for the emergence of complexity in a number of systems (Hubel and Wiesel 1962; Jortner et al. 2007; Kobatake and Tanaka 1994; Riesenhuber and Poggio 2002). Thus further convergence of afferents to NCM and CLM, as well as nonlinearities such as high spike thresholds (Escabi et al. 2005), nonlinear intensity dependence (Ahrens et al. 2008; Nagel and Doupe 2008), or sensitivity to multiple spectrotemporal features (Atencio et al. 2008) could build more complex receptive fields and contribute to further increases in selectivity in the songbird auditory system. Feedback mechanisms may also be involved, as CMM makes extensive projections back to CLM and NCM (Vates et al. 1996). FRFs could provide a valuable tool in exploring the emergence of selectivity in these areas.

Emergence of selectivity for complex objects in CMM

Although the FRF accounted for a much larger proportion of the variance in CMM responses than any previous model, substantial nonlinearities remained. These were observed as facilitative and suppressive interactions between notes. More than 30% of the note responses were significantly suppressed in the context of the motif, and 19% of the note responses were facilitated. Although facilitation could be caused by inappropriate segmentation of the motifs, for instance if an excitatory spectrotemporal pattern is split across features, suppression suggests long-range interactions between notes that would otherwise be excitatory. The higher prevalence of suppression thus provides further support for the idea that notes encapsulate the excitatory features for CMM neurons.

Suppressive interactions may be involved in the selectivity of CMM neurons. They are a potential mechanism for temporal combination sensitivity, which has been commonly observed in the song system (Margoliash 1983; Margoliash and Fortune 1992), and here would probably involve lateral inhibition between CMM neurons. In support of this model, selective cells appeared to be drawn primarily from a single class of neurons (Fig. 4); these may be principal neurons. Other neurons were significantly less selective and exhibited much narrower spike shapes and higher spontaneous rates, properties often associated with interneurons (McCormick et al. 1985; Rauske et al. 2003). Lateral inhibition can sharpen tuning and increase the precision of responses to preferred stimuli (Tan et al. 2004; Wehr and Zador 2003), which could explain the correlation between selectivity and spike timing precision observed here (Fig. 5C).

Effects of behavioral state

Neurons in this study were recorded under restrained, unanesthetized conditions and also under urethane anesthesia. In contrast to a previous study (Capsius and Leppelsack 1999), we saw no significant differences in spontaneous firing rate or the proportion of auditory units. However, our data are from well-isolated single units that were only presented with conspecific stimuli, whereas the earlier study measured multiunit responses to a wide variety of natural and synthetic stimuli. If there are enough neurons in CMM that are not strongly auditory but show state-dependent changes in their activity, this could account for the differences between our results. Spike timing precision and intertrial correlation were significantly higher in the unanesthetized state, which is consistent with the effects of anesthesia on the balance of excitation and inhibition, and the effects of this balance on spike reliability and timing (McCormick et al. 1985; Tan et al. 2004; Wehr and Zador 2003). In addition, facilitative interactions between notes were more likely in unanesthetized birds, which suggests increased selectivity for complex spectrotemporal patterns. We observed no significant difference in motif selectivity between states, but given the high variance in SI across neurons, the sample size may not be large enough to draw any firm conclusions.

Our results suggest a hierarchical model of auditory processing in which neurons in CMM represent the identity of complex auditory objects through a linear combination of inputs tuned to specific auditory features. These auditory features are present in notes, suggesting that auditory processing in starlings may be specialized for elements that have salience for both the production and interpretation of vocal communication. Our data show parallels to the emergence of selectivity for complex features in primary and secondary areas of other modalities (Jortner et al. 2007; Kobatake and Tanaka 1994; Riesenhuber and Poggio 2002), suggesting common mechanisms for the detection of patterned information in sensory data.

Acknowledgments

We thank S. M. Sherman and J. N. MacLean for comments on a previous version of the manuscript.

Grants

This work was supported by National Institute of Deafness and Other Communication Disorders Grants DC-007206 and F32 DC-008752 and National Science Foundation Grant DMS-07-06048.

References


