Quantifying unit isolation.

In order to make our criterion for classifying recordings explicit, we used a simple method for quantifying the isolation of action potential waveforms. This technique is based on principal components analysis (PCA), a mathematical technique commonly used for waveform classification. As shown in supplemental Figure 1a, a threshold (red line) was used to select both spike and noise waveforms from singing-related activity. These waveforms (supplemental Figure 1b) were then analyzed using PCA, which identifies the dimensions of variability (or principal components) that account for the greatest amount of total waveform variability. By projecting each neural waveform onto each of the two components that account for the most variability (PC$_1$ and PC$_2$), we obtained a 2-dimensional representation of the distribution of waveforms (supplemental Figure 1c). Across the entire dataset, the first two components captured 64.0 +/- 7.2% (mean +/- SD) of the total variance of waveforms in each recording (70.6 +/- 8.9% for single units, 62.7 +/- 6.0% for multiunit sites, see below). An automated nearest-neighbor clustering algorithm (kmeans.m in MATLAB, The MathWorks, Natick, MA) was then applied to the 2-dimensional data to assign each waveform to a cluster. The number of clusters used by the algorithm was set by hand. In the majority of cases, two clusters (a "spike" cluster and a "noise" cluster) were selected. In a small number of cases, three clusters were selected to capture waveforms belonging to two distinct spikes in addition to the noise cluster. In these cases, only the waveforms belonging to the larger spike were used. This cluster classification is represented by the colors of the
datapoints in supplemental Figure 1d. Each cluster was then fit with a 2-D gaussian (ellipses in supplemental Figure 1d) in order to estimate its mean and variance.

The goal of this analysis was to establish a scalar measurement of unit isolation. We found that the extent of overlap between the gaussian cluster fits served as a reliable indicator of unit isolation. Overlap was quantified by computing the probability with which a point from one cluster would be miscategorized by the nearest-neighbor algorithm as belonging to the other cluster. We generated 10,000 points from each cluster (according to that cluster’s gaussian fit), reran the nearest-neighbor algorithm, and measured the frequency with which synthetic points were miscategorized. In the example shown in supplemental Figure 1 a-e, the gaussian fits had an overlap (or “isolation error”) of 0.0013. Examination of our data suggested that an isolation error of 0.01 was a reasonable threshold for classifying a recording as single-unit. Accordingly, clusters with isolation errors of less than 0.01 were classified as single-unit, and clusters with errors of greater than 0.01 were classified as multiunit. Two additional examples of isolation measurement are shown in supplemental Figure 1 f and g.
Supplemental Figure 1. Isolation quantification technique. (a) One second of neural data from Bird 1, Unit 8. A threshold (red line) was used to collect voltage waveforms. (b) Collected waveforms. Times are relative to the peak of the negative excursion that crossed the voltage threshold. (c) Principal components representation of waveforms using the first two components. (d) A nearest-neighbor algorithm classified waveforms into two clusters, corresponding to “spike” (green) and “noise” waveforms (red). These clusters were fit with 2-D gaussians (ellipses show 2 S.D.) and cluster overlap (isolation error) was estimated. Since the isolation error is less than the threshold value of 0.01, this recording is classified as single-unit. (e) Waveforms colored by cluster assignment. Here the waveforms plotted in (b) are color-coded according to their cluster assignment. Note that while some “noise” traces (red) do contain spikes, these spikes were not responsible for the threshold crossing, and are separately represented as green traces with spikes centered at time zero. For clarity only 300 waveforms from each cluster are plotted in (b) and (e). Neural data and isolation errors for an additional single unit and one multiunit recording are shown in (f) and (g), respectively.
Quantitative analysis of spectral variability

In our analysis of RA's contribution to trial-by-trial variations in the acoustic structure of song syllables, it is important that we choose acoustic measures that capture a significant portion of behavioral variability. Failing to do so could cause us to underestimate RA's influence on behavior, since a correlation test requires the presence of both neural and behavioral variability in order to be meaningful. The spectral structure of song is extremely complex, offering many potential measures of vocal output (Tchernichovski et al., 2001). Although behavioral studies, stimulation experiments, and *in vitro* measurements of the syrinx have suggested that pitch, amplitude, and spectral entropy are tightly controlled song parameters (and therefore likely to be under RA's influence), we wanted to investigate explicitly the nature of spectral variation within each syllable.

With these issues in mind, we performed principal components analysis (PCA) on the power spectrum of each syllable in our dataset in order to characterize the important dimensions of behavioral variability. PCA identifies the axes of variation in a data set that account for the greatest amount variability. By performing PCA on the power spectra recorded from all renditions of a single syllable, we identify the axes of variation (that is, the deviations from the syllable's mean spectrum) that make the greatest contributions to the total spectral variation.

Supplemental Figure 2 (a) and (b) illustrate PCA as performed on two example syllables. For each syllable in our dataset, we measured the power spectrum at the same measurement time (dashed black lines at left) used to compute pitch, amplitude,
and entropy in the original analysis (see Methods). Power spectra were sampled at 71 evenly-spaced frequencies between 1-10 kHz. Mean spectra for the two example syllables are shown in the middle panels of (a) and (b). For each syllable, this procedure resulted in an $n \times 71$ matrix, where $n$ is the number of instances of the syllable in question. We then performed PCA on each matrix to derive a set of 71 orthogonal basis vectors (the principal components) in which PC1 is the direction along which spectral variance is maximized, PC2 is the second-most variable direction, and so on. The fraction of the total variance explained by each of the 71 principal components for the example syllables is shown at right in (a) and (b).

**Identification of dominant axes of spectral variation:**

These two examples are typical of our dataset in that the first 1-3 principal components accounted for much more of the total variance than did any of the remaining components (group data are shown in supplemental Figure 2 c and d). These data indicate that while spectral variations were not easily reduced to a compact representation (as would have been the case if, for example, 95% or more of the cumulative variance had been accounted for by a small number of components), in each syllable a few dimensions of variability accounted for a relatively large amount of the total variability.

**Quantifying similarity between PCs and changes in pitch, amplitude, and entropy:**

We often found that the dominant principal components (those describing the greatest fraction of behavioral variation) resembled changes in pitch, amplitude, or
entropy. Below, we show components for which this was the case and illustrate how we quantified the similarity between principal components and these three acoustic features.

Supplemental Figure 3a shows an example in which the dominant principal component resembled a change in pitch. This can be seen by noting that positive weights along this component correspond to a downward shift in the spectral peaks of the syllable (red trace in the bottom panel, representing the mean spectrum + 2*PC1). Conversely, negative weights along this component reflect an upward shift in pitch (blue line in bottom panel).

To quantify the degree to which this principal component described a pitch shift, we measured its similarity to a “synthetic component” generated by explicitly varying the pitch of the mean power spectrum. Synthetic pitch components were generated by first linearly stretching (or contracting) the mean power spectrum across the frequency axis to simulate an increase (or decrease) in pitch. The mean spectrum was then subtracted from the pitch-shifted spectrum to generate a synthetic component representing the change in power at each frequency resulting from the pitch shift. The best-fit synthetic component (red dashed line in middle panel of supplemental Figure 3b) was found by varying the stretching factor (pitch shift) until the closest possible match to the real component was achieved. The similarity between real and synthetic components was then computed using the cosine-similarity measure, which measures the cosine of the angle between the two vectors in 71-dimensional space (and ranges from –1.0 to 1.0). In the case illustrated in supplemental Figure 3b, PC1 and the synthetic amplitude change have a cosine-similarity score of 0.99.
Supplemental Figure 3b illustrates the power spectrum (top) and first principal component (middle) for a case in which PC1 described a change in amplitude. Positive weights along this component reflect a subtraction of power at each frequency (red line in bottom panel). Conversely, negative weights along this component reflect an increase in amplitude (blue line in bottom panel). The blue dashed line in the middle panel of supplemental Figure 3a shows the synthetic component explicitly simulating a change in amplitude that best fits PC1. Synthetic amplitude components were generated by simulating a constant amplitude offset at each frequency. The best-fit synthetic component was found by varying the offset until the closest possible match to the real component was achieved. In the case illustrated in supplemental Figure 3a, PC1 and the synthetic amplitude change have a cosine-similarity score of 0.96.

Supplemental Figure 3c illustrates the power spectrum (top) and second principal component (middle) for a case in which PC2 described a change in spectral entropy. Positive weights along this component reduce the power at the spectral peaks (dotted vertical lines) and increase the power in the troughs between the peaks, as shown in the bottom panel. Although changes in spectral entropy (defined as  \( \Sigma p \log(p) \), see Methods) could in principle be achieved by many different changes to spectral structure, we consistently found principle components of the form shown in supplemental Figure 3c. Synthetic entropy components were therefore generated by summing a scalar offset with gaussian distributions centered on each harmonic peak present in the spectrum:
Where $PC_{\text{synthetic entropy}}(f)$ is the value of the synthetic component at a given frequency and $h_x$ are the $x$ harmonics of the syllable. The parameters of this equation are constrained such that (1) $|\alpha| < 0.66*|\beta|$ and (2) $\alpha$ and $\beta$ are of opposite sign. These constraints are necessary to ensure that the gaussian components contribute to the overall shape of the synthetic component (i.e. they ensure that synthetic entropy components are not allowed to approximate synthetic amplitude components, which have the form $PC_{\text{synthetic amplitude}} = \alpha$). In the case illustrated in supplemental Figure 3c, the real and synthetic components had a cosine similarity score of 0.93.

**Classification of the dominant components:**

We then asked how frequently the principal components explaining the greatest amount of behavioral variation were well-described by a synthetic pitch, amplitude, or entropy component. To do this, we restricted our analysis to PCs that explained at least 10% of the total variance of the syllable ($PC_{>10\%}$). For each $PC_{>10\%}$, we found the best-fit synthetic components describing changes in pitch, amplitude, and entropy and computed the cosine-similarity for each. If cosine-similarity exceeded a threshold of 0.8, the $PC_{>10\%}$ was classified as “congruent” to the relevant synthetic component.

To verify that our categorization of principal components as “congruent” with pitch, amplitude, or entropy is meaningful, we compared the amount of behavioral variation along the direction of each $PC_{>10\%}$ with the amount of behavioral variation...
along the direction of the best-fit congruent synthetic component. We found that the fraction of the total variability explained by the congruent synthetic components was not significantly different from the fraction explained by the corresponding PC>10% (p=0.35, Wilcoxon signed-rank test).

If a PC>10% was congruent with more than one best-fit synthetic component, then the component with the greater cosine-similarity score was selected. This occurred in 22% of cases and almost always resulted from a PC>10% being found congruent with both amplitude and entropy. Using this criterion, we found that the majority (70.2%) of PC>10% were congruent with either pitch, amplitude, or entropy (supplemental Figure 4a), showing that the most dominant dimensions of behavioral variability usually reflect variations in one of these parameters. Moreover, in cases where PC>10% were not congruent with pitch, amplitude, or entropy, the PC>10% often resembled a linear combination of two of these features (not shown).

Conversely, supplemental Figure 4b shows that when all PCs (not just those explaining more than 10% of the total variance) are classified by cosine-similarity, components congruent with pitch, amplitude, and entropy explain a much larger proportion of behavioral variation than do components that are not congruent with these parameters. This result demonstrates that among all dimensions of spectral variation yielded by PCA, the components describing changes in these three parameters had much more explanatory power than other components. Together, these results suggest that pitch, amplitude, and entropy are indeed important components of behavioral variability, and therefore are reasonable features to use in characterizing the contributions of RA activity to behavioral variation.
Primary and PCA-based behavioral analysis yield similar results:

Finally, we wanted to ensure that our decision to quantify correlations between neural activity and the measured values of pitch, amplitude, and entropy (rather than correlations between neural activity and loadings along the PC>10%) did not result in our underestimating the strength of neural-behavioral correlations. Supplemental Figure 4c shows that the $r^2$ values of correlations between neural activity and the measured behavioral parameters are not significantly different from the strength of correlations between neural activity and loadings along the PC>10%.
Supplemental Figure 2: PCA-based analysis. (a) and (b) show the spectra of two example syllables. Note that the syllable shown in (a) is syllable "E" from Bird 1, also illustrated in Figure 5 in the main text. For each syllable, we computed the power spectrum at the same measurement time used for the primary analysis (of pitch, amplitude, and entropy) in the main text (dashed lines in spectrograms at left). Each power spectrum consisted of measurements of the log spectral power at each of 71 equally-spaced frequencies between 1-10 kHz. The mean power spectra for the two example syllables are shown in the middle panels of (a) and (b). PCA was then used to find the dimensions of variation (or "components") that captured the greatest, second greatest, etc. fraction of the total variation of each syllable. The fractions of the total variance accounted for by each component of the two example syllables is shown in the right panels of (a) and (b). (c) shows the mean +/- SD of this measure for all 83 syllables in our dataset. Inset shows the first 5 components in greater detail. (d) shows the same data as (c) expressed as a cumulative fraction.
Supplemental Figure 3: PCs approximating changes in pitch, amplitude, and entropy. (a) shows the mean power spectrum (top), first principal component (middle, black trace), and deviations from the mean spectrum along the first principal component (bottom) for the syllable shown in supplemental Figure 2a. The dashed red line in the center panel shows the synthetic pitch component that best fits the principal component (see text). Vertical dashed lines indicate spectral peaks at the first and second harmonics. (b) shows the same measures for the first principal component of the syllable shown in supplemental Figure 2b. The dashed blue line shows the synthetic amplitude component that best fits the principal component. Other conventions as in (a). (c) shows the same measures from the second principal component of the syllable shown in supplemental Figure 2a. The dashed green line shows the synthetic entropy component that best fits the principal component.
**Supplemental Figure 4** (a) Distribution and classification of principal components explaining at least 10% of spectral variation (PC>10%). Blue shading indicates a PC>10% classified as congruent with pitch, red indicates a PC>10% congruent with amplitude, green indicates a PC>10% congruent with entropy, and black indicates PC>10% that were not congruent with any of the three parameters. (b) Cumulative distributions of the fraction of the total variance explained by all principal components. Data shown are for all principal components, regardless of how much spectral variance they explain. Components are divided into those congruent with pitch (blue), amplitude (red), entropy (green), and those not congruent with any of these three parameters (black). (c) Cumulative distributions of $r^2$ values for significant correlations between neural activity and acoustic output for each acoustic feature. In each plot, $r^2$ values are shown for two different methods of quantifying acoustic output on each trial. Solid lines show the correlation strengths resulting from measuring pitch, amplitude, or entropy as described in the main text. Dotted lines show the $r^2$ values resulting from using the loadings along the PC>10% congruent with the relevant acoustic measure.
Testing the timescale of premotor encoding

Work in diverse sensory and motor systems has shown that spike trains can carry information on multiple time scales (Theunissen and Miller, 1995; Tiesinga et al., 2008). At one extreme, sensory input or motor output might be encoded by the total number of spikes produced on a given trial (rate coding). At the other extreme, millisecond-scale differences in spike timing might encode important task parameters.

Here, we compare the predictive power of premotor encoding at a range of timescales. For each instance where a single unit was active prior to a syllable (one "case," see Methods), we fit the data with five different linear models that range from a simple rate coding scheme (counting the total number of spikes in the 40 msec premotor window) to a model that divides the premotor window into 1 msec-wide bins in order to ask whether fine temporal structure encodes motor output. The models take the form

\[ b = \sum_{i=1}^{n} a_i s_i + c \]

where \( b \) is the behavioral parameter to be fit (pitch, amplitude, or entropy), \( s \) is the number of spikes falling in each of \( n \) consecutive bins spanning the 40 msec premotor window, and \( [a_1, a_2, a_3, a_4, a_5, a_6, c] \) are the fit parameters of the model. The five models tested have \( n = [1, 2, 4, 8, 40] \), corresponding to bins with widths of 40, 20, 10, 5, and 1 msec respectively, as illustrated in Supplemental Figure 5a.

For each case and behavioral parameter, we used cross-validation to assess which model has the most predictive power while protecting against overfitting. We randomly
split the data into a training set consisting of 80% of the data used to fit the model parameters, which were then used to predict the data in a testing set consisting of the remaining 20% of the data. Repeating the procedure 1000 times for each model (splitting the data randomly on each iteration) yielded a distribution of error (mean squared error). The "best" model was defined as the model with the lowest mean error.

Supplemental Figure 5 shows that for all three behavioral parameters, counting the number of spikes in a single 40 msec-wide bin \( (n = 1) \) in the above equation is the best predictor of the neural-behavioral relationship in a majority of cases, and represents the best model more than twice as frequently as any of the other tested encoding schemes. These results suggest that in most cases the fine temporal structure of spikes does not encode the acoustic structure of song (at least not in the linear fashion described by the above equation). Consequently, models that use fine temporal structure to predict behavior presumably produce higher cross-validation error because free parameters are fit to non-informative features of spike trains. Based on these results, we used spike count in a single 40 msec window – the tested model with the most predictive power in the majority of cases– to examine correlations between neural activity and song.
Supplemental Figure 5. Cross-validation analysis. (a) Five models with different binsizes. Each model divides the 40 msec-long premotor window (horizontal red line) into a different number of bins. Song output is modeled as a linear function of the number of spikes falling within each bin (see text). For the example spike train (vertical red lines), the model with a single bin would count four spikes. For the model with two 20 msec-long bins, three spikes fall in the first bin, and one spike falls in the second bin, and so on. (b) Performance of the five models. As described in the text, we used a cross-validation approach to determine which model is best able to predict the neural-behavioral relationship in each case where a neuron is active prior to a syllable. Bar plots show the fraction of cases in which each model is the best predictor of pitch, amplitude, and entropy.
**Multiple comparisons:**

Our goal is to determine whether trial-by-trial variations in premotor neural activity are correlated with behavioral variations in a significant number of cases. This presents a multiple comparison problem: as the number of tests increases, so will the number of false positives (cases in which a significant correlation is found but none exists). For example, if we are performing \( k \) independent correlation tests (at \( p<0.05 \)) on data in which no correlations exist, the odds of finding at least one spuriously significant correlation are \([1-0.95^k]\). Put another way (in terms of the expectation value for the number of false positives), for \( k=100 \) tests, we would expect to find 5 false positives.

There are several approaches to dealing with this problem. One approach (employed by the Bonferroni correction and related techniques) is to control the probability that any test is found to be significant. More formally, for \( k \) tests, these techniques control the probability of rejecting at least one of \( k \) null hypotheses when all \( k \) null hypotheses are true (Westfall and Young, 1993). With this type of correction, finding any test significant provides evidence (at a confidence level determined by the controlled false-positive rate) of a relationship between the variables in question.

An alternate technique is to test whether the proportion of significant correlations is greater than that expected by chance. As mentioned above, 100 independent tests of uncorrelated variables will on average yield 5 false positives. We could then use a binomial test to determine whether the proportion of significant correlations in the empirical data is significantly greater than the proportion expected by chance.

While both techniques are in principle correct (provided that all tests are independent, a topic discussed at length below), they answer slightly different questions
about our dataset. Applying the Bonferroni correction by reducing each p-value threshold within each neuron (e.g. by a factor of 18 if a neuron is subjected to 18 tests) still results in many significant correlations with pitch, amplitude, and entropy (76, 76, and 47, respectively). Furthermore, in the Bonferroni-corrected dataset significant correlations with pitch have a positive sign in 68% of cases (the asymmetry is significant at p=0.001), whereas no significant asymmetry is found in correlations with entropy and amplitude. Similar results are obtained from an even more conservative version of the Bonferroni correction, in which the threshold p-value for every test is reduced by a factor of 2115 (the total number of tests performed on all 145 neurons). These results indicate that at least some significant correlations are present in our data.

Because one goal of our paper is to estimate the proportion of RA units encoding the acoustic output at any given time, we also employed a proportion-based technique, since it tells us about the fraction of the neural population encoding behavioral variation. This approach is detailed below.

**Non-independent tests:**

While the proportion-based (binomial) test describe above is applicable when all tests being performed are independent, our analysis is potentially complicated by correlations among the three acoustic parameters measured from each syllable (behavioral-behavioral correlations) and by correlations between neural activity during consecutive syllables (neural-neural correlations) (see Results). Correlations of this sort raise the possibility that our tests for correlations between premotor activity and acoustic output (neural-behavioral correlations) might not be independent. In the extreme case, if
the pitch, amplitude, and entropy of each syllable (for example) were perfectly correlated, then the three correlations between neural activity and each parameter only truly constitute a single independent test. The apparent statistical power of a binomial test (to determine whether the proportion of correlations is significantly greater than 5%) might therefore be artificially inflated.

To illustrate this, consider a hypothetical case in which 20 correlation tests are performed, where all 20 tests are independent, and of which 2 are found to be significant. Although 2/20 > 5%, this proportion fails the binomial test (p=0.08). Now, imagine that for each of the 20 original tests, 9 other tests are performed with outcomes that are perfectly correlated with the corresponding test from the original 20. In this case, although there are only 20 truly independent cases, we would find that 20/200 tests were significant, a proportion that easily passes the binomial test (p=0.001).

Estimating the null distribution:

It is therefore necessary to correct for the consequences of neural-neural and behavioral-behavioral correlations. Our approach is to use a resampling technique to create an artificial dataset in which all neural-behavioral (NB) correlations are broken, but all neural-neural (NN) and behavioral-behavioral (BB) correlations are preserved. We then perform correlation tests on these resampled datasets and note the proportion of cases with significant correlations. By performing this procedure many times, we can estimate the distribution of proportions of significant correlations under the null hypothesis, and then ask whether the proportion of significant correlations in the real dataset is beyond the 95th percentile of the null distribution.
Details of the resampling-based approach:

First, consider a bird that sings only one syllable, from which we have recorded 10 neurons. For one of these neurons, let N and B be matrices of neural and behavioral data, respectively. The rows of N and B correspond to trials, and the columns correspond to the tests being performed. N<sub>k,i</sub> is therefore the premotor neural activity (# spikes) during trial k of test i, and B<sub>k,i</sub> is the value of the appropriate acoustic parameter during trial k. If there are n examples of the syllable, N and B will each have n rows and 3 columns, representing the three acoustic measurements (pitch, amplitude, and entropy). (Note that if the bird sings only one syllable, the columns of N are identical: N<sub>k,i</sub> = N<sub>k,j</sub> for all k, since the three different behavioral measures are being compared to the same premotor neural activity.)

Our goal is to create a dataset in which any neural-behavioral correlation between columns N<sub>i</sub> and B<sub>i</sub> are broken, but correlations between columns N<sub>i</sub>, N<sub>j</sub>, i ≠ j (NN correlations) and B<sub>i</sub>, B<sub>j</sub>, i ≠ j (BB correlations) are preserved. To accomplish this, we permute (shuffle) the rows of N, resulting in N<sup>p</sup>. (The same results could be achieved by permuting the rows of B.) We then compute correlations between the three paired columns N<sup>p</sup><sub>i</sub> and B<sub>i</sub> and count the number of tests that achieve significance (at p<0.05).

Since our hypothetical dataset consists of ten such neurons, running the above-described procedure on each neuron would yield 30 tests, of which some subset will achieve significance. We then record this proportion and repeat the procedure many times to generate a distribution of proportions of significant correlations under the null
hypothesis. We can then reject the null hypothesis if the proportion of significant cases in the original dataset lies beyond the 95\textsuperscript{th} percentile of the null distribution.

**Multiple syllables:**

Now we consider how to perform this procedure when a bird's repertoire consists of more than one syllable, as is the case in our data (Bengalese finch song typically contains 5-10 distinct syllables). Consider a case in which a bird's song contains 5 syllables, labeled "ABCDE". If the bird always sings these syllables in this order, an example song might be "ABCDE-ABCDE-ABCDE" (hyphens are inserted for visual clarity and do not represent syllables) and adapting the above procedure is straightforward. Both the N and B matrices, rather than having 3 columns (3 acoustic measurements of 1 syllable), will have 15 columns (3 acoustic measurements x 5 syllables). Each row of N and B will correspond to one rendition of the motif "ABCDE."

Permuting the rows of N will remove NB correlations, while preserving BB and NN correlations. Note that when multiple syllables are analyzed, NN and BB correlations describe neural and acoustic correlations across, as well as within, syllables.

Our analysis is complicated, however, by the fact that syllable order in Bengalese finch song, while highly patterned, is seldom as stereotyped as in the above example. A more typical syllable order for a Bengalese finch song bout is "AB-AB-ABC-ABCDE-ABCDE-AB" (again, hyphens inserted for visual clarity only). Because of inconsistencies in the order and prevalence of syllables, it is not possible to gather all neural data into one neural matrix and one behavioral matrix. Instead, data from each syllable must be collected into a separate pair of matrices. Then, in each run of the
resampling algorithm, we permute the rows of all neural matrices. However, it would be incorrect to permute the neural matrices independently: to do so would destroy any NN and BB correlations between, for example, syllable "A" and syllable "B". Rather, we permute the neural matrices in a manner that, as closely as possible, preserves relationships between consecutive syllables in the original data.

Row permutation for complex syllable patterns:

For each recorded neuron, we divide the recorded syllables into "segments," where each segment is the longest possible sequence of syllables in which no syllable is repeated. Here, segment # is indicated with subscript for an example sequence:

A₁B₁ - A₂B₂ - A₃B₃C₃ - A₄B₄C₄D₄E₄ - A₅B₅C₅D₅E₅ - A₆B₆ - A₇B₇C₇D₇E₇

Then, on each resampling trial, we randomly permute the order of segments. For example, the segment reordering on one resampling trial might be: [6 2 5 7 4 3 1]. The reordered (permuted) song would look like this:

A₆B₆ - A₂B₂ - A₅B₅C₅D₅E₅ - A₇B₇C₇D₇E₇ - A₄B₄C₄D₄E₄ - A₃B₃C₃ - A₁B₁

Note that while the order of the segments has been permuted, local adjacencies (that is, the order of consecutive syllables within segments) are preserved.
We then use this reordering of segment sequence to shuffle the rows of the neural matrices:

Permutated order for each syllable:

- Reorder the 7 rows of $N_A$ by $[6257431]$
- Reorder the 7 rows of $N_B$ by $[6257431]$
- Reorder the 4 rows of $N_C$ by $[3421]$
- Reorder the 3 rows of $N_D$ by $[231]$
- Reorder the 3 rows of $N_E$ by $[231]$

The procedure allows us to permute the matrices of neural data in as consistent a fashion as possible given that not every syllable is included in every song segment. Note that in cases where all segments are identical, as in

$$A_1B_1C_1D_1E_1- A_2B_2C_2D_2E_2- A_3B_3C_3D_3E_3- \ldots$$

this method is equivalent to collecting all data into one neural matrix and one behavioral matrix, and permuting the rows of the neural matrix.

**Results of resampling analysis:**

Using the above-described techniques, we permuted our entire dataset 1000 times, creating a distribution of the number of significant correlations (out of the total of 2115 cases) expected under the null hypothesis. When tests for pitch, amplitude, and entropy were combined, the 95th percentile of this distribution fell at 122 cases, (or 122/2115 = 5.8% of the total number of tests), as shown by the dashed black line in supplemental Figure 6a. When the combined resampled distribution was separated into tests for pitch,
amplitude, and entropy, the critical values were at 46, 44, and 45 significant tests out of 705 (6.5%, 6.2%, and 6.4%, respectively; dashed black lines in supplemental Figure 6b,c, and d). Since these values are quite similar, we used the most conservative of these figures (6.5%) as the significance threshold in Figure 7a in the main text.

The number of significant correlations in our dataset exceeded the relevant significance threshold in all cases (169, 175, and 137 for pitch, amplitude, and entropy respectively; dashed red lines in supplemental Figure 6b,c, and d). Furthermore, the number of significant correlations not only exceeded significance thresholds but fell beyond the range of all 1000 resampled datasets, whether tests for pitch, amplitude, and entropy were considered either together or separately. The observed proportions were thus significantly (p<0.001) greater than those expected by chance.

Supplemental Figure 6: Results of resampling analysis. Each plot shows the distribution of the number of significant correlations in each of the 1000 permutations of the empirical data. The 95th percentile of this distribution (dashed black lines) is the threshold for significance. In all cases, the number of significant correlations in the empirical data (dashed red lines) was beyond this threshold. Permutation tests were conducted using the full dataset, combining tests for pitch, amplitude, and entropy, as described in the text and shown in (a). (b), (c), and (d) show the results of separating the empirical and resampled data into separate distributions for each acoustic parameter.
Supplemental Figure 7. Post-song inhibition. (a) Rasters show the activity of a single unit from Bird 4 aligned to the onset (left) and offset (right) of song. As was typical in our data, this unit switched from regular tonic activity (far left) to bursty firing several seconds prior to song onset. Following song offset, spiking was inhibited for approximately 500 msec, after which regular tonic firing resumed. (b) Mean firing rate for the unit shown in (a). Before song-related bursting begins, this unit had a baseline rate of 34 Hz (dashed red line). We quantified post-song inhibition by comparing the baseline rate to the mean firing rate in a window 100-400 msec after song offset (blue box). (c) Group data. We included only those units for which we collected at least 10 song offsets after which song did not resume for at least 3 seconds (n=43). Of these, 36 units (red dots) had significantly lower mean firing rates in the post-offset window, 1 unit (blue dot) had greater activity, and 6 units (white dots) were not significantly different (2-tailed t-test, p<0.05). An arrowhead indicates the unit shown in (a) and (b).
**Putative interneurons vs. putative projection neurons**

In contrast to the tonic activity characteristic of the majority of units (Figure 3a in the main text), a small subset of recordings (1 single-unit, 3 multiunit) had very low or no spiking activity when the bird was at rest and displayed bursty spiking activity during song. The spike width-at-half-height of the sole single unit of this type (103 µsec) was narrower than spike widths of all other single units (164 +/- 31 µsec, mean +/- S.D.). The similarity of the spike widths and activity patterns in this subset of units to a similar class in the zebra finch (Spiro et al., 1999; Leonardo and Fee, 2005) suggests that these four units are interneurons, and that the main body of our recordings are from projection neurons. The four putative interneurons were not included in further analysis. See supplemental Figure 8b for an example of a recording from a putative interneuron alongside a recording from a putative projection neuron (supplemental Figure 8a).
Supplemental Figure 8. Firing patterns of a putative projection neuron and a putative interneuron. (a) Raw sound amplitude trace (top) and a recording of a putative projection neuron (bottom) from Bird 2. Putative projection neurons are spontaneously active when the bird is not singing, displaying characteristic evenly spaced spikes, such as those seen in the last second of the neural trace. The bursty activity preceding the song is likely related to the production of three introductory notes, marked with green asterisks. (b) A putative interneuron from Bird 3. Neurons of this type are not active when the bird is at rest, but fire bursts of activity during song. Putative projection neurons and putative interneurons made up 97% and 3% of units sampled, respectively.
**Sparse distribution of significant correlations**

While most units are correlated with at least one acoustic parameter at some point during song, significant correlations are sparsely distributed (see Figure 6 in the main text and supplemental Figure 9). That is, units are typically active before multiple syllables, but are significantly correlated with acoustic output in only a fraction of these cases. This sparse distribution might reflect dynamic changes in the strength of covariation between recorded neurons and the ensemble of RA neurons controlling each acoustic feature. Such changes would presumably lead to across-syllable differences in the correlation between that cell’s activity and motor output.

The sparse distribution of significant correlations might also result from nonlinearities either in the brainstem targets of RA or in the syrinx itself. The activity of a group of brainstem motor neurons or a syringeal muscle might influence pitch (for example) during some syllables but not others. Variations in the activity of RA neurons driving these motor structures would therefore only produce pitch variations during a subset of syllables.

In addition to being sparsely distributed across syllables, significant correlations are also distributed across acoustic properties such that one unit can be correlated with more than one acoustic feature. This might reflect either the multiple actions of individual syringeal muscles or the connectivity of RA neurons to the motor neuron pool. EMG studies have shown that activation levels of the syringeal muscles controlling amplitude are also correlated with pitch (Goller and Suthers, 1996), suggesting that individual muscles can contribute to the control of multiple acoustic features. Furthermore, although RA has a roughly myotopic organization (Vicario, 1991), single
RA neurons might activate motor neurons controlling multiple muscles, thereby affecting multiple aspects of song (Wild, 1993).

**Supplemental Figure 9 (on following 2 pages)** Significant correlations of premotor activity with acoustic structure across all birds. Each box shows data from one bird in our dataset, following the same conventions as Figure 6 in the main text (which showed the data from Bird 1). Note that different birds can have different numbers of rows (reflecting differences in the number of units recorded) and different numbers of columns (reflecting differences in the number of syllables sung by each bird).
**RA activity in directed vs. undirected song**

Song is produced both in social isolation ("undirected song") and during courtship interactions ("directed song"). Previous studies have established that directed song is less variable than undirected song (Sossinka and Bohner, 1980; Kao and Brainard, 2006; Sakata et al., 2008). Furthermore, lesion, inactivation, and stimulation studies suggest that some of the increased behavioral variation observed during undirected song results from lMAN injecting neural variation into RA (Kao et al., 2005; Olveczky et al., 2005; Kao and Brainard, 2006). Together with the results of the current study, these prior findings suggest that the increased behavioral variability during undirected song might be driven by changes in the overall level of variability in RA activity across social contexts.

Although our dataset consists almost entirely of recordings during undirected song, in a small number of cases in one bird we were able to obtain both directed and undirected song while recording in RA. Two such recording sites yielded sufficient data in both conditions to allow for comparison across social contexts. As shown in supplemental Figure 10, neural variability (CV) was indeed significantly lower in the directed condition. While preliminary, these results suggest that not only do trial-by-trial variations in RA activity drive variations in song (the main finding in our study), but also that modulations in the overall level of RA variability are responsible for social context-dependent changes in song.
Supplemental Figure 10. Neural variability across social context. (a) Spiking activity during undirected (middle) and directed (bottom) song for Bird 3, unit 8. Bouts of directed and undirected song were interleaved during data collection and are plotted separately for visual clarity only. Other plotting conventions as in Figure 4 in the main text. (b) Variability of spiking activity in directed vs. undirected song. Each point plots the CV (SD/mean) of the number of spikes in the premotor window before one syllable. Only syllables with mean activity >25 Hz are analyzed (see Methods). Squares plot data from the unit shown in (a), circles plot data from Bird 3, unit 3. The p-value is from a Wilcoxon signed-rank test.
Correlations extend across time

When quantifying the relationship between neural activity and song output, we restricted our analysis of neural activity to a 40 msec premotor window preceding each syllable. This window likely encompasses the latencies with which spikes in RA directly (via synapses with motor neurons in the brainstem) influence behavior. Using this restricted time window allows us to examine the relationship between neural activity and acoustic output in the context of a single syllable.

However, variations in both premotor activity and acoustic output might have a timecourse longer than a single syllable (Glaze and Troyer, 2006), and an increase in firing rate or pitch (for example) in one syllable might correlate with a similar increase in the next syllable. In the extreme case, consecutive syllables could be perfectly correlated, both in terms of premotor activity and acoustic output. In this situation a 40 msec premotor window, although it includes the "true" causal latency of RA neurons, would have no more or less predictive power than any other window, and acoustic output would be predicted equally well by neural activity following, rather than preceding, the syllable. At the other extreme, variation in one syllable could be completely independent of variation in neighboring syllables. In this case, neural activity outside of the true premotor window would be a poor predictor of behavioral output.

To examine whether behavioral correlations extend across time, we asked whether the acoustic properties of a given syllable are correlated with the acoustic properties of the next syllable. We found this type of correlation to be widespread. The pitch of a given syllable was significantly (p<0.05) correlated with the pitch of the next syllable in 23.8% of cases. Similar measurements of amplitude and entropy yielded significant correlations in 48.5% and 21.4% of cases, respectively.
We then performed a similar analysis of premotor activity to ask whether neural activity is also correlated across time. As with acoustic output, we found a substantial number of correlations. The number of premotor spikes before a given syllable and the number of spikes before the next syllable were significantly correlated in 32.4% of cases.

The prevalence and strength of these correlations show that motor activity during consecutive syllables is neither perfectly correlated nor completely independent. Since consecutive syllables are not perfectly correlated, neural activity in the 40 msec premotor window we used should predict behavior better than activity taken from other times relative to the syllable. Since consecutive syllables are not independent, however, we expect that using other “premotor” windows to predict behavior would have some statistical power, though not as much as using the original premotor window.

We tested this prediction by repeating our analysis using two alternate premotor windows (supplemental Figure 11). In one analysis, we examined correlations between acoustic data and premotor neural activity taken from the previous syllable (light gray bars). In a second alternate analysis, neural data were taken from the syllable after the one from which premotor neural data were taken (dark gray bars). Comparison of the results with those of the original analysis (black bars) confirms the prediction that these alternate premotor windows have some predictive power, but significantly less power than the window used in the original analysis.
**Supplemental Figure 11** Correlations with adjacent syllables. Each bar shows the percent of cases correlated with a particular acoustic feature. The black bars show the results of the main analysis (also shown in Figure 7a in the main text), in which the acoustic features of a given syllable are regressed against the premotor activity before that same syllable. Also shown are the results two alternate analyses, in which the neural data are taken from the premotor window preceding the previous syllable (light gray bars) or next syllable (dark gray bars). In all nine cases, the percent of cases correlated is significantly greater than chance (dashed line). Asterisks indicate that the proportion of cases correlated with neural data from the same syllable is significantly higher than those derived from using either of the two alternate premotor windows (Z-test for proportions, p<0.0001).
Literature Cited