Quantitative assessment of vocal development in the zebra finch using self-organizing neural networks

Petr Janata[a)]
Department of Organismal Biology and Anatomy, University of Chicago, Chicago, Illinois 60637

(Received 19 April 2001; revised 14 August 2001; accepted 20 August 2001)

To understand the mechanisms of song learning by songbirds it is necessary to have in hand tools for extracting, describing, and quantifying features of the developing vocalizations. The extremely large number of vocalizations produced by juvenile zebra finches and the variability in these vocalizations during the sensorimotor learning period preclude manual scoring methods. Here we describe an approach for classification of vocalizations produced during sensorimotor learning based on self-organizing neural networks. This approach allowed us to construct probability distributions of spectrotemporal features recorded on each day. By training the network with samples obtained across the course of vocal development in individual birds, we observed developmental trajectories of these features. The emergence of stereotypy in sequences of song elements was captured by computing the entropy in the matrices of first- and second-order transition probabilities. Self-organizing maps may assist in classifying large libraries of zebra finch vocalizations and shedding light on mechanisms of vocal development. © *2001 Acoustical Society of America.* [DOI: 10.1121/1.1412446]

PACS numbers: 43.80.Ka [WA]

I. INTRODUCTION

Juvenile zebra finch (*Taeniopygia guttata*) males undergo a period of vocal development between ~30 and 90 days of age during which the spectrotemporal properties of their vocalizations change significantly. The developmental progression is typically divided into three stages: subsong, plastic song, and crystallized song (Arnold, 1975; Zann, 1996). During subsong, which lasts from approximately 30 to 50 d, the vocalizations are generally quiet, sustained, and without regular repetition of identifiable spectral features. Plastic song (50–80 d) is characterized by the emergence of identifiable song elements, such as harmonic stacks, whistles, and frequency sweeps that are stable across many bouts of singing. Additionally, the ordering of song elements begins to assume a more stable structure. Finally, the adult crystallized song consists of a set of spectrotemporal song elements, referred to as notes and syllables, which are arranged into fixed sequences called motifs, phrases, or strophes.

The morphology of notes, syllables, and motifs is usually quantified by human observers along several feature dimensions (Scharff and Nottebohm, 1991), and song elements are classified based on these features. Such manual approaches are extremely labor intensive. Partially automated methods have been developed for classification of song elements (Anderson *et al.*, 1996; Kogan and Margoliash, 1998). Nonetheless, these approaches still require selection of templates and training based on pre-selected vocalization examples. Another method for automated feature extraction has been developed recently for quantifying the similarity of songs, e.g., songs produced by a tutor and a pupil (Tchernichovski *et al.*, 2001, 2000). The advantage of the latter method is that it derives similarity indices from a set of derived spectrotemporal features without making any assumptions about song-element boundaries. Automated methods have the advantage of eliminating human subjectivity from the similarity judgments as well as their ability to quickly compare multiple exemplars. Both human and automated classification methods work well in the case of crystallized song which is characterized by a relatively circumscribed set of distinct song elements.

The quantification and classification of subsong and plastic song vocalization features present significant challenges, however. The increased variability in spectrotemporal characteristics of subsong and plastic song, as well as the extremely large number of vocalizations produced on any given day, preclude human scoring. Consequently, selection of ''representative'' exemplars from these developmental stages for use in behavioral or neurophysiological experiments is somewhat idiosyncratic. An alternative approach would be to segment and classify, with minimal human intervention, the entire corpus of vocalizations recorded for an individual bird. Such an approach would allow one to determine the prevalence of a spectrotemporal pattern on any given day, and to track the emergence and disappearance of spectrotemporal patterns across vocal development. Selection of representative vocalization exemplars could then be based on objective statistical principles.

As an initial step toward this goal, we decided to establish the feasibility of using a self-organizing neural network algorithm to extract and cluster the spectrotemporal patterns encountered across the various stages of vocal development in individual zebra finches. Such an approach has been used to classify cries of human infants (Schonweiler *et al.*, 1996), speech sounds in general (Leinonen *et al.*, 1993, 1992), as well as musical instrument timbres (Toiviainen, 1996).

[a)]Present address: Department of Psychological and Brain Sciences, 6207 Moore Hall, Dartmouth College, Hanover, NH 03755. Electronic mail: petr.janata@dartmouth.edu

TABLE I. Summary of birds used in the experiment.

Bird	Age (d) at isolation	Age at last recording	Total #fragments analyzed
nj6	29	51	51 226
nj7	31	41	122 545
nj9	34	51	154 319
nj10[a]	30	91	562 077
nj11[a]	30	162	523 414
nj12	30	44	219 880
nj13	30	49	159 268
nj14	30	43	99 702
nj17[a]	34	82	567 558
nj19	30	50	222 925
nj22	37	53	199 198
nj23	54	58	60 168
nj24	55	62	143 728
nj25	60	65	49 016
nj26	34	46	176 378
zf_bk480	120+		41 200
zf_bk520	120+		15 056
zf_bk526	120+		48 191

[a]Indicates a juvenile whose vocalizations were recorded through adulthood.

II. METHODS

A. Experimental animals

The vocalizations of 18 male zebra finches (15 juveniles, 3 adults) were studied. Twelve juveniles were obtained from their home cages in our breeding colony at approx. 30 d of age (range: 30–37 d, mean 32 d). These birds were used for companion neurophysiological experiments. Each of these birds was removed at a different stage of vocal development and not returned to the experiment. Thus all recordings were obtained prior to any neurophysiological recording. The vocal development of three of these birds (nj10, nj11, nj17) was tracked into adulthood. Three additional juveniles were removed from their home cages between 54 and 60 days of age, several days prior to neurophysiological experimentation. As their vocalizations were recorded for several days, their data were included in order to increase the sample size in the 55–65 d age range. The vocalizations of three adult birds (>120 d) were recorded for 4–9 days so that the data analysis procedures described below could be assessed and tested using crystallized song. The age and duration of isolation of each bird is summarized in Table I.

Until the time of removal, juveniles were housed with both parents and any siblings. Following removal from the home cage, each bird was housed alone in a sound-attenuating chamber (Industrial Acoustics Corp.) with unrestricted access to food and water, and was maintained on a 14/10 h light/dark cycle. All animals were housed and treated according to protocols approved by the University of Chicago Institutional Animal Care and Use Committee.

B. Song collection

Sounds in the chamber were monitored continuously by means of a microphone (Model 33-2011, Realistic) suspended above the cage in the sound-isolation box. The signal was amplified and filtered (500 Hz high-pass; 10 000 Hz low-pass) with custom-built electronics (JFI Electronics, University of Chicago). The signals were digitized with 16-bit resolution at 20 000 samples/s (atMIO16x card, National Instruments) using custom software (Amish Dave, University of Chicago). Those signals that exceeded a specified amplitude threshold [typically twice the ambient root-mean-square (rms) of the signal] at least once during a 30 ms window in 10 out of 12 consecutive windows were recorded as an entry to computer disk. Recording of the entry stopped when the signal failed to cross threshold for 300 ms. Both short (<1 s) and long (>20 s) vocalization periods were captured with these settings. On rare occasions, the recording system would fail, resulting in gaps of one or two days in the vocalization database for any given bird.

Prior to the automated data analysis of the vocalizations, the spectrograms of all entries in the data files were visually inspected. Many entries consisted primarily of artifacts, e.g., cage noises, wing-flapping, and rustling of the food dish. These entries were excluded from the final dataset. Although this step was extremely time consuming, typically requiring 1–2 h of manual scoring for each day's vocalizations from a single bird, it was necessary in order to reduce the size of the original dataset to fit within computational constraints. Final reduced datasets ranged in size from several hundred megabytes to several gigabytes. The overall duration of identified song fragment sequences extracted from these datasets averaged 6.5 h/bird.

C. Automatic song parsing

All data analyses in this report were scripted in MATLAB (Mathworks, Natick, MA), and used functions in the Signal Processing, Neural Networks, and Statistics Toolboxes. Analyses were performed using a computer with 500 Mb RAM, running Linux on a 450 MHz Pentium II processor. The amplitude envelope of the waveform recorded in each entry was used to identify acoustic fragments that could serve as input data to the self-organizing map (SOM) algorithm. The signal was full-wave rectified and low-pass filtered (150 Hz) using a 5-pole Butterworth filter. A heuristic was empirically established to find those samples in the rectified and filtered waveforms that might constitute an amplitude peak (acoustic fragment). For most recorded entries, in which both sounds and extensive silent periods were present, the threshold criterion was set to be the median value in the signal. In some entries few silent periods were present, causing the criterion value to be set too high, resulting in the loss of many valid entries. Thus when the ratio of the mean and median values for the entry was <2, the criterion value was set to be one-fourth of the median value. Runs of samples that exceeded the threshold continuously for at least 10.5 ms were tagged as acoustic fragments that would enter into subsequent analyses. The continuity threshold was selected after inspecting the parsed data of several birds. It eliminated a large number of fragments that appeared unrelated to vocalizations while retaining the very short whistles that where observed in the vocalizations of some birds. In addition to retaining the waveform of each fragment, the onset and offset timing information about each fragment was preserved for subsequent use in identifying fragment sequences.

D. Self-organizing map training

A self-organizing neural network was presented with randomly selected exemplars from the bird's vocalization library in order to determine a mapping of acoustic features onto a vector of output units. The training set consisted of 20% of the total number of fragments for each bird taken across all days. To facilitate equal representation of vocalizations produced on different days, the maximum allowable number of fragments from each day contributing to the training set was equal across days. If the number of fragments for any given day was smaller than the daily allocation, all available fragments from that day were used. This happened only rarely, typically in the initial days of isolation in juveniles. Normally, a random sample of fragments was chosen from each day.

For two of the juvenile birds, whose vocalizations were monitored into adulthood (nj10, nj11), fragments recorded every second day between the ages of 50 and 60 d and every fifth day between 60 and 90 d were entered into the analyses. The sparser sampling was deemed adequate given the greater stability of vocalizations in these age ranges, and it precluded disproportionately weighting the random sample of training exemplars toward these ages.

1. Preprocessing of acoustical fragments

Input vectors to the neural network were time-frequency representations (spectrograms) of the acoustic fragments. Input vectors to the SOM were required to be of equal length, so it was necessary to specify a maximum fragment duration for each bird. Distributions of fragment duration (e.g., Fig. 2) showed that the proportion of long fragments was small. Thus in the interest of computational efficiency, fragments exceeding a criterion threshold were excluded from the analysis. For each bird, the threshold was fixed. Averaged across birds, the thresholds were 286.7 ± 58 ms (mean$\pm$std. dev.). On average, 98.78 ± 1.11% of the fragments for a bird were shorter than the criterion and included in the training and classification sets.

Each fragment in the training set constituted a single input vector to the training algorithm. First, the fragment was filtered with a fifth-order Butterworth filter (800 Hz high-pass; 8000 Hz low-pass settings). Fragments shorter than the established maximum fragment duration for each bird were padded with zeroes to achieve the proper length. Next, a spectrogram of the fragment was computed (specgram function in MATLAB) using a window length of 12.8 ms with 75% overlap between successive windows. A Hanning window was applied to each portion of the waveform before the Fast-Fourier Transform (FFT) was computed. In order to increase the temporal resolution in the input vector, while keeping the input vector's size tractable, values in successive pairs of frequency bins of the spectrogram were averaged, e.g., bin 1 and 2, bin 3 and 4, etc., thus yielding an effective frequency resolution in the spectrogram of 156.25 Hz/band. Only frequency bins in the range from 800 to 8000 were included in the input vector, as these were within the bandpass region of the filtering stage described above. The modified spectrogram was then ''unfolded'' to create a one-dimensional vector in which the spectra of successive time windows were laid end-to-end. Thus the length of the input vector corresponded to the number of averaged frequency bins (46 bins) multiplied by the number of time windows (e.g., 125 time windows), where the number of time windows differed for each bird depending on the fragment duration cutoff. Each input vector was normalized by the maximum value in that vector so that all input vector values would fall within a range from 0 to 1.

2. Network parameters

Self-organizing maps for each bird were created using the SOM functions in the Neural Networks Toolbox (Revision 1.3) in MATLAB. Briefly, the architecture consisted of a one-dimensional input layer connected to a one-dimensional output layer through a single layer of weights. Several output layer sizes and topologies were explored in several juvenile and adult birds to determine whether higher-dimensionality in the output layer facilitated classification of the song fragments. Output unit topology did not appear to influence the distributions of correlations between input vectors and the weight vectors connecting them to the winning output units. Therefore, for ease in displaying and interpreting the weight matrices, we settled on linearly arrayed output units. For adult birds we used 64 output units, and for juveniles we used 200. We used a smaller output vector for adults because the spectrotemporal variability in crystallized song is smaller, and presumably adequately represented with a smaller number of output units, than is the variability in juvenile subsong and plastic song.

Every input unit element was connected by a weight to every element in the output vector. Thus the weight matrix for a juvenile bird who had 5750 elements in the input vector contained 1 150 000 elements. Weights were initialized to random values. The weight matrix was updated through a competitive (''winner-take-all'') learning algorithm. Default values were used for the learning rates during the ''ordering'' phase (starting value of 0.9) and ''tuning'' phase (0.02). During the ordering phase, the size of the neighborhood in which weights were modified was gradually reduced in equal steps from the maximum distance between output units to a neighborhood of one unit. Similarly, the learning rate was reduced in equal steps from the starting value to the tuning phase value. Two-thirds of the training set were randomly selected and used in the ''ordering'' phase and the other third was used for the ''tuning'' phase.

E. Classification of song fragments and characterization of output unit loadings

Once the network had been trained for each bird, all identified fragments for the bird including those in the training set were classified. Each fragment was transformed into an input vector representation and correlated (Pearson correlation coefficient) with the weight vectors mapping the input vector to the output vector. The output unit associated with the vector of weights that correlated most highly with the fragment's spectrogram representation was chosen as the winning output.

Daily loading matrices were constructed for each bird by tallying the number of times each output unit was activated by that day's song fragments and dividing by the total number of fragments produced during that day.

F. Characterization of fragment sequences

1. Sequence identification

After the recordings had been parsed, sequences of acoustic fragments were identified as follows. Each entry in the recordings that contained multiple fragments was used for this purpose. A fragment was included in a sequence if its onset occurred within a criterion inter-fragment interval (IFI), measured from the offset of the preceding fragment. The criterion IFI was arbitrarily selected based on the IFI distribution for each bird, and was chosen to fall along the long tails of the distribution. In most cases, IFIs of 200 ms were used. This value was based on the empirical observation that the silence between fragments within motifs (in adults) or bouts (in juveniles) was less than 200 ms, and that longer intervals represented motif or bout boundaries.

Following SOM training and classification of every fragment, the identified fragment sequences were recoded as sequences of output units by replacing the identity of each fragment in the sequence with the output unit that it was classified under. In those cases where the fragment was not associated with an output unit of the network, i.e., if the duration of the fragment was too long, the fragment was assigned to an extra element in the output vector specifically used for these cases.

2. Transition probability matrices and entropy estimation

Once sequences of output unit activations had been identified it was possible to construct transition probability matrices (TPMs). For each day's vocalizations, a first-order TPM was constructed by tallying all first-order transitions. Each row in the TPM indexed the first of two sequence elements and the column indexed the second element. For example, the sequence {34, 10, 22, 34, 10} consists of four first-order transitions (34,10; 10,22, etc.) and would increment values in three elements of the TPM. The TPMs reflect the most frequent transitions between pairs of acoustic fragments. Second- and third-order TPMs were also computed for fragment triplets and quadruplets, respectively. Each row in a second-order TPM indexed a pair of fragments, and the different elements in that row indicated the overall likelihood of observing each of the different fragments following the particular pair of fragments. Third-order TPMs were similarly constructed.

Structure in each day's TPM, P, for each bird was quantified with the information theoretic measure of entropy, H:

$$H = \sum_{i=1}^{N} \sum_{j=1}^{N} -P_{i,j} \cdot \log_2 P_{i,j},$$

A normalized entropy value, H^*, was obtained using the total number of nonzero elements in the TPM:

$$H^* = \frac{H}{\log_2 g},$$

where g is the number of nonzero elements in the TPM.

The maximum possible normalized entropy value was $H^* = 1$, regardless of the number of nonzero elements in the TPM. This value would be obtained if all observed transitions were equally likely to occur. In order to estimate whether the structure in the observed TPMs differed from random probability distributions, we calculated simulated entropy values that would be expected given random sets of transition probabilities. For these simulations, we used 20 random vectors containing the same number of elements as the number of nonzero entries in the daily TPM for each bird.

III. RESULTS

A. Parser performance

The number of fragments identified for each bird is shown in Table I. Figure 1 shows examples of the parser's performance on the song of one juvenile (nj14) recorded at 32 and 43 days of age. Both the juvenile and the specific examples were selected randomly. The parser's performance did not always match the parsing by a human scorer. In cases where the amplitude threshold was set too high, a syllable consisting of two notes would be split into its constituent notes, whereas in other instances the syllable would be retained as a single unit. Similarly, if the heuristic resulted in a threshold that was set too low for a given entry, several seemingly separate elements would be grouped into a single element. Detailed inspection of the waveforms showed that often the transition between what appeared to be two closely apposed notes in a syllable were in fact separated by a very brief low-amplitude period. In such cases, the parser correctly identified the two acoustic events as separate, even though the expert human scorer of zebra finch song would tend to integrate the two events into a single, higher-order, event. Overall, we felt that the occasional ''errors'' of the parser were mitigated by its ability to efficiently process the immense volume of the data according to strict objective criteria. For example, the fragments shown in Fig. 1(B) represent merely 0.1% of the total number of fragments identified for this bird.

B. Duration statistics of zebra finch song fragments

One characteristic of crystallized zebra finch song is stable song element duration. This is exemplified in Fig. 2(A), in which the distributions of song fragment durations remained stable across four consecutive days. Not surprisingly, songs of different adults are characterized by different song fragment duration distributions [Figs. 2(A), (B)]. The cumulative duration distributions for 13 juveniles are shown in Fig. 2(C). Many of the distributions show peaks at various durations. However, the cumulative distributions mask any daily variation in fragment duration that may occur across the course of song development. Figure 2(D) illustrates the variability in fragment duration across song development in two juveniles. Both examples illustrate that by 75 days of

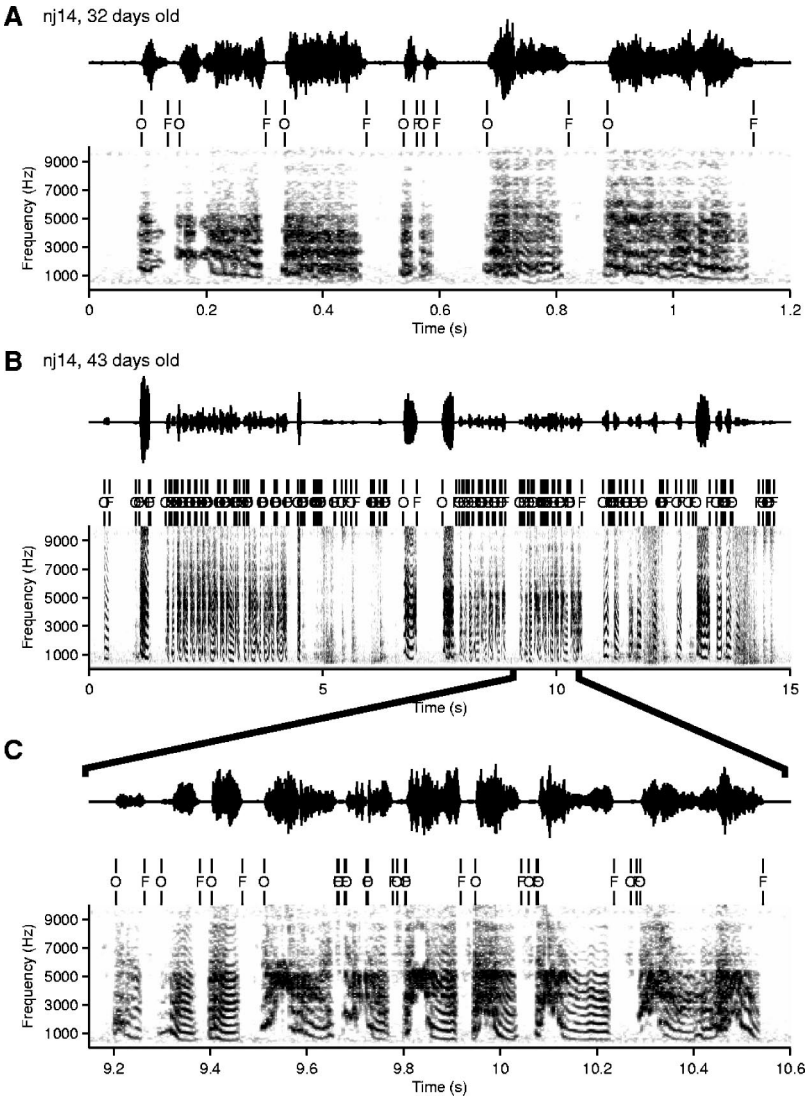

FIG. 1. Example of the parsing algorithm's performance. Two vocalization epochs from different developmental stages of the same zebra finch were selected at random. In each panel, the oscillogram at the top shows the amplitude fluctuations in the vocalization. Below it are shown the onset (O) and offset (F) marks for each fragment found by the parsing algorithm. The spectrogram is shown at the bottom. (A) A 1.2 s epoch of subsong recorded at 32 d of age. (B) A 15 s example of plastic song recorded from the same bird at 43 d of age. Cage noises, are evident in the recordings between 5 and 6 s (hopping), and again around 14 s (wing-flapping). The parsing algorithm had no way of distinguishing between vocalizations and cage noises. The 146 fragments shown in this epoch represent 0.1% of the total number of fragments identified for this bird. (C) An expanded view of the parser's output corresponding to a ~1.5 s segment in (B).

age the daily duration distributions contained a small number of distinct peaks. In the case of nj11, fragment durations were more uniformly distributed prior to day 45. The duration images also show that some peaks in the duration distributions shifted gradually along smooth trajectories after they initially formed, and some trajectories appeared to bifurcate.

C. Properties of the SOMs

The primary goal in utilizing a self-organizing network for the analysis of juvenile song fragments was to obtain an automatic classification of the various spectrotemporal characteristics present in the extremely large fragment dataset. Once the weight matrix linking the spectrogram representation with output categories had been established using a random subset (20%) of the fragments, the relative abundance (loading) of fragments in each output category was determined by correlating every fragment with each row in the weight matrix and assigning it to the output unit linked to the most highly correlated row in the weight matrix. The loading on each output unit could then be examined as a function of the bird's age.

The SOM approach was first tested on crystallized songs from adult zebra finches. Figure 3(A) illustrates a weight matrix for a single zebra finch, and the loadings on each output unit during each of the four days that this bird's song was recorded. Inspection of the weight matrix shows that clusters of adjacent output units coded similar features in the input vectors. For example, the rows of weights corresponding to output units 1–20 look very similar to each other, as do weights corresponding to output units 22–39, 40–42, and 57–64. Figure 3(B) illustrates that rows in the weight matrix form a very literal representation of the spectrotemporal features of the input vectors. Each panel shows a row (one-dimensional vector combining ''frequency'' and ''time'') of the weight matrix reshaped as a spectrogram (a frequency $\times$time matrix). The resulting spectrograms show identifiable zebra finch song elements. Since a row of weights maps the entire input vector onto a single output unit, the weights in a row of the matrix can be thought of as the components of a feature detector. Thus if an input vector, representing a set of spectrotemporal features, is correlated strongly with the row of weights, the output unit corresponding to that row of weights will be strongly activated.

Figure 4 shows weight matrices and daily weight matrix

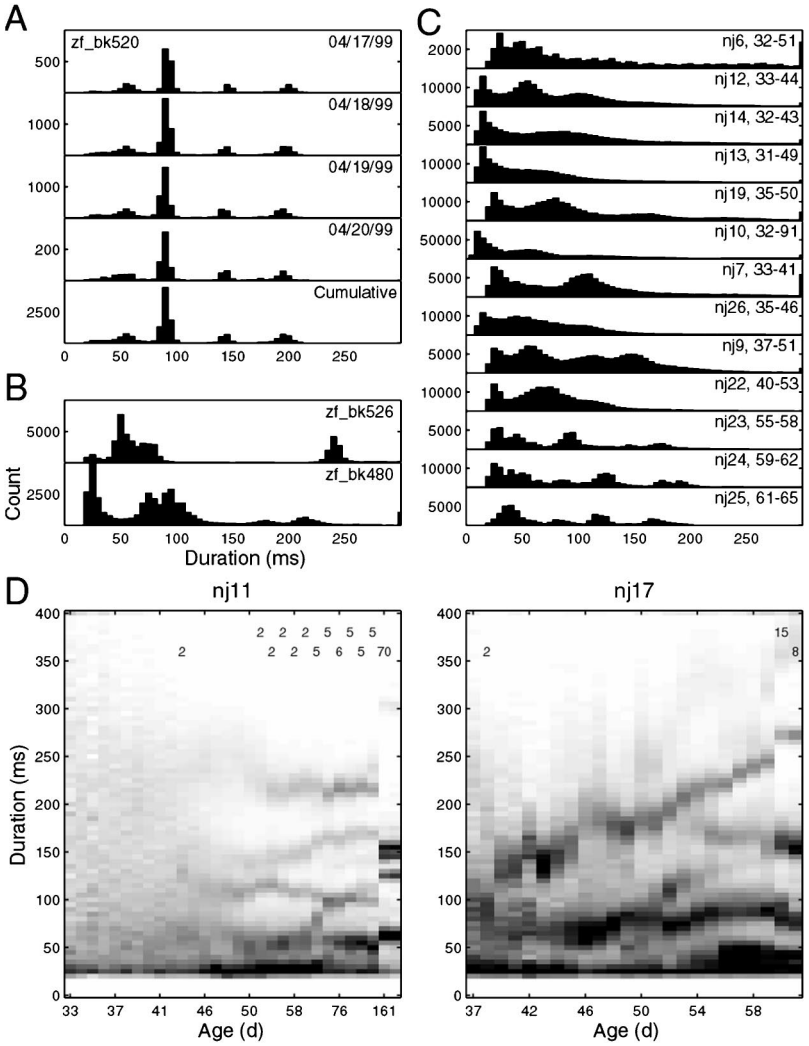

FIG. 2. Distributions of song fragment durations in adult and juvenile zebra finches. (A) Histograms showing the number of fragments observed for durations marked along the abscissa. Duration histograms for fragments recorded on each of four days from an adult zebra finch show little variation. (B) Cumulative duration histograms for two other adults. (C) Cumulative duration histrograms for 13 juvenile zebra finches. Each panel corresponds to the data for a single bird. The age range over during which the fragments contributing to the histogram were recorded is indicated at the top right of each panel. For example, 'nj9, 37–51', indicates fragments were recorded from bird, nj9, between 37 and 51 days of age. (D) Images of daily fragment duration distributions trace the evolution of fragment duration structure for two birds whose song was recorded across the period of sensorimotor learning. Gray scale intensity reflects the proportion of fragments for each duration on each day. The numbers at the top of each matrix indicate the number of days that were skipped between recording of fragments in the column with the number and the preceding column.

loadings for the three juveniles whose vocalizations were recorded over the course of their song development. Given the length of the input vector, the details of the spectrotemporal properties encoded by any given row of weights are lost when the weight matrix is viewed as a whole. Nonetheless, several properties of the weight matrix can be discerned at the coarse level. For instance, the lengths of the gray streaks in the weight matrices indicate which fragment durations different output units became sensitive to. Thus the gray-scale intensity profiles depict the overall temporal weighting functions applied to each fragment when calculating the best match. Short and long fragments tended to be represented at opposite ends of the output unit array.

Despite the coarse features represented in the overall view of the weight matrices, individual output units were sensitive to detailed spectrotemporal patterns (as shown in Fig. 3). The insets in Figs. 4(A) and (C) provide another example of the spectral features represented by the weights in four consecutive time windows. Figure 4(B) insets illustrate that nearby output units represent similar inputs, whose spectra during the same time windows differ primarily in the presence of a small peak at around 4 kHz in the plot of inset ii (see arrows). The daily output unit loading image for nj11 shows that the fragments activating the output units whose weight vector segments are plotted in the insets were recorded primarily between the ages of 47 and 51 d.

Together, the loading and weight matrices provide information about what the most common spectrotemporal features were in the song fragments produced at each stage of vocal development. Most striking about the loading matrices was the abruptness and magnitude with which some song features (as represented by the weight vectors) appeared. For example, for nj10, output units 170–190 were not loaded prior to 50 d, after which different members of the set were loaded for the remainder of the recording period. Conversely, output units that were loaded highly initially e.g., nj10, units 150–155 between 32 and 37 d, were loaded weakly or not at all following 60 d. For any given bird, the abrupt transitions did not all occur on the same day. For nj10, different spectrotemporal characteristics appeared at approximately 41, 44, 48, 50, and 54 days of age.

Figure 5 illustrates developmental trajectories captured by the weight matrices. For each bird, a portion of the weight matrix was selected that showed a gradual change in the distribution of loadings on adjacent output units across the course of several days. In other words, if adjacent output units represent subtle differences in their respective weight matrix rows (best-fitting spectrotemporal features), then loading of adjacent output units on successive days may represent gradual change in one or more features of song frag-

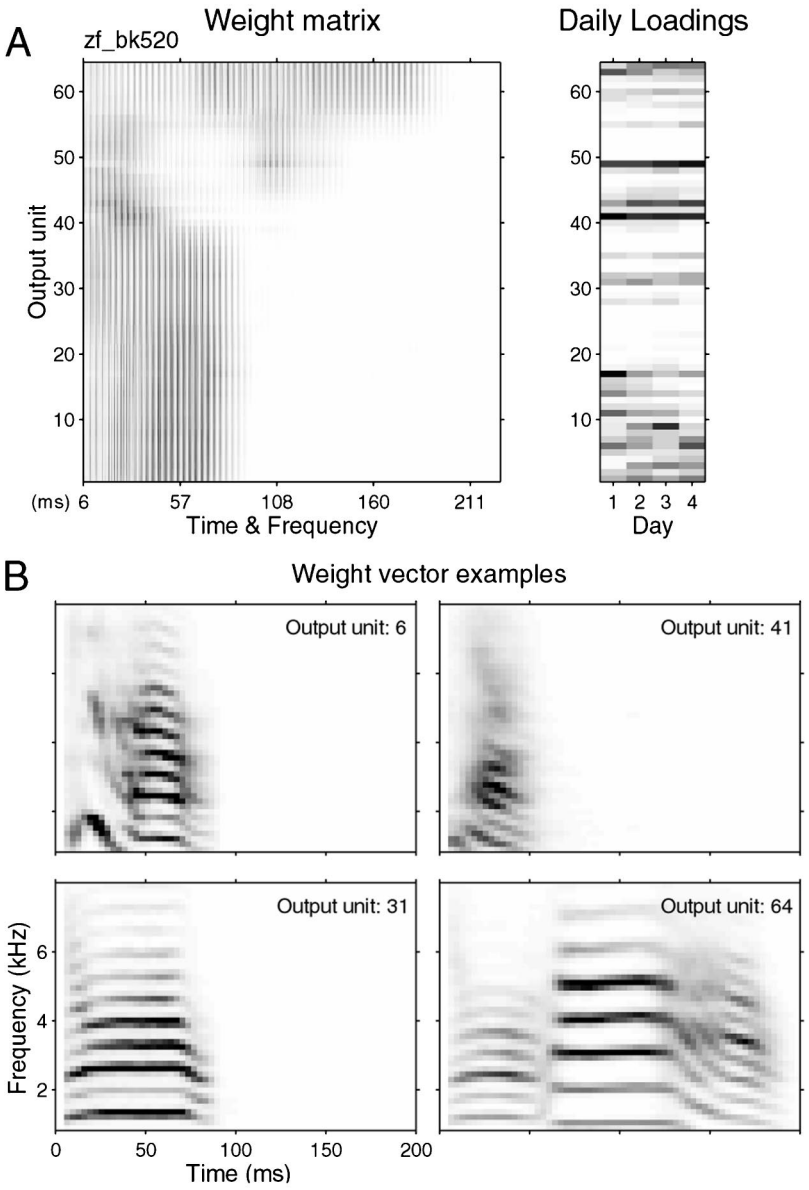

FIG. 3. SOM weight matrix and daily weight matrix output-unit loadings for an adult zebra finch. (A) The weight matrix shows the strength of association between all input units (columns labeled ''Time & Frequency'') and all output units (rows). Weight values are represented in gray scale intensity. The input layer corresponds to a spectrogram with the spectra of successive time windows laid end to end. The vertical striation in the weight matrices is a consequence of the ''unfolded spectrogram'' input representation. To the right of the weight matrix is a loading matrix which shows how often each output unit was activated on each day. Each row in the loading matrix corresponds to an output unit and is aligned with the corresponding output unit in the weight matrix to the left. The columns represent days. The gray scale intensity of each element in the matrix represents the proportion of fragments recorded on a particular day that was most strongly correlated with the output unit represented by that row. Darker values indicate a higher proportion. (B) Each panel shows the weight values of a row in the weight matrix in (A), rearranged into a two-dimensional spectrogram representation. These images make evident that the weights in the SOM adapt to represent specific features in the input vectors and that the output units serve as ''feature detectors'' for these specific features.

ments that most strongly activate those output units. Each of the song fragments shown in Fig. 5 was the fragment that correlated most highly with the particular output unit on the specified day. In the case of nj11, the harmonic stacks become more distinct with increasing age. The song fragments for nj17 show a more complex pattern of change, including both a lengthening of the component note elements and continued differentiation of the spectral features in the second half of the song fragment.

D. Analysis of produced sequences

Development of zebra finch song is characterized not only by the emergence and crystallization of the spectrotemporal features of individual song elements, but also by arrangement of these song elements into fixed, stereotyped sequences. In this study, song fragments were identified as belonging to the same sequence if the time between the end of one fragment and the start of the next was less than 200 ms. Classification of fragments into categories using the SOM facilitated the quantification of emerging structure in fragment sequences because of the sheer number of fragments and sequences that could be labeled automatically. All the sequences recorded on any given day were used to construct first-order TPMs which summarized for each output unit the likelihood that it would be followed by itself or some other output unit. Because many output units were not loaded on any given day, and because transitions were not observed between all possible pairs of output units, the TPMs were rather sparse for any given day. The most dense TPMs showed at least one transition for 12 500 out of 40 000 possible transitions, i.e., ~31% of the entries in a TPM had nonzero values. The number of overall fragments produced during a day and the number of nonzero entries in the TPM were significantly correlated (Fig. 6). A test of the difference in regression coefficients for the young and older birds showed that for any given number of fragments produced, however, the TPMs for older birds showed significantly fewer transitions (nonzero entries) $[F(1,240)=9.764,\ p$

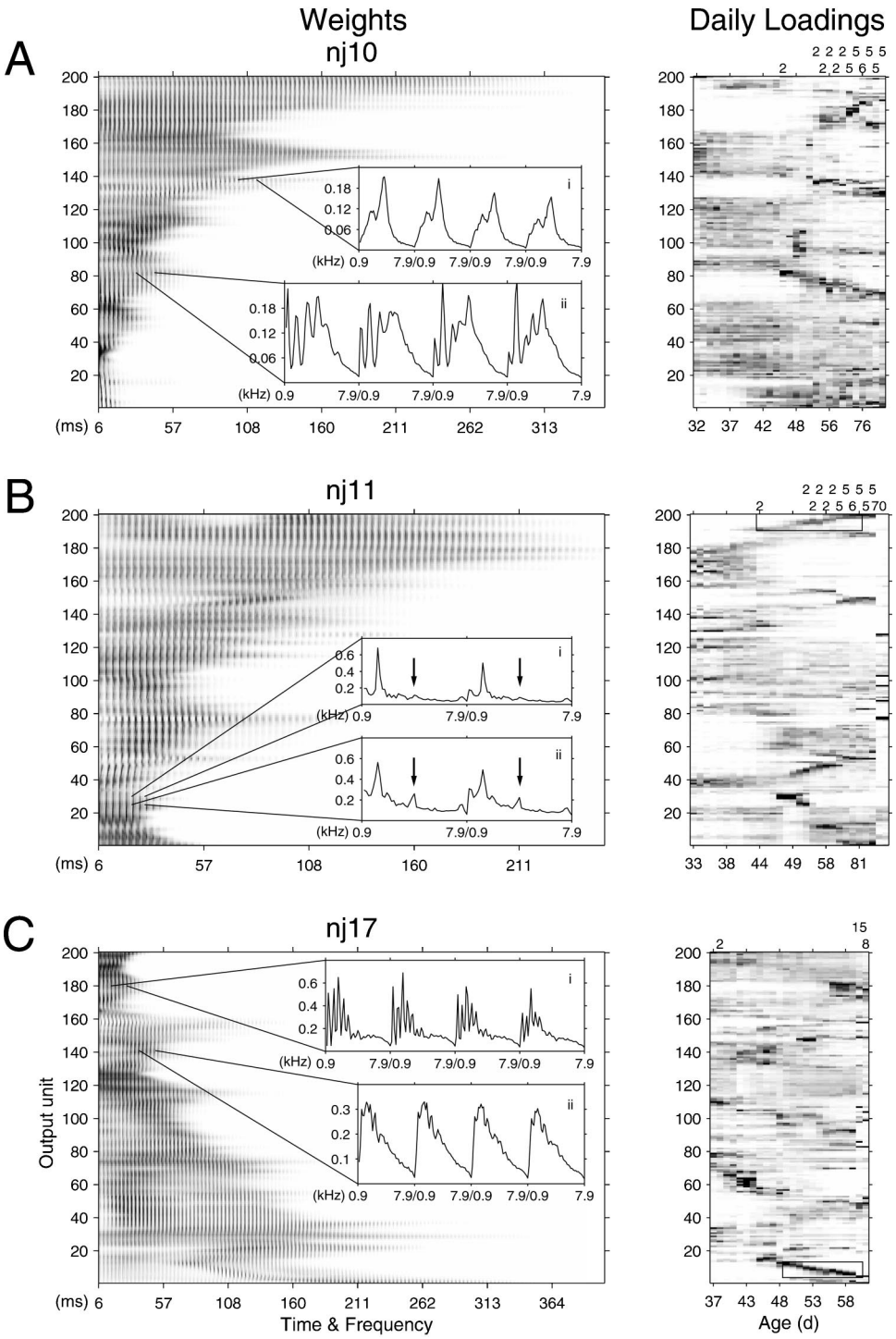

FIG. 4. SOM weight matrices and daily weight matrix output-unit loadings for three zebra finch juveniles whose vocal production was recorded across the bulk of the sensorimotor learning period. See Fig. 3 for a description of how to interpret the weight matrices and loading matrices. Each inset in the weight matrices plots the weights linking a section of the temporo-spectral input representation with an output unit. For example, inset A-i shows the spectral features in four successive time windows occurring at ~108 ms from fragment onset to which the subset of weights is sensitive. The matrices of daily output unit loadings (shown at the right) indicate that some output units represent song fragments produced in early stages of song development, whereas others encode fragments produced at later stages. For example, the output unit linked to the weights shown in inset C-i was primarily activated after the age of 56, whereas the output unit linked to the weights shown in C-ii was activated primarily by fragments produced between ages 43 and 48. The numbers above each loading matrix indicate the number of days that were skipped between recording of fragments in the column with the number and the preceding column. The boxes in the loading matrices for nj11 and nj17 enclose regions that form the basis for Fig. 5.

<0.002]. This indicated that a smaller set of transitions occurred more often in older birds, as would be expected of more stereotyped sequences.

We quantified the amount of structure in a TPM by computing the entropy in the TPM. Entropy is maximal if all transitions are equally likely to occur. Because the calculated entropy value was normalized with the number of nonzero elements in the TPM, i.e., the maximum entropy given the number of nonzero elements, the entropy values ranged from 0 to 1. Entropy in the daily loading matrices, i.e., in the distribution of probabilities of activating any given output unit increased until day 38 and then stabilized, indicating an increase in the diversity of activated output units. Entropy in the first- and second-order TPMs decreased with increasing age (Fig. 7, circles and diamonds). The entropy of the TPMs was compared with entropy values that would be obtained by assigning random probability values to the nonzero entries in the observed first-order TPMs (with the constraint that the probabilities sum to 1). Somewhat surprisingly, the entropy of the observed first-order TPMs was higher (0.9864 $\pm$0.0038 std. dev.) than for random first-order TPMs (0.9702 $\pm$0.0022 std. dev.) in the age range of 31–35 days. After 46 d of age, the entropy of the observed first-order TPMs was substantially lower than the entropy for random TPMs, and

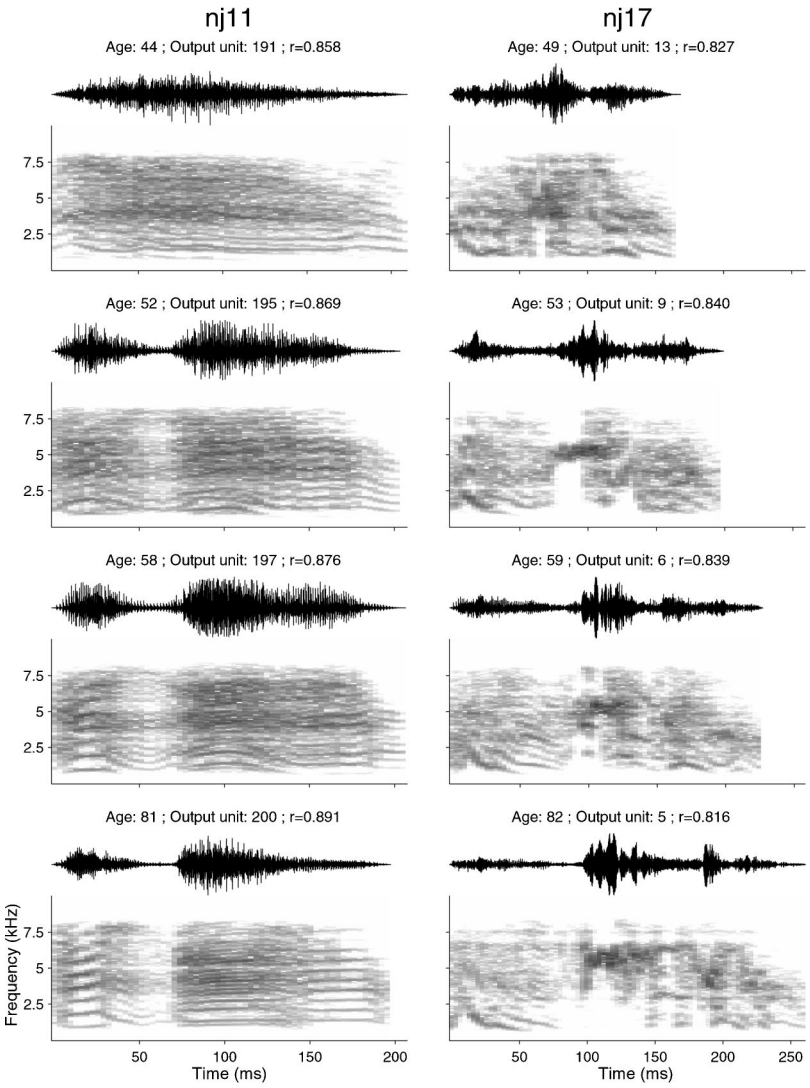

FIG. 5. Examples of song fragments that loaded nearby output units on successive days. Oscillograms and spectrograms on the left show the song fragment that was most highly correlated with the weights associated with the specified output unit on each of four days for bird, nj11. The age of the bird, output unit (row in the weight matrices shown in Fig. 4), and the magnitude of the correlation are specified in the title of each plot. Song fragments for bird, nj17, are shown on the right. Note the similarity and development of note features as the birds age (top to bottom).

decreased to ~0.7 in nj11 who was recorded until 162 days of age (data not shown). Second-order entropy also started to decrease after 46 days of age (Fig. 7, diamonds). The average TPM entropy for adults was 0.881 ± 0.003 (std. dev.), 0.948 ± 0.018, and 0.983 ± 0.009 for first-, second-, and third-order transitions, respectively.

IV. CONCLUSIONS

Using a simple input representation of zebra finch vocalizations (the amplitude component of FFT-based spectrograms), and a simple self-organizing neural network architecture consisting of a single weight layer and one-dimensional output vector, we generated maps of individual zebra finch vocalization histories. When reconstituted as spectrograms, rows of connection weights mapping the spectrotemporal input vector to output units appeared as plausible song elements. This indicated that the SOMs had extracted the most prominent spectrotemporal features in the song fragment database for each bird. Using the SOMs, automated classification of tens to hundreds of thousands of song fragments from individual birds enabled us to generate a statistical description of which features were present when during vocal development. The method identified the emergence and disappearance of spectrotemporal features that had come to be represented in the weight matrix. In many cases, adjacent output units of the SOMs were heavily loaded on successive days, forming identifiable trajectories in the daily output loading matrices. Trajectories in these matrices appear to represent development trajectories of spectrotemporal features of song elements.

Tens of thousands of sequences were automatically labeled and transition probabilities between sequence elements were calculated. The entropy in the first-order transition probability matrices decreased with increasing age of the bird, indicating that the ordering of song elements, as represented in the SOM output layer, became less random as the bird's vocalizations developed. This is in agreement with qualitative observations of increased sequence stereotypy as the zebra finch crystallizes his song. The drop in first-order TPM entropy, averaged across animals, around day 45 corresponds well to the observed juncture between subsong and plastic song stages of song development (Zann, 1996).

While the SOM approach provides a convenient means of reducing extremely large datasets of vocalizations, what insights into the vocal development process do the resulting SOMs and derived sequence entropy measures provide? As

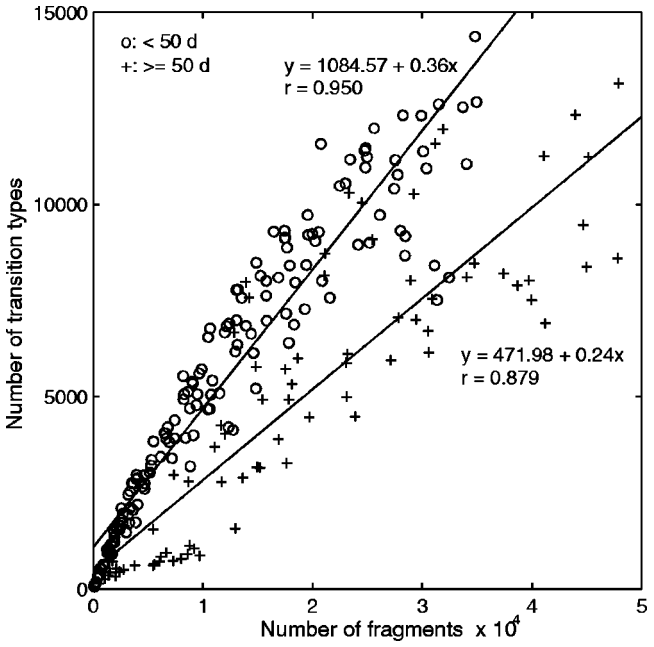

FIG. 6. Relationship between the number of fragments recorded during a day and the number of different first-order transitions between output units activated by those fragments. Circles represent data for fragments produced before 50 d of age and crosses correspond to fragments produced after 50 d of age.

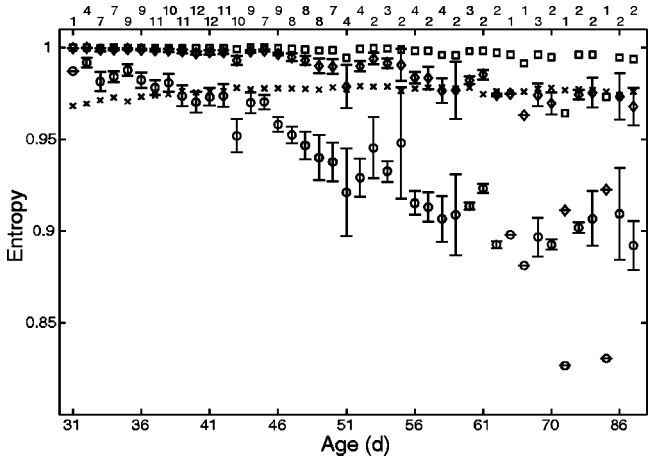

FIG. 7. Entropy of song fragment sequences as a function of age. Circles indicate average entropies of the daily first-order transition probability matrices (TPMs), diamonds correspond to entropy in second-order TPMs, and squares correspond to entropy in third-order TPMs. The entropy of a TPM was normalized using the number of nonzero entries in the TPM. If all observed transitions were equally likely, entropy would equal 1. Crosses indicate simulated entropy values that are obtained if the distributions of probabilities in the first-order TPMs are randomly determined using the number of non-zero entries in the observed data. Error bars represent ± 1 S.E.M. In order to reduce clutter in the figure error bars are not shown for the simulated and third-order data because they rarely extend beyond the bounds of the plotted symbols. The number of birds contributing to the entropy estimate for each day is shown above the plot.

indicated above, fragment sequences labeled using the SOM showed increased stereotypy with increasing age, mirroring qualitative descriptions of zebra finch vocal development. To our knowledge, the SOM results provide the first quantitative estimates of when changes in sequence structure occur, based on a nearly exhaustive sampling of the vocalization history. Aside from the ability to quantify changes in sequence structure, the SOM weight matrices and associated daily loading matrices suggest that two types of developmental phenomena are captured by the SOM approach. The first type represents developmental trajectories that arise from gradual changes in spectrotemporal features across several days. This type of trajectory is captured by virtue of the SOM algorithm modifying not only the weights between the input vector and the most highly activated output unit, but also the weights of neighboring units. This leads to a clustering of output units whose connection weights are very similar and represent similar spectrotemporal features in the input layer. The observation that adjacent output units were maximally loaded on successive days in the daily loading matrices indicates that spectrotemporal features were changing subtly across time.

The other type of phenomenon is the sudden appearance of new spectrotemporal features. This was observed as heavy output unit loading starting on one day and then continuing on successive days, with no loading of the output unit on previous days. In other words, the spectrotemporal features represented by the output unit came into sudden existence in the vocalization database. It was possible to represent such features in the weight matrices because fragments sampled equally from all stages of development were presented randomly to the neural network during the training phase. Thus all frequently occurring spectrotemporal feature categories had a chance of establishing themselves in the weight matrix.

The notion that the both the gradual trajectories and sudden loadings in the daily weight matrix loading maps represent actual modes of vocal development in the zebra finch derives from the recent work of Tchernichovski and colleagues (2001). In their analysis, zebra finch vocalizations are decomposed into four feature parameters: Wiener entropy, spectral continuity, pitch, and frequency modulation (Tchernichovski *et al.*, 2000). At any given developmental stage, a vocalization is described by the relative magnitudes of these four parameters. A similarity index for vocalizations recorded at different developmental stages is computed by comparing the distributions of values on these four parameters. When assessing the imitation of tutor song elements presented to juvenile birds under highly constrained operant learning conditions, they found that song learning, i.e., modification of the four song features, proceeded along ''direct'' and ''indirect'' routes (Tchernichovski *et al.*, 2001). In the ''direct'' imitation trajectories, the features changed gradually, whereas in the ''indirect'' trajectories, the pitch feature would change gradually until a critical point at which the period would suddenly double. Further work is needed to establish whether the trajectories observed by Tchernichovski *et al.* are similar to the trajectories appearing in the daily weight matrix loadings.

A potentially promising approach would be to merge elements of both methods. For example, input vectors could be built for each fragment from values on each of the four feature dimensions described above. The role of the SOM would be to extract the organization of this four-dimensional feature space across the developmental history of the individual bird. One advantage of this approach might be a re-

duction in the computation time resulting from a reduction in the size of the input layer because each time window in the fragment would be represented by only four feature parameters rather than a larger number of frequency bins.

In performing the analyses described in this paper we identified several steps that call for further improvement. First among them is eliminating the need to manually separate those recorded entries containing song from those containing predominantly calls and cage noises. In some cases, entries with cage noises and/or calls represented 50% of the recordings. This resulted in several problems. First, if entries containing cage noises and calls were parsed and added to the fragment database, the database became unmanageably large requiring several gigabytes of disk storage for each bird and several hundred megabytes of RAM for handling the data structures. More significantly, if allowed into the final dataset, cage noises and calls would comprise a disproportionate amount of the SOM training set, since training was based on a random sample of 20% of the acoustical fragments in the final dataset. Thus we found it necessary to discard entries containing primarily cage noises and calls. For a trained scorer this process took approximately 1 h for each day's recordings from a single bird.

In our analyses, we adopted an architecture with output units arranged along a single dimension. We did this primarily because the weight matrix is easier to look at and understand with this type of architecture. However, it is reasonable to assume that the optimal organization of spectrotemporal features in developing zebra finch song might be captured more accurately with higher dimensionality in the output layer. As a preliminary test of this idea, we used the data from an adult bird to train SOMs with either a one-dimensional output vector with 64 units, or a six-dimensional output vector with 2 units along each dimension (64 units total). Following training, we correlated each fragment with the respective weight matrices and constructed the distributions of maximal correlations. There appeared to be no difference in the correlation distributions (data not shown) so we did not pursue this issue further and chose instead to illustrate the method on the one-dimensional case.

Given our initial results, we believe that self-organizing neural networks promise to be a useful tool for the objective categorization of zebra finch vocalizations recorded over the course of vocal development.

ACKNOWLEDGMENTS

The research was supported by the following NIH grants: F32 NS10395 to P.J., R01 NS25677 to Daniel Margoliash, and P50 NS17778-18 to Jamshed Bharucha. The help of Alphi P. Elackattu in the screening of recordings was greatly appreciated. Patrice Adret and Timothy Q. Gentner provided helpful comments on an earlier version of the manuscript.

Anderson, S. E., Dave, A. S., and Margoliash, D. (**1996**). ''Template-based automatic recognition of birdsong syllables from continuous recordings,'' J. Acoust. Soc. Am. **100**, 1209–1219.

Arnold, A. P. (**1975**). ''The effects of castration on song development in zebra finches (*Poephila guttata*),'' J. Exp. Zool. **191**, 261–278.

Kogan, J. A., and Margoliash, D. (**1998**). ''Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study,'' J. Acoust. Soc. Am. **103**, 2185–2196.

Leinonen, L., Hiltunen, T., Torkkola, K., and Kangas, J. (**1993**). ''Self-organized acoustic feature map in detection of fricative-vowel coarticulation,'' J. Acoust. Soc. Am. **93**, 3468–3474.

Leinonen, L., Kangas, J., Torkkola, K., and Juvas, A. (**1992**). ''Dysphonia detected by pattern recognition of spectral composition,'' J. Speech Hear. Res. **35**, 287–295.

Scharff, C., and Nottebohm, F. (**1991**). ''A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: Implications for vocal learning,'' J. Neurosci. **11**, 2896–2913.

Schonweiler, R., Kaese, S., Moller, S., Rinscheid, A., and Ptok, M. (**1996**). ''Neuronal networks and self-organizing maps: New computer techniques in the acoustic evaluation of the infant cry,'' Int. J. Pediat. Otorhinolaryngol. **38**, 1–11.

Tchernichovski, O., Mitra, P. P., Lints, T., and Nottebohm, F. (**2001**). ''Dynamics of the vocal imitation process: How a zebra finch learns its song,'' Science **291**, 2564–2569.

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., and Mitra, P. P. (**2000**). ''A procedure for an automated measurement of song similarity,'' Anim. Behav. **59**, 1167–1176.

Toiviainen, P. (**1996**). ''Optimizing auditory images and distance metrics for self-organizing timbre maps,'' Journal of New Music Research **25**, 1–30.

Zann (**1996**). *The Zebra Finch* (Oxford University Press, New York).