Statistical learning of recurring sound patterns encodes auditory objects in songbird forebrain

Kai Lu[a] and David S. Vicario[b,1]

[a]Institute for Systems Research, University of Maryland, College Park, MD 20740; and [b]Psychology Department, Rutgers University, Piscataway, NJ 08854

Edited* by Mortimer Mishkin, National Institute for Mental Health, Bethesda, MD, and approved August 25, 2014 (received for review June 27, 2014)

Auditory neurophysiology has demonstrated how basic acoustic features are mapped in the brain, but it is still not clear how multiple sound components are integrated over time and recognized as an object. We investigated the role of statistical learning in encoding the sequential features of complex sounds by recording neuronal responses bilaterally in the auditory forebrain of awake songbirds that were passively exposed to long sound streams. These streams contained sequential regularities, and were similar to streams used in human infants to demonstrate statistical learning for speech sounds. For stimulus patterns with contiguous transitions and with nonadjacent elements, single and multiunit responses reflected neuronal discrimination of the familiar patterns from novel patterns. In addition, discrimination of nonadjacent patterns was stronger in the right hemisphere than in the left, and may reflect an effect of top-down modulation that is lateralized. Responses to recurring patterns showed stimulus-specific adaptation, a sparsening of neural activity that may contribute to encoding invariants in the sound stream and that appears to increase coding efficiency for the familiar stimuli across the population of neurons recorded. As auditory information about the world must be received serially over time, recognition of complex auditory objects may depend on this type of mnemonic process to create and differentiate representations of recently heard sounds.

multielectrode | electrophysiology | single-unit | memory | novelty

A central question in neuroscience is how a sensory system combines the transduced features of a sensory experience into a unitary object that is recognized and available for further processing. In the field of auditory research, progress has been made on the mechanisms that encode basic acoustic features, but it still not clear how multiple sound components are integrated over time and recognized as a single object. A striking example is that, after only 2 min of passive exposure to a continuous stream of artificial words (nonsense sound sequences), infants extract and distinguish these "words" from other sequences based on the transition probability between syllables; this provides an explanation of how actual words may be initially recognized by infants, even though obvious boundaries (e.g., interword pauses) are absent in natural speech (1). Further work with similar passive exposure paradigms showed that human adults and infants could also learn to distinguish consistent but nonadjacent patterns (2, 3), and were thus able to learn invariant relationships despite variable intervening sounds. These experiments demonstrate that the brain can form a record of the recurring patterns in an ongoing stimulus stream. This occurs spontaneously in the absence of reinforcement or of any cues to what should be learned, and is referred to as statistical learning. The ability to extract invariant patterns through passive exposure is not unique to the speech/language system, because behavioral studies with similar paradigms in nonhuman primates, rats, and birds produced related results (4–7). In addition, human subjects can also learn the statistical regularities in nonlinguistic materials (8, 9).

Despite the many techniques used to explore the mechanisms of statistical learning in audition, [e.g., EEG, functional MRI (fMRI)] (10–11), study at the single-neuron level remains to be done. We have addressed this problem by studying neurons in two auditory areas of the songbird forebrain where responses are modulated by passive experience: the caudomedial nidopallium (NCM) and the caudolateral mesopallium (CLM), which receive inputs from the thalamorecipient field L, and thus may correspond to superficial layers of mammalian A1 or to secondary auditory areas (12). In these areas, neural responses show persistent adaptation to the playback of specific sounds, independent of intervening stimuli, a phenomenon referred to as stimulus-specific adaptation (SSA) (13, 14). This form of SSA has not been described in rodents or nonhuman primates to our knowledge (15); it is very long-lasting (hours to days) and is sensitive to the order of elements within each stimulus, as well as to their acoustic structure. It thus appears to reflect recognition of compound auditory objects through a process of statistical learning.

In the present study, we used bilateral multichannel recordings in the caudal forebrain of awake birds in passive-exposure paradigms (based on those used in infant speech studies) to define neural mechanisms of statistical learning in the auditory system. We found that effects of passive exposure could be detected in the firing patterns of neurons in NCM and CLM. Responses to familiar invariant patterns were reduced relative to novel patterns for stimuli with contiguous transitions and for stimuli with nonadjacent elements. This reduction appears to reflect SSA and is associated with increased neural discrimination between familiar stimuli. Furthermore, differential responses to familiar nonadjacent patterns showed a difference between the two hemispheres; this learning effect was seen only in the right hemisphere and may reflect a top-down modulation that is lateralized.

Significance

Human infants and adults can extract statistical regularities from a continuous sound stream even when passively exposed. This type of unreinforced learning is important for language acquisition and auditory perception. We exposed songbirds, a well-developed model for studying perception and production of learned vocal signals, to sound streams and then recorded neural activity from the auditory forebrain. Responses demonstrated neuronal recognition of complex recurring patterns present in the stream, even for patterns made of nonadjacent sounds. These results show that passive exposure can create and differentiate representations of recently heard sounds in the neuronal population. Because auditory information about the world is received serially over time, recognition of complex auditory objects may depend on this type of ongoing memory process.

NEUROSCIENCE

PSYCHOLOGICAL AND COGNITIVE SCIENCES

Author contributions: K.L. and D.S.V. designed research; K.L. performed research; K.L. and D.S.V. analyzed data; and K.L. and D.S.V. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. Email: vicario@rci.rutgers.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1412109111

Results

Statistical Learning of Transition Probability. Four experiments assessed whether the songbird auditory forebrain records transition probabilities in a long sound stream during passive listening in the absence of reinforcement. Methods and terminology were adopted from human infant research (1). In experiment 1, awake adult zebra finches ($n = 5$) first heard eight repetitions of a continuous stream containing six artificial "words" (each composed of six synthetic syllables; Fig. 1*D* and *Methods*) in a shuffled order. Immediately (<5 min) after the conclusion of last repetition of the continuous stream during awake electrophysiological recording, each bird was exposed to 12 different stimuli on individual trials [6-s interstimulus interval (ISI)] that included six "words" from the familiar stream and six novel "nonwords" made from the same syllables arranged in different order. No syllable transitions were shared between familiar words and nonwords. Responses to these test stimuli were obtained from single-unit ($n = 99$) and multiunit spike recordings at 75 sites in NCM (45 single units; 40 sites) and CLM (54 single units; 35 sites). The large majority (63%; 62 of 99) of the single units showed higher responses (spike rate) to nonwords over words. For single unit responses, the difference between nonwords and words was highly significant (Wilcoxon $z = -3.43$; $P < 0.001$; Fig. 1*E*, blue bar). When multiunit data (spikes that crossed a threshold; *Methods*) from each site were analyzed, 79% of sites (59 of 75) showed higher responses (spike rate) to nonwords over words (Fig. 1*F*, blue bar), and the difference between nonwords and words was again highly significant (Wilcoxon $z = -5.87$; $P < 0.001$). Overall, these results show that nonwords elicit larger responses, providing evidence that they are distinguished as novel and consistent with the observation that infants showed longer head-turning time to nonwords than words in a similar exposure paradigm (1). Thus, the bird's auditory system encodes information about the invariant syllable order of each word during passive exposure to the continuous syllable stream for 60 min; this is a form of statistical learning.

Experiment 2 assessed neural responses to additional invariant structure that is present in the long continuous stream of words. In the shuffled word sequence of experiment 1, pairs of different words can follow each other more than once. This effectively exposes the bird to "part-words," which consist of the last three syllables of each word and first three syllables of each other word (Fig. 1*D*). Although the frequencies of most syllable transitions in part-words were the same as those in the two adjoining words, the syllable transition in the middle of each part-word had a 20% probability of occurrence in the word stream. To determine whether the brain could use the relative frequency of syllable pairs to distinguish boundaries between words from syllable transitions within each word (1), we first exposed a group of naïve birds ($n = 5$) to the long word stream, as in the first experiment, then tested with part-words as well as words (as in experiment 1).

The results showed higher responses to part-words than words in 61% of single units (66 of 109; Fig. 1*E*, red bar) and 74% of multiunits (55 of 74; Fig. 1*F*, red bar). These differences were significant in single-unit (Wilcoxon $z = -3.64$; $P < 0.001$) and multiunit data (Wilcoxon $z = -4.68$; $P < 0.001$). This difference is striking because words and part-words differ only in the probability of the transition between the third and fourth syllable in a given stimulus (100% for each word vs. 20% for each part-words). This result demonstrates that songbird neurons are sensitive to the transition probabilities, not just the order of syllables they hear during passive exposure to the sound stream.

In experiment 3, to rule out the possibility that the observed effect was caused by a nonspecific factor, we tested a third group

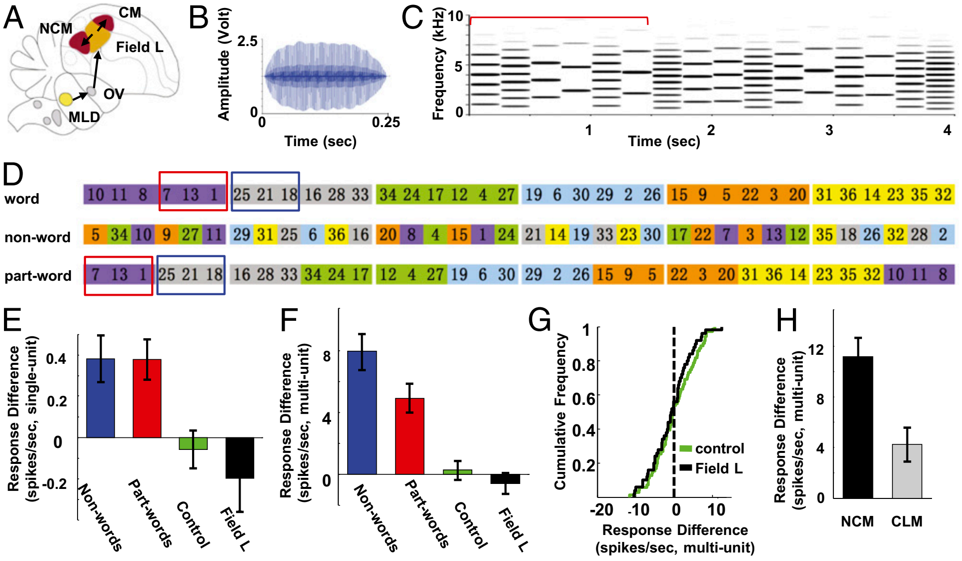

Fig. 1. Experimental design, stimuli, and responses in stimulus stream experiments. (*A*) Ascending auditory pathway in songbirds. Auditory nuclei of avian hindbrain innervate nucleus mesencephalicus lateralis, pars dorsalis (MLd; inferior colliculus homolog). MLd innervates nucleus ovoidalis (OV; medial geniculate homolog). OV projects to forebrain field L2 (analog of A1, layer IV). Field L2 innervates L1 and L3, which in turn project to NCM and the caudal mesopallium (CM). (Modified with permission from ref. 12.) (*B*) Example of a synthesized zebra finch song syllable. (*C*) A segment of the long word stream made up of syllables (like the one in *B*) with different fundamental frequencies. The red bracket indicates the start and the end of one artificial six-syllable word. (*D*) The order of syllables in words, nonwords, and part-words. Syllables within a word are shown in the same color. Each syllable is labeled with a number indicating its order in pitch (low to high). Nonword sequences were made from the same syllables as words, but in a changed order. Part-words consist of the three final syllables of one word (red box) and the first three syllables of the next word (blue box). (*E*) Differences in single-unit responses between nonwords and part-words vs. words. Differences were significant for both nonwords (blue) and part-words (red) in NCM/CLM of birds exposed to the word stream, but not in field L (black) or in control birds (green) without preexposure. (*F*) The overall pattern of effects seen for single-unit data (*E*) was also seen for multiunit recordings; nonwords also showed a significantly larger effect than part-words. (*G*) Cumulative frequency distributions of differences in responses between words and part-words were symmetrically distributed around zero in NCM/CLM of control birds (green) and in field L (black) neurons. (*H*) NCM and CLM showed significant differences for nonwords vs. words in single-unit data, but NCM showed larger differences than CLM only in multiunit data. Error bars show ± SEM.

of naïve birds ($n = 5$) as controls. They were not exposed to the long sound stream, but were tested with the same stimuli as in experiment 2 as described earlier. There was no significant difference in responses between part-words and words in single-unit data (Wilcoxon $z = -0.27$; $P = 0.784$, Fig. 1*E*, green bar) or multiunit data (Wilcoxon $z = -0.38$; $P = 0.705$; Fig. 1*F*, green bar). Differences in responses between words and part-words were symmetrically distributed around zero (Fig. 1*G*, green trace).

In the results presented so far, data acquired from NCM and CLM showed similar patterns and were pooled together for analysis. To determine whether the two areas showed subtle response differences, we also analyzed multiunit and single-unit data from the three experiments with two-way repeated-measures ANOVAs in which NCM vs. CLM and left vs. right hemispheres were treated as two independent factors and responses to words and to nonwords or part-words were dependent measures. We found a significant interaction between NCM/CLM and responses to words and nonwords only in the analysis of multiunit data from the first experiment [words vs. nonwords; $F(1,71) = 10.64$; $P = 0.002$]. Further testing showed that the difference in responses between words and nonwords was significantly higher in NCM than in CLM (two-sample Kolmogorov–Smirnov test, $P < 0.001$; Fig. 1*H*), although both showed significant effects ($P < 0.01$). Although this is consistent with the observation that that NCM represents familiarity more strongly than CLM (16), the difference between NCM and CLM was not confirmed in the analysis of single-unit data from the same experiment or in the other two experiments.

In experiment 4, we explored whether the effects of passive exposure on differentiation between words and part-words seen in NCM and CLM could also be observed in recordings from the major source of their auditory input, the field L complex. This complex includes field L2, which is analogous to the thalamorecipient layer of A1 in mammals and projects indirectly across one or a few synapses in fields L3 and L1 to NCM and CLM, respectively (17). These birds ($n = 8$) were exposed to the same word stream, then tested with words and part-words as in experiment 2. In field L recordings, we found no significant difference in responses between part-words and words in single-unit data (Wilcoxon $z = -1.58$; $P = 0.115$; Fig. 1*E*, black bar) or multiunit data (Wilcoxon $z = -0.81$; $P = 0.420$; Fig. 1*F*, black bar). Differences in responses between words and part-words were symmetrically distributed around zero (Fig. 1*G*, black trace).

In sum, the first series of four experiments showed that, after 60 min of passive exposure to a long continuous sound stream, neurons in NCM and CLM have higher responses to stimuli with lower transition probability than to stimuli with higher transition probability in that sound stream. This effect is absent in a control group (experiment 3) that received no exposure. Furthermore, this effect is not seen in the field L complex, which is the main source of auditory input to NCM and CLM. Therefore, we conclude that areas NCM and CLM of the auditory forebrain may represent the first stage of the process of extracting invariant sound patterns (in this case, first-order conditionals) presented in an ongoing acoustic stream.

Decorrelation of Auditory Responses After Passive Exposure. Our observations show that auditory responses to high-probability familiar patterns were lower than responses to novel patterns. This reduction in responses is very likely a result of SSA, which is documented in these songbird auditory areas. However, it is still unknown whether the lower responses to familiar patterns convey information relevant to stimulus identification and/or discrimination. One possible mechanism is that the reduced response—sparser firing—to familiar sound sequences may reflect improved coding efficiency (18). We speculated that the decreased response amplitude with adaptation would result in a lower correlation of responses between neurons for any given stimulus and thus reduce redundancy. A similar hypothesis has been extensively studied in visual research and provides a good explanation for some visual aftereffects that follow repeated exposure to certain patterns (19).

To explore these possibilities, we reanalyzed single-unit response data for part-words vs. words in experiment 2 (birds with sound exposure) and experiment 3 (naïve animals). Multiunit data were not studied because each site can include responses from neurons with different selectivity. We first compared the amplitudes of responses to all stimuli in the two conditions (Fig. 2*A*). Absolute response amplitudes (i.e., spike rates) in naïve birds were significantly higher than in birds that heard the sound stream (Kolmogorov–Smirnov test, $P < 0.001$), presumably because all stimuli were novel. We then compared the correlations between responses elicited from different pairs of simultaneously recorded neurons as follows. For each stimulus, we calculated poststimulus time histograms (PSTHs, 10-ms bins) from the first 10 trials for all single units in each naïve (example in Fig. 2*B*) and each sound-exposed (Fig. 2*C*) bird, then computed correlation coefficient between PSTHs for each pair of neurons recorded simultaneously from each bird. Correlation coefficients obtained from all neuron pairs were first averaged for each stimulus tested in each bird, and then the grand means for all stimuli in experiments 2 and 3 were compared. In sound-exposed birds, correlation coefficients between PSTHs were significantly lower than in naïve birds (Kolmogorov–Smirnov test, $P < 0.001$; Fig. 2*D*). Therefore, the adaptation that occurs during passive exposure appears to underlie decorrelation in responses of different neurons. This might reflect more efficient coding and ultimately contribute to stimulus discrimination on the population level; however, a more conclusive understanding of the effect of decorrelation at the population level would require analysis of the correlation of response noise across sites (20), which is beyond the scope of this study.

Lateralization and Learning Nonadjacent Patterns. In experiments 1–4, the words made familiar by passive exposure were all fixed sequences of six syllables. However, in the natural environment, ongoing sounds from different sources often overlap in time, and may interfere. Components of another source may overlap a target sound in a variable way. Thus, the invariant pattern of the target may have to be detected from the recurring relationship of nonadjacent components. This process is likely to be essential for recognizing auditory objects and may contribute to stream segregation for episodic sounds. To explore this, in

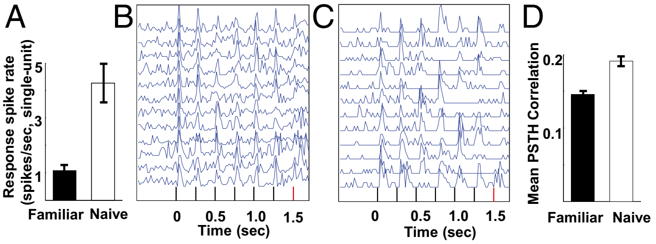

Fig. 2. Comparisons of response amplitudes and correlation coefficients between naïve birds and birds exposed to the stimulus stream. (*A*) Mean response amplitude to all stimuli in the test phase. Response amplitude was significantly lower in birds that had heard the stimulus stream (solid bar) in the part-word experiment than in naïve birds (open bar) in a control experiment. (*B*) Representative PSTHs of responses to 12 stimuli (six words and six part-words) in one neuron recorded from a naïve bird. For most stimuli, PSTHs showed clear peaks in response to each syllable onset (black vertical ticks) and to final offset (red tick). (*C*) Representative PSTHs of responses to the same 12 stimuli as in *A* in a neuron recorded from a bird exposed to the stimulus stream. Response peaks at syllable onsets showed heterogeneous response patterns for different stimuli. (*D*) Mean of correlation coefficients between response PSTH waveforms in birds exposed to the stimulus stream and in naïve birds. Correlation coefficients were significantly lower in birds that had heard the stimulus stream. Error bars show ± SEM.

experiment 5, we made a more stringent test of statistical extraction of invariant acoustic structure by songbird neurons by testing with second-order conditionals (i.e., nonadjacent regularities). Methods and terminology were borrowed from human psychophysics work (2). We constructed a sound stream consisting of syllable triplets in which the first and the last syllables were fixed and the middle syllable was variable (Fig. 3*A*), with forms, e.g., aXd, bXe, cXf (Fig. 3*B* and *Methods*). As in the previous experiments, birds (n = 12) were passively exposed to the shuffled triplet stream, then underwent electrophysiological recording in NCM and CLM. For testing, the stimulus set included 24 familiar triplets from the passive exposure stream and 24 novel triplets. Each novel triplet was matched to a familiar triplet by having the same first and middle syllables or the same middle and last syllables, so that the position of any given syllable and the order between two adjacent syllables were the same, and the only difference was in the fixed relationship between the first and last syllables for familiar triplets.

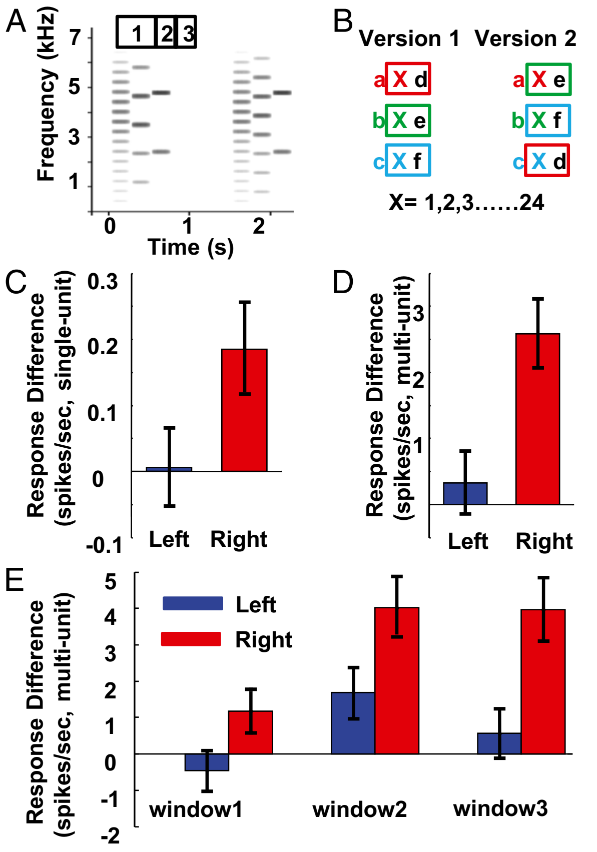

Fig. 3. Nonadjacent stimuli and neural responses in experiment 5. (*A*) Example of triplet stimuli. Triplets are made from syllables like those in Fig. 1 *B* and *C*. The two triplets shown share the same first and last syllable, but the middle syllable is variable. Numbered bars above the first triplet show the timing of three response windows (see text and *E*). (*B*) Structure of the stimulus set. One version of triplets was heard during passive exposure, and both versions were played in the testing phase. The two versions share the same sounds in different combinations. Letters with same color indicate triplets that share the first and second syllable. Boxes with same color indicate triplets that share the second and third syllable. (*C*) Differences in single-unit responses between novel and familiar triplets were significant only in the right hemisphere. (*D*) Differences in multiunit responses between novel and familiar triplets in each hemisphere were only significant in the right hemisphere. (*E*) Differences in responses between novel and familiar triplets in the three response windows. No significant difference was found in window 1 for either hemisphere. Significant differences were found in window 2 for both hemispheres. In window 3, significant differences were seen only in the right hemisphere. Error bars show ± SEM.

Responses to triplet stimuli were obtained from 222 single units at 172 recording sites. Slightly more than half (54%; 120 of 222) of these single units showed higher spike rates to novel triplets than familiar triplets, and the difference in responses showed a nonsignificant trend (Wilcoxon $z = -1.90$; $P = 0.057$). When recordings from each site were analyzed as multiunit data, 63% of sites (107 of 172) showed higher response to novel triplets than familiar triplets, and the difference was significant (Wilcoxon $z = -3.36$; $P < 0.001$). When single and multiunit data were analyzed in a two-way ANOVA in which NCM vs. CLM and left vs. right hemispheres were treated as two factors, with responses to familiar and novel triplets as a repeated measure, a significant interaction was found between hemisphere and familiar vs. novel triplets [$F(1,168) = 10.41$; $P < 0.002$]. Further tests that compared responses to familiar and novel triplets for each single- and multiunit recording showed no difference in the left hemisphere [single units (Fig. 3*C*, blue bar), $n = 121$, Wilcoxon $z = -0.17$, $P = 0.869$; multiunit (Fig. 3*D*, blue bar), $n = 93$, Wilcoxon $z = -0.58$, $P = 0.562$]. In contrast, more than 61% (62 of 101) of single units and 71% (56 of 79) of multiunits in the right hemisphere showed stronger responses to novel triplets (single units, Fig. 3*C*, red bars; multiunits, Fig. 3*D*, red bars). The right hemisphere had significantly higher responses to novel than familiar triplets (single units, Wilcoxon $z = -2.57$, $P = 0.010$; multiunits, Wilcoxon $z = -4.29$, $P < 0.001$). These data show that neurons in NCM and CLM can record invariant patterns of nonadjacent sounds during passive exposure. Intriguingly, this process is seen in the right but not the left hemisphere. Similar results were seen in both regions studied; there was no significant interaction between NCM/CLM and the main effect [single-unit, $F(1,218) = 0.41$, $P = 0.523$; multiunit, $F(1,168) = 0.43$, $P = 0.512$] or interaction between NCM/CLM and the lateralization effect [single-unit, $F(1,218) = 0.95$, $P = 0.331$; multiunit, $F(1,168) = 0.01$, $P = 0.927$].

Temporal Characteristics of the Response to Novel Triplets. In a further analysis, we explored the time course of the response to novel nonadjacent patterns by using multiunit data to measure response patterns at the population level. We predicted that the difference in neural responses to novel and familiar triplets would occur at the onset of the last syllable of each triplet, because it is only then that the difference occurs and can be detected. We tested the difference in responses in each hemisphere during three time windows: (*i*) the first two syllables, (*ii*) the last syllable, and (*iii*) the 250-ms period starting at the offset of the last syllable (Fig. 3*A*, numbered boxes above sonogram). In window 1, no significant differences between novel and familiar triplets were observed in either hemisphere, (left, Wilcoxon $z = -0.75$, $P = 0.454$; right, Wilcoxon $z = -1.77$, $P = 0.077$). In window 2, left and right hemispheres showed significant differences between novel and familiar triplets (left, Wilcoxon $z = -2.45$, $P = 0.014$; right, Wilcoxon $z = -4.42$, $P < 0.001$). Thus, this time-window analysis revealed an effect in the left hemisphere not seen in the data summed across the entire response. In window 3, only the right hemisphere showed significant differences between novel and familiar triplets (left, Wilcoxon $z = -0.64$, $P = 0.521$; right, Wilcoxon $z = -3.72$, $P < 0.001$). In addition, when differences between novel and familiar triplets were compared between the two hemispheres in each of the three windows, only window 3 showed a significant difference (two-sample Kolmogorov–Smirnov test, $P = 0.046$), indicating that most of the difference between hemispheres occurs in this late period. Therefore, a modulation after the offset of the last syllable contributes to the hemispheric difference. This late timing might reflect top-down modulation of auditory responses or a memory retrieval process that affects the two hemispheres differently.

Discussion

The present experiments demonstrate that the auditory forebrain of songbirds extracts and records first-order transition probabilities from a continuous stream of sounds without cues or reinforcement. This is seen as a higher response amplitude to novel over familiar patterns and is absent in recordings from field L, which provides the major input to NCM and CLM. Thus, NCM and CLM may be the first stage where neural extraction of invariant patterns occurs, without any necessary role for reinforcement. In addition, the temporal structure of neural responses to stimuli made familiar by passive exposure became less similar between neurons, suggesting enhanced coding efficiency. We also found that second-order conditional (nonadjacent) patterns were extracted and recorded from streams of triplet stimuli; this occurred primarily in the right hemisphere and included late response components that suggest a top-down influence on auditory responses.

In recent years, neurons in the auditory forebrain have been recognized as much more than fixed filters extracting simple acoustic features. Accumulating evidence indicates that responses in the auditory forebrain encode the behavioral salience (21, 22) and probability of sounds (23). Our results provide evidence that auditory responses are further modulated by statistical regularities experienced through passive sound exposure. Recurring patterns in the acoustic mixture (whether they occur in contiguous sequences or are separated by other variable sounds) are extracted and recognized, as evidenced by the reduced size of auditory responses to specific familiar compared with novel patterns. This rapid, experience-dependent plasticity in auditory responses may have two important implications: First, the reduced responses may effectively lower the probability that stimuli with familiar regularities will engage attention, and thus result in reduced behavioral responsiveness to these patterns, as has been observed in human infants, nonhuman primates, and birds (1, 4, 7). Concomitantly, attentional resources may remain available when novel, variant, and potentially meaningful acoustic events are detected in the acoustic mixture. This combination of mechanisms may contribute to perception in complex auditory scenes (24) and is consistent with the hypothesis that sound memories may underlie templates that contribute to a schema-based auditory scene analysis (25). A recent study in human psychophysics showed that repetitive exposure enhanced recovery of auditory objects in overlapping mixtures (26), consistent with the hypothesis outlined here. Second, beyond a possible attentional effect, response reduction—sparsening of spike patterns that may reflect increased coding efficiency—may in turn function to increase discriminability of familiar patterns. Indeed, our results show that reduced responses were associated with decorrelation of the temporal structure of responses between different neurons, potentially increasing stimulus separation at the population level.

Our results also imply a possible mechanism that integrates temporally independent sounds into a perceptual object. In our design, familiar and novel stimuli share the same acoustic features during exposure; all that differs is the sequential relationship (transitional or nonadjacent) between distinct artificial syllables over a temporal scale of hundreds of milliseconds. Thus, discrimination of familiar from novel stimuli in the testing phase implies that a "memory" of sequential relationships between different sounds is formed during the passive exposure phase. We propose that this memory for sound patterns may work as a template that contributes to integrating multiple sounds into a perceptual auditory object. Furthermore, our results suggest two mechanisms that may contribute to this integration. The first-order contiguous invariant pattern may be processed locally in NCM and CLM, with SSA as the most likely underlying mechanism. Our data show no hemispheric difference for this type of processing. In contrast, the two hemispheres appear to play different roles in the extraction of second-order nonadjacent patterns. The response to the final syllable of the triplets (which distinguishes the novel stimulus) showed two significant components. The first, which may underlie the initial differentiation between novel and familiar patterns, was present in both hemispheres (reflecting SSA), but was stronger in the right. The second component, after the offset of the final syllable, was larger and only present in the right hemisphere. This later component may represent a further stage of the pattern extraction process that reflects the influence of top-down input from brain regions beyond the caudal auditory areas studied. Structures in the songbird forebrain that could be involved in the top-down process include (*i*) vocal motor structures, e.g., HVC and the lateral magnocellular nucleus of the anterior nidopallium, that contribute to vocal learning (27–29) and perception of grammatical structures (2, 7) and possibly (*ii*) the hippocampus, involved in statistical learning in vision (30), but not yet explored in songbirds. In any case, the recognition of nonadjacent patterns occurs primarily in the right hemisphere, which is known to have higher responses to conspecific vocalizations and adapt more rapidly than the left hemisphere (31).

The exact relationship between our results and observations using EEG and fMRI methods in humans hearing artificial words embedded in a continuous stream is unclear. Observations from EEG (10) and fMRI (11) works suggest that some brain areas beyond the auditory system may be involved in auditory object recognition. Our timing data suggest the existence of similar influences in the songbird auditory forebrain, so the paradigm we have developed may provide a useful model to explore these modulatory effects at the neurophysiological level.

Taken together, our results demonstrate how processes of statistical learning that detect and store invariant structures in the ongoing auditory stream could contribute to assembling auditory sequences into compound auditory objects. SSA to recurring patterns appears to play a role in a continuous process that creates and differentiates representations of recently heard sounds across the population of neurons we recorded. Because auditory information about the world must be received serially over time, recognition of complex auditory objects may depend on this type of mnemonic process.

Methods

Subjects. Experiments used adult male zebra finches (N = 35) bred in our aviary or obtained from the Rockefeller University Field Research Center. Birds were housed on a 12:12-h light-dark cycle in a general aviary, where they could see and hear other birds. Food and water were provided ad libitum. All procedures conformed to a protocol approved by the institutional animal care and use committee of Rutgers University.

Stimuli. In experiments 1, 2, and 4, birds were preexposed to a continuous sound stream that contained 50 repeats of six different artificial words in a shuffled order. Each word consisted of six synthetic syllables in a fixed sequence without gaps, chosen randomly from a larger set of 36 syllables that differed only in fundamental frequency. No gaps or other acoustic cues indicated the start or end of each word. Two such streams containing different words were used in different subjects in a balanced design. Each syllable was a harmonic stack (250-ms duration) with a fundamental frequency that ranged from 400 to 2,384 Hz in the first set and from 416 to 2,430 Hz in the second set, with equal logarithmic increments in the two sets. The spectral emphasis and amplitude envelope of each syllable were derived from an average of natural zebra finch calls (24) to optimize responsiveness. Syllables were equated for rms amplitude, and stimuli were played back in a free field (60 dB sound pressure level at the bird's ears). In experiment 5, birds were preexposed to a triplet stream that contained six repeats of 72 different triplets (Fig. 3*A*) in a shuffled order. Triplet stimuli were assembled from the same artificial syllables described earlier, with no gap between syllables. Although the first and last syllables had fixed pitches (a, 400 Hz; b, 952 Hz; c, 1,801 Hz; d, 646 Hz; e, 1,331 Hz; f, 2,384 Hz), the middle syllables were 24 syllables randomly chosen from the syllable pool. There was a 750-ms gap between triplets in the stream. Two versions of the triplets (Fig. 3*B*) and corresponding triplet streams were constructed from the same 30 syllables. In the first, triplets took the forms aXd, bXe, and cXf; in the second, triplets took the forms aXe, bXf, and cXd. Animals were initially exposed to one of the triplet streams and then were tested with triplets from both versions. Triplets were balanced across animals.

Surgery. In preparation for electrophysiological recording, each animal was anesthetized with isoflurane (2% in oxygen) and placed in a stereotaxic apparatus. Marcaine (0.04 mL, 0.25%) was injected under the scalp, the skin was incised, and a craniotomy in the first layer of the skull exposed the bifurcation of the midsagittal sinus. Then, four smaller openings were made in the inner bone layer over NCM and CLM in both hemispheres. Dental cement was used to attach a metal post to the skull rostral to the opening and to form a chamber around the recording area. The chamber was then sealed with silicone elastomer (Kwiksil; World Precision Instruments). Metacam (0.04 mL, 5 mg/mL, i.m.) was given to relieve postsurgical pain. Anesthesia was discontinued and the bird allowed was allowed to recover under a heat lamp.

Passive Exposure Procedures. Two days after initial surgery (to allow for full recovery from anesthesia), each bird was isolated in a cage in a walk-in soundproof booth (IAC). In experiments 1, 2, and 4, birds were exposed to the continuous stream of artificial words in a shuffled order; the stream lasted 7.5 min and was repeated seven times with 10-min gaps. The bird was then restrained in a comfortable tube, the implanted head pin was fixed to the stereotaxic frame, and the bird was prepared for electrophysiological recording. When all microelectrodes were in place (<1 h) the stream was played one more time, immediately followed by testing with the appropriate auditory stimuli. In experiment 3, birds received no exposure. In experiment 5, as described earlier, unrestrained birds were first exposed to the triplet stream (Fig. 3*A*), which lasted 10.8 min and was repeated three times with 10-min gaps. Then the bird was restrained and heard the triplet stream one more time at the start of electrophysiological testing.

Electrophysiology and Stimulus Presentation. Recordings were made at 16 sites (n = 4 each in NCM and CLM in both hemispheres) using glass-insulated platinum/tungsten microelectrodes (2–3 MΩ impedance) independently advanced by a multielectrode microdrive (Eckhorn; Thomas Recording). Signals were amplified (×19,000) and filtered (bandpass 0.5–5 kHz), then acquired at 25 kHz per channel using Spike 2 software [Cambridge Electronic Design (CED)]. White noise stimuli with the amplitude envelope of canary song were presented to search for responsive sites. When all electrodes showed responses, the bird heard a final playback of the long word stream or the triplet series, followed by the testing stimuli. In experiments 1–4, each stimulus word was presented individually at 6 s ISI during testing. The stimulus set included six novel nonwords and six words or part-words (*Results*) from the familiar stream. Words and nonwords were repeated 20 times each in shuffled order. In experiment 5, each testing stimulus set included 24 triplets from the familiar stream and 24 novel triplets. Each stimulus was repeated 20 times in shuffled order at 6 s ISI. At the end of the session, eight small electrolytic lesions (20 μA for 15s) were made to enable histological reconstruction of recording sites.

Histology. At the conclusion of the experiment, the animal was killed with an overdose of pentobarbital, then perfused with saline solution and paraformaldehyde. Sagittal sections were cut at 50 μm on a Vibratome, then stained with cresyl violet. Lesion sites were confirmed histologically to be in NCM and CLM by using cytoarchitectonic landmarks.

Analysis of Neural Responses. The spike waveforms of single units were detected by using template-based digital clustering algorithms implemented in Spike2 software (CED). Single-units were validated by analysis of the interspike interval histograms. To be accepted, a unit had to have contamination rate (<2% of interspike intervals under 2 ms, corresponding to spike rates > 500 Hz). The same recordings were also analyzed as multiunit spiking activity by counting all spike waveforms that crossed a threshold (2 SDs above baseline). The spikes of a single unit typically represent only a small percentage of all multiunit spikes at each recording site, so we report single-unit and multiunit data. Spike rates were calculated in spikes per second over a control window (500 ms before stimulus onset) and over a response window (from stimulus onset to offset plus 250 ms) on each trial. Response amplitude was quantified as the difference between the spike rate in the response and control windows. For each stimulus, response rates were then averaged across the first seven trials for experiments 1–4 and the first five trials for experiment 5. For statistical analysis, the mean responses to all familiar and novel stimuli were calculated for each single- or multiunit channel, and compared as repeated measures by using the Wilcoxon matched-pairs test, with $P < 0.05$ as the criterion for significance. Two-sample Kolmogorov–Smirnov tests were used to compare responses across different brain regions.

ACKNOWLEDGMENTS. We thank Tom Ziv and Manda Pierce for technical assistance and Nikolas Francis, Dana Strait, and Mimi Phan for helpful comments on the manuscript. This work was supported by National Institutes of Health Grant DC008854 (to D.S.V.).

1. Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274(5294):1926–1928.
2. Gómez RL (2002) Variability and detection of invariant structure. *Psychol Sci* 13(5):431–436.
3. Newport EL, Aslin RN (2004) Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognit Psychol* 48(2):127–162.
4. Hauser MD, Newport EL, Aslin RN (2001) Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition* 78(3):B53–B64.
5. Newport EL, Hauser MD, Spaepen G, Aslin RN (2004) Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognit Psychol* 49(2):85–117.
6. Toro JM, Trobalón JB (2005) Statistical computations over a speech stream in a rodent. *Percept Psychophys* 67(5):867–875.
7. Abe K, Watanabe D (2011) Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neurosci* 14(8):1067–74.
8. Saffran JR, Johnson EK, Aslin RN, Newport EL (1999) Statistical learning of tone sequences by human infants and adults. *Cognition* 70(1):27–52.
9. Gebhart AL, Newport EL, Aslin RN (2009) Statistical learning of adjacent and non-adjacent dependencies among nonlinguistic sounds. *Psychon Bull Rev* 16(3):486–490.
10. Cunillera T, et al. (2009) Time course and functional neuroanatomy of speech segmentation in adults. *Neuroimage* 48(3):541–553.
11. McNealy K, Mazziotta JC, Dapretto M (2006) Cracking the language code: Neural mechanisms underlying speech parsing. *J Neurosci* 26(29):7629–7639.
12. Theunissen FE, Shaevitz SS (2006) Auditory processing of vocal sounds in birds. *Curr Opin Neurobiol* 16(4):400–407.
13. Chew SJ, Mello C, Nottebohm F, Jarvis E, Vicario DS (1995) Decrements in auditory responses to a repeated conspecific song are long-lasting and require two periods of protein synthesis in the songbird forebrain. *Proc Natl Acad Sci USA* 92(8):3406–3410.
14. Chew SJ, Vicario DS, Nottebohm F (1996) A large-capacity memory system that recognizes the calls and songs of individual birds. *Proc Natl Acad Sci USA* 93(5):1950–1955.
15. Scott BH, Mishkin M, Yin P (2012) Monkeys have a limited form of short-term memory in audition. *Proc Natl Acad Sci USA* 109(30):12237–12241.
16. Thompson JV, Gentner TQ (2010) Song recognition learning and stimulus-specific weakening of neural responses in the avian auditory forebrain. *J Neurophysiol* 103(4): 1785–1797.
17. Vates GE, Broome BM, Mello CV, Nottebohm F (1996) Auditory pathways of caudal telencephalon and their relation to the song system of adult male zebra finches. *J Comp Neurol* 366(4):613–642.
18. Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287(5456):1273–1276.
19. Barlow HB (1997) The knowledge used in vision and where it comes from. *Philos Trans R Soc Lond B Biol Sci* 352(1358):1141–1147.
20. Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. *Nat Rev Neurosci* 7(5):358–366.
21. Weinberger NM (2007) Auditory associative memory and representational plasticity in the primary auditory cortex. *Hear Res* 229(1-2):54–68.
22. Fritz JB, Elhilali M, David SV, Shamma SA (2007) Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? *Hear Res* 229(1-2): 186–203.
23. Ulanovsky N, Las L, Nelken I (2003) Processing of low-probability sounds by cortical neurons. *Nat Neurosci* 6(4):391–398.
24. Lu K, Vicario DS (2011) Toward a neurobiology of auditory object perception: What can we learn from the songbird forebrain? *Current Zoology* 57(6):671–683.
25. Bregman AS (1990) *Auditory Scene Analysis* (MIT Press, Cambridge, MA).
26. McDermott JH, Wrobleski D, Oxenham AJ (2011) Recovering sound sources from embedded repetition. *Proc Natl Acad Sci USA* 108(3):1188–1193.
27. Brenowitz EA (1991) Altered perception of species-specific song by female birds after lesions of a forebrain nucleus. *Science* 251(4991):303–305.
28. Vicario DS, Raksin JN, Naqvi NH, Thande N, Simpson HB (2002) The relationship between perception and production in songbird vocal imitation: What learned calls can teach us. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 188(11-12):897–908.
29. Raksin JN, Glaze CM, Smith S, Schmidt MF (2012) Linear and nonlinear auditory response properties of interneurons in a high-order avian vocal motor nucleus during wakefulness. *J Neurophysiol* 107(8):2185–2201.
30. Turk-Browne NB, Scholl BJ, Johnson MK, Chun MM (2010) Implicit perceptual anticipation triggered by statistical learning. *J Neurosci* 30(33):11177–11187.
31. Phan ML, Vicario DS (2010) Hemispheric differences in processing of vocalizations depend on early experience. *Proc Natl Acad Sci USA* 107(5):2301–2306.