Temporal scales of auditory objects underlying birdsong vocal recognition

Timothy Q. Gentner[a)]
Department of Psychology, Neurosciences Graduate Program, University of California, San Diego, La Jolla, California 92093

(Received 20 November 2007; revised 5 May 2008; accepted 28 May 2008)

Vocal recognition is common among songbirds, and provides an excellent model system to study the perceptual and neurobiological mechanisms for processing natural vocal communication signals. Male European starlings, a species of songbird, learn to recognize the songs of multiple conspecific males by attending to stereotyped acoustic patterns, and these learned patterns elicit selective neuronal responses in auditory forebrain neurons. The present study investigates the perceptual grouping of spectrotemporal acoustic patterns in starling song at multiple temporal scales. The results show that permutations in sequencing of submotif acoustic features have significant effects on song recognition, and that these effects are specific to songs that comprise learned motifs. The observations suggest that (1) motifs form auditory objects embedded in a hierarchy of acoustic patterns, (2) that object-based song perception emerges without explicit reinforcement, and (3) that multiple temporal scales within the acoustic pattern hierarchy convey information about the individual identity of the singer. The authors discuss the results in the context of auditory object formation and talker recognition. © *2008 Acoustical Society of America.*

[DOI: 10.1121/1.2945705]

PACS number(s): 43.80.Lb, 43.66.Gf [JAS] Pages: 1350–1359

I. INTRODUCTION

Male European starlings, *Sturnus vulgaris*, produce behaviorally relevant vocalizations (songs) that are spectrally and temporally complex. Both male and female adult starlings can learn to recognize large sets of these songs, even when sung by several different conspecific males (Gentner *et al.*, 2000). This recognition learning, in turn, drives neuronal plasticity in regions of the auditory forebrain analogous to mammalian auditory cortex, where neurons respond most strongly to the songs that birds have learned to recognize (Gentner and Margoliash, 2003). Several lines of evidence point to the importance of short, stereotyped, spectrotemporal patterns called motifs, in guiding song recognition. For example, one can closely control behavioral and physiological responses to songs by manipulation of song acoustics at the motif level, which equates to timescales on the order of several 100's of milliseconds (Gentner, 2004). It is not understood, however, how spectrotemporal acoustic structures defined on more precise timescales contribute to song recognition. One hypothesis is that song recognition is driven explicitly by submotif level features, rather than whole motifs. Alternatively, information about the identity of the singer may be coded at multiple temporal scales within songs. As an initial step in understanding the detailed relationships between vocal recognition and song acoustics we report here on how the short timescale acoustic structure of songs governs recognition behavior.

Starling song is hierarchically structured. Males tend to sing in long continuous episodes called *bouts*. Song bouts are composed of much shorter acoustic units referred to as *motifs* (Adret-Hausberger and Jenkins, 1988; Eens *et al.*, 1991) (Fig. 1) that, in turn, are composed of still shorter units called *notes*. Notes can be broadly classified by the presence of continuous energy in their spectrotemporal representations. The note pattern within a given motif is largely stereotyped across successive motif renditions, and each motif is often repeated two or more times in a song before a different motif is sung. Thus, starling songs appear (acoustically) as sequences of iterated motifs, where each motif is a spectrotemporally complex event (Fig. 1). Different song bouts from the same male are not necessarily composed of the same set of motifs. A complete repertoire of motifs can, however, be characterized over many song bouts, and for a mature male starling can exceed 50 or more unique motifs (Eens *et al.*, 1989; Eens, 1992; 1997; Gentner and Hulse, 1998).

Although some sharing of motifs does occur among captive males (Hausberger and Cousillas, 1995; Hausberger, 1997), the motif repertoires of different males living in the wild are generally unique (Adret-Hausberger and Jenkins, 1988; Eens *et al.*, 1989, 1991; Chaiken *et al.*, 1993; Gentner and Hulse, 1998). Thus, learning which males sing which motifs can provide a diagnostic cue for individual recognition. At least to a first approximation, this strategy does a good job of describing how starlings learn to recognize conspecific songs. Using operant trainings techniques, starlings can easily learn to recognize many songs sung by different individuals, and can maintain this accurate recognition when classifying novel song bouts from the training singers (Gentner and Hulse, 1998; Gentner *et al.*, 2000). When the novel song bouts have *no* motifs in common with the training songs, however, performance in this recognition task falls to chance (Gentner *et al.*, 2000). Also consistent with a motif-

[a)]Electronic mail: tgentner@ucsd.edu

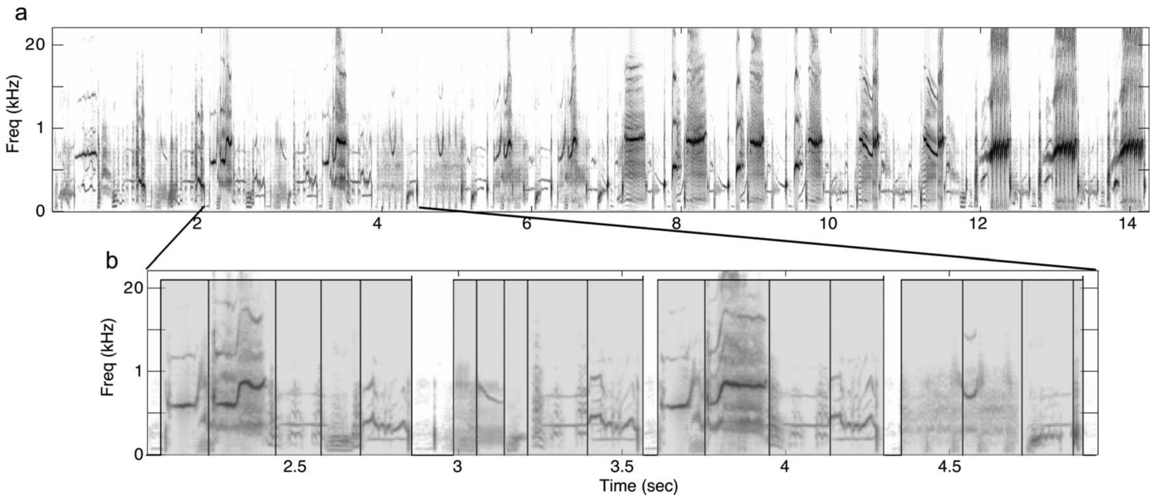

FIG. 1. Hierarchical organization of starling song. Spectrograms showing spectral power as a function of time for (a) a 14-s long excerpt from a much longer bout of singing by one male European starling and (b) a region from the same song with the temporal axis expanded. Motifs in (b) are denoted by the gray shading, and the boundaries of submotif features (see Sec. II) are shown by the overlaid black boxes. Note the highly variable but still repetitive structure, and the hierarchy of groupings for spectrotemporal acoustic patterns.

memorization strategy, the recognition of hybrid, chimeric, songs based on the natural songs of two familiar males shows a linear relationship to the relative proportions of familiar motifs in each bout (Gentner and Hulse, 2000).

Taken together, the results of these studies suggest that when starlings learn to recognize conspecific songs from different singers, they memorize large numbers of unique motifs corresponding to individual singers. The spectrotemporal structure and scale of motifs can vary widely, however, even within a single bird, and the apparent recognition of "whole motifs" may result from recognition of spectrally and/or temporally restricted acoustic features that are diagnostic of (or uniquely covariant with) the larger event (motif). That is, starlings might learn notes rather than motifs. Although motifs form an operationally and phenomenologically useful unit for manipulations that alter song recognition, the necessity of the motif as the minimal perceptual unit for song recognition has not been established empirically. Here we explore the temporal lower bound on acoustic pattern recognition in starlings by testing the recognition of conspecific songs that have been systematically manipulated at submotif temporal scales.

II. GENERAL METHODS

A. Subjects

Seven adult European starlings, *Sturnus vulgaris*, served as subjects in this study. All subjects were wild and caught in southern California in May 2006. All had full adult plumage at the time of capture, and thus were at least one year old. From the time of capture until their use in this study, all subjects were housed in large mixed sex, conspecific aviaries with *ad libitum* access to food and water. The photoperiod in the aviary and the testing chambers followed the seasonal variation in local sunrise and sunset times. No significant sex differences have been observed in previous studies of individual vocal recognition (Gentner *et al.*, 2000), and the sex of subjects in this study was not controlled.

B. Apparatus

Starlings learned to classify the training stimuli using a custom-built operant apparatus, housed in a 61 × 81 × 56 cm inner diameter sound attenuation chamber (Acoustic Systems). Inside the chamber, a subject was held in a weld-wire cage (41 × 41 × 35 cm) that permitted access to a 30 × 30 cm operant panel mounted on one wall. The operant panel contained three circular response ports spaced 6 cm center-to-center, aligned in a row with the center of each port roughly 14 cm off the floor of the cage and with the whole row centered on the width of the panel. Each response port was a PVC housed opening in the panel fitted with an IR receiver and transmitter that detected when the bird broke the plane of the response port with its beak. This "poke-hole" design allows starlings to probe the apparatus with their beak, in a manner akin to their natural appetitive foraging behavior. Independently controlled light emitting diodes (LEDs) could illuminate each response port from the rear. Directly below the center port, in the section of the cage floor immediately adjacent to the panel, a fourth PVC lined opening provided access to food. A remotely controlled hopper, positioned behind the panel, moved the food into and out of the subject's reach beneath the opening. Acoustic stimuli were delivered through a small full-range audio speaker mounted roughly 30 cm behind the panel and out of the subject's view. The sound-pressure level (SPL) inside all chambers was calibrated to the same standard broadband signal. Custom software monitored the subject's responses, and controlled the LEDs, food hoppers, chamber light, and auditory stimulus presentation according to procedural contingencies.

C. Stimuli

1. Song recording

Recordings of four male European starlings were used to generate all the stimuli for this experiment. The procedures for obtaining high-quality song recordings from male starlings have been detailed elsewhere (Gentner and Hulse, 1998). Briefly, a minimum of 0.5 h of song was recorded from each male when housed individually in a large sound-attenuating chamber. During recording, males had visual and auditory access to a female starling (the same female was used to induce song from all the males). All the songs were recorded on digital audiotape (16 bit, 44.1 kHz) using the same microphone (Sennheiser ME66-K6), and high-pass filtered at 250 Hz to remove low frequency background noise. The multiple songs of each bird were parsed into roughly 15-s exemplars of continuous singing taken from the beginning, middle, or end of a typically much longer song bout, and then sorted into sets based on the presence or absence of motifs that were shared with other 15-s song exemplars from the same bird. Human observers labeled the motifs in each 15-s exemplar. These same stimuli have been used to explore the role of motif familiarity in the recognition of individual songs in several studies (Gentner and Hulse, 1998; 2000; Gentner *et al.*, 2000). None of the males whose songs were used to generate the stimuli for the present study served as subjects in the operant testing described here.

2. Baseline training stimuli

To avoid issues related to psuedoreplication (Kroodsma, 1989), we used three different stimulus sets for the baseline song classification. Each stimulus set consisted of eight song exemplars drawn from the library of song bouts sampled from a single bird, and eight exemplars drawn from the songs of another bird (16 exemplars total). The singer of each set of songs and the assignment of those songs as either S+ or S− was counterbalanced across test subjects. Each exemplar was 15 ± 0.5 s of continuous song taken from either the beginning, middle, or end of a song bout, as described previously. Many of the exemplars sampled from the beginning of a song bout included whistles, along with other "warble" motifs [i.e., "variable" motifs, rattles, and high-frequency motifs, for motif nomenclature see Adret-Hausberger and Jenkins (1988) and Eens *et al.* (1991)]. Those sampled at later time points in a bout comprised only warble song motifs. Previous data indicate that recognition is easily learned with this length of a song exemplar, and is unaffected by the relative position within a longer song bout from which the exemplar is sampled and/or the broader motif classes it may or may not contain (Gentner and Hulse, 1998).

3. Submotif permutations

From each of the four original sets of 15-s song exemplars we selected 16 additional 15-s song stimuli that served as the basis for further testing. Eight of these 16 song stimuli had motifs in common with the baseline training exemplars from the same singer and eight were composed entirely of novel motifs sung by the same male (i.e. motifs that did not appear in any baseline training songs). We refer to these two types of song stimuli as "familiar-motif" and "unfamiliar-motif" song, respectively. The familiar-motif songs shared, on average, 89.1% of their motifs with the baseline training songs.

We parsed the 8 familiar-motif and 8 unfamiliar-motif songs from each singer into constituent submotif segments with a semiautomated procedure, detailed as follows. We converted the waveform of each 15-s song stimulus to the frequency domain, by computing the spectrogram of the original stimulus (9.27 ms window, 90% overlap; 512 fft points; see Fig. 2). We obtained an estimate of the noise over a single time slice in a set of songs by manually identifying periods of silence in the songs from each individual, and then computed the mean magnitude of the power spectra (512 fft points) for the "silence" intervals separately for each bird. We then recursively compared each time slice of the stimulus to the silence interval (by taking the normalized inner product) to find points where the song changed from "signal" to silence. On each recursive call, the similarity threshold used to classify the given time slice as signal or "noise" was progressively lowered so that we detected increasingly subtle temporal breaks within the segments yielded by the preceding call. We halted the recursion when we reached a minimum threshold value (held constant for all songs), and marked the start and stop times for all segments parsed in this way. In principle, of course, the recursive parsing procedure could be carried on until arbitrarily small segments were delineated (up to the temporal resolution of the spectrogram).

Using dynamic time warping (DTW) (see Anderson *et al.*, 1996; Kogan and Margoliash, 1998) we compared each parsed song segment to a library of segment templates for that bird, and determined the template (or set of tem-

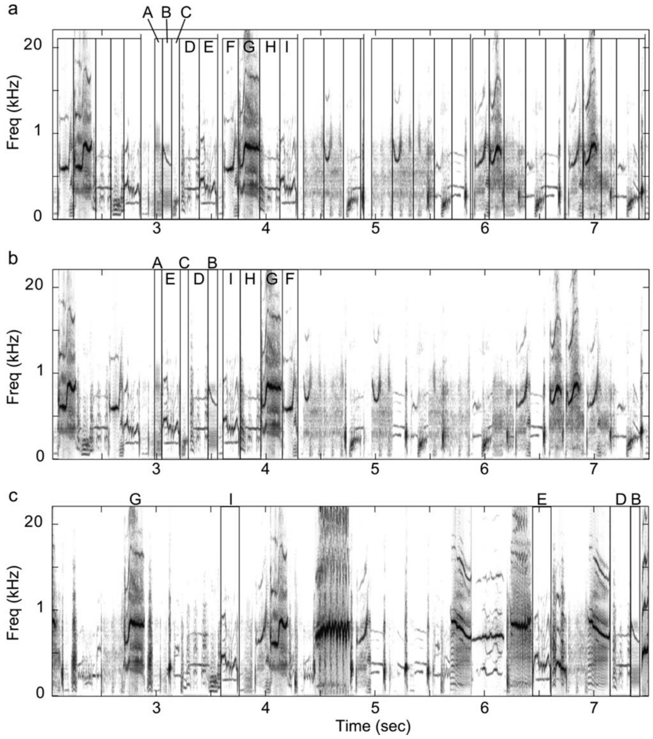

FIG. 2. Submotif song features and permutations. Example spectrograms from excerpts of three test exemplars where ordering of the submotif features, denoted by the black overlaid boxes, is (a) unpermuted (i.e., naturally ordered), (b) permuted within the temporal boundaries of their original motif, or (c) permuted over the entire exemplar. The submotif features for the second and third motifs in (a) are labeled with letters, so that their permuted positions in (b) and (c) can be more easily seen. Because only an excerpt of the entire stimulus exemplar is shown, not all of the submotif features labeled in (a) appear in (c).

plates) that yielded the closest match. From this match we then modified (and confirmed by hand) the segment start and stop times in the parsed songs as necessary to ensure that different iterations of the same motif were parsed into similar sets of segments. The reference library of segment templates for each bird was obtained from a random sample of all the motifs that that bird produced, and optimized so that it produced a minimum error when classifying (using similar DTW techniques) all of the motifs in that bird's repertoire (data not published).

To permute the temporal sequence of each song, we shuffled the order of submotif segments either (1) within the temporal boundaries of their original motif [Fig. 2(b)], or (2) over the whole song [Fig. 2(c)]. In the former case, we constrained the permutation so that the original ordering of segments within a motif was not replicated. The segmentation, dynamic time warping, and permutation routines were written in MATLAB (v7.4). The threshold level for noise/signal and the resulting minimum segment duration were established empirically as part of a larger (unpublished) endeavor using DTW to optimally match a minimum number of spectral-temporal templates to all the submotif components in the songs from a single male starling. We extracted a total of 5141 submotif segments from all of the songs. The mean (±sem) number of motif and submotif segments extracted from a given 15-s song stimulus was 15.61 ± 0.39 and 80.33 ± 1.94, respectively. The mean (±sem) submotif segment duration was 154.66 ± 1.24 ms with a range from 10 to 663 ms. The mean (±sem) motif duration over all song stimuli was 873.12 ± 27.74 ms, with a range from 191 to 2762 ms.

For each of the 16 test songs from a given singer we created four permutations with submotif segments randomly distributed over the whole song and four permutations with submotif segments reordered within their original motif boundaries. This yielded a total of 256 permutation sequences per baseline stimulus set. Because many more than four permutations are possible, we selected a subset that coved the possible range of permutations as uniformly as possible assuming all possible positions for a given submotif segment are equally salient (which is a strong but valid *a priori* assumption given the available data). To quantify the complexity of any given permutation we took the number of submotif segment transitions that differed from the original sequence, normalized by the number of possible transitions. Suppose S is the original stimulus and T is the permuted stimulus and that $S(i)$ represents ith feature in S, then the "R distance" is calculated as the number of times $S(i+1)$ does not follow $S(i)$ in T. For example, when $S=[1\ 2\ 3\ 4\ 5]$ and $T=[4\ 5\ 1\ 2\ 3]$, the R distance for $T=1/(5-1)=0.25$, as only one transition (from 3 to 4) is absent from T. If $T=[3\ 1\ 5\ 2\ 4]$, the R distance is $4/(5-1)=1$ as all 4 original transitions are absent.

D. Procedure

1. Shaping

Subjects learned to work the apparatus through a series of successive shaping procedures. Upon initially entering the operant chamber, the subject was given unrestricted access to the food hopper, and then taught through autoshaping to peck the center port to gain access to the food. Once the subject pecked reliably at the center port to obtain food, the center LED ceased flashing, while the requirement to peck at the same location remained in effect. Shortly thereafter, pecks to the center port initiated the presentation of a song stimulus, and the trial proceeded as described in Sec. II D 2. In all cases, initial shaping occurred within one to two days, and was followed immediately by the start of song recognition training.

2. Song recognition training

Each subject was trained initially to classify sixteen 15-s song exemplars (8 exemplars from 2 singers) using a "go-nogo" operant procedure. In this procedure, subjects initiated a trial by pecking at the center response port to trigger the immediate presentation of a training song. Following stimulus presentation the animal was required to either peck the center response port again, or to withhold responses altogether. Responses to half of the stimuli (S+) were reinforced positively with 2-s access to the food hopper. Responses to the other half of the stimuli (S−) were punished by extinguishing the house light for 2–10 s and denying food access. Failure to respond to either S+ or S− stimuli had no operant consequence. For performance evaluation, we considered a response to an S+ stimulus and the withholding of a response to an S− stimulus as "correct." Withholding a response to an S+ stimulus and responding to an S− stimulus were considered "incorrect." Subjects could freely peck at the center response port throughout stimulus presentation, but only the first response within a 2-s response window beginning at stimulus offset triggered reinforcement or punishment. Responses prior to completion of the stimulus were ignored.

The stimulus exemplar presented on any given trial was selected randomly with uniform probability from the pool of all 16 stimuli the animal was learning to classify. The intertrial interval was 2 s. Water was always available. Subjects were on a closed economy during training, with daily sessions lasting from sunrise to sunset, and each subject could run as few or as many trials as they were able. Food intake was monitored daily to ensure each subject's well being. The explicit pairings of songs for baseline training was counterbalanced across subjects. All procedures were approved by the UCSD institutional animal care and use committee whose policies are consistent with the Ethical Principles of the ASA.

3. Test procedure

Prior to initiation of the first test session, the rate of food reinforcement for correct responses to S+ stimuli was lowered from 100% (where it had been during baseline training) to 80%, and the rate of "punishment" (dimmed house lights) for incorrect responses to S− stimuli was lowered to 95%. After performance again stabilized, typically within one or two sessions, we began presenting test stimuli on roughly 20% of the trials. The test stimuli were 32 naturally ordered songs composed of either familiar or novel motifs and 256 other versions of these songs (8/singer/familiarity type) where the submotif ordering was permuted (see Sec. II C 3). The test stimulus for a given trial was selected randomly from the set of all possible test stimuli for that subject, and balanced so that equal numbers of permuted and unpermuted songs were presented. We reinforced responses to test stimuli nondifferentially regardless of accuracy as follows: each response to a test stimulus had a 40% chance of eliciting a food reward, a 40% chance of eliciting punishment (timeout without food), and a 20% chance of eliciting no operant consequence. Because reinforcement of the test stimuli was random and nondifferential with respect to response outcome, subjects had no opportunity to learn to associate a given test stimulus with a given response. Thus, the correct classification of test stimuli can be taken as strong evidence for generalization rather than learning rote sets of specific training exemplars. If there was no generalization, classification accuracy would be at chance and all responses would be the same.

E. Analysis

We used d' to estimate the sensitivity for classification of baseline training song stimuli, and the various test stimuli as given by

$$d' = z(H) - z(F),$$

where H gives the proportion of responses to an S+ stimulus, F gives the proportion of responses to an S− stimulus, and $z(\,)$ denotes the z score of those random variables. The measure d' is convenient because it eliminates any biases in the response rates (e.g., due to guessing) that may vary across individuals and within individuals over time. To gauge the effect of various song manipulations during the test sessions, we compared d' values for different stimulus classes using repeated measures analysis of variance (ANOVA), and where appropriate used post-hoc analyses to quantify the significance of specific differences between mean d' measures. Identical analyses conducted on mean percent correct scores yielded the same results.

III. RESULTS

All seven subjects easily learned the initial song classification task, sorting the songs of two conspecific males into separate classes with high accuracy. The mean performance over all subjects showed significant improvement over the course of training ($F_{(6,39)}=7.65$, $p<0.0001$, repeated measures ANOVA), with individual birds requiring 1300–3800 trials to achieve reliably accurate classification (mean d' over five consecutive blocks >1.0). At asymptote, the mean (±sem) d' for all birds was 3.6 ± 0.32 (Fig. 3), and the percent correct combined for both classes of songs was 87.5 ± 0.01.

Once subjects achieved stable and accurate performance on the baseline training songs, we presented test stimuli on a subset of all trials (see Sec. II D 3). As expected from previous studies, subjects were significantly better at classifying novel songs composed of familiar motifs compared to those composed entirely of unfamiliar motifs from the same singers (i.e., motifs that did not appear in any of the baseline training songs). The mean (±sem) d' for the naturally ordered familiar-motif songs was 1.47 ± 0.21, whereas that for the naturally ordered unfamiliar-motif songs was 0.59 ± 0.13, and these values were significantly different from each other ($F_{(1,6)}=28.66$, $p<0.005$, repeated measures ANOVA). Interestingly, responding to both classes of naturally ordered songs was significantly above chance ($d'=0$) ($t=7.07$, $p<0.0005$; $t=4.60$, $p<0.005$, for songs with familiar and unfamiliar motifs, respectively). The mean proportions of correct responses to the test songs composed of familiar and unfamiliar motifs were 0.70 ± 0.04 and 0.59 ± 0.03, respectively, and both means are significantly above chance (t-test, $p\leq0.02$, both cases).

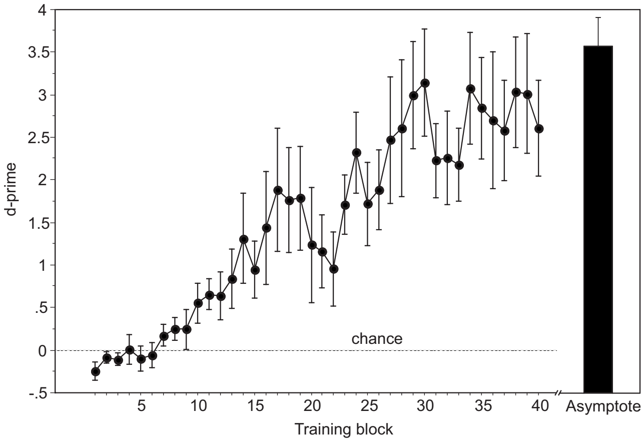

FIG. 3. Acquisition of song recognition. Mean *d*-prime (±sem) shown over the course of successive 100-trial training blocks showing the gradual improvement in classification of songs from two different singers. Performance is above chance after seven blocks, which includes acquisition of all operant task requirements (e.g., when to peck) as well as knowledge about the stimuli. Mean (±sem) performance at asymptote, just prior to testing (see methods), is shown by the black bar on the right.

The foregoing results are consistent with previous studies (Gentner and Hulse, 1998; Gentner *et al.*, 2000) in showing a clear advantage for song recognition when familiar motifs are present in novel, to-be-recognized, song bouts compared to when a bout is composed entirely of novel motifs sung by an otherwise familiar singer. This recognition advantage may result from a representation of motifs as holistic acoustic patterns or from the representation of diagnostic acoustic features centered below the level (or temporal scale) of the motif. Previous manipulations, that have only altered songs at the level of the motif, cannot distinguish between these two recognition strategies. To test the availability of acoustic pattern information at temporal scales shorter than a single motif we also asked birds to classify versions of the naturally ordered test stimuli that had been parsed into submotif features, whose order was permuted either within a motif or within the entire song (see Sec. II C 3). Permuted and naturally ordered songs were presented in the same test sessions.

The permuted songs were significantly more difficult for the subjects to recognize than the naturally ordered songs ($F_{(2,12)}=9.91$, $p<0.005$, main effect of permutation), but this effect was restricted to the familiar songs. That is, permuting

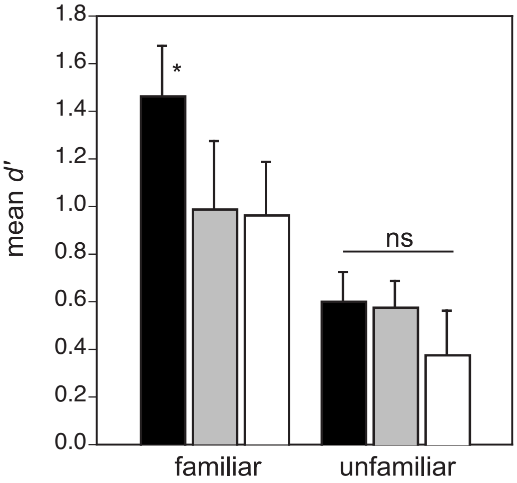

FIG. 4. Permutation test results. Mean (±sem) classification performance for the various test stimuli composed of submotif features derived from familiar (left-most bars) and unfamiliar (right-most bars) motifs. Black bars show the classification of stimuli in which the sequence of submotif features is not permuted from its natural order, gray bars when the permutation displacement is constrained by the temporal boundaries of the original motif (see methods), and white bars when the permutation is allowed over the entire exemplar. The asterisk denotes a significant difference from the other familiar-motif songs.

the submotif features in songs composed of familiar motifs lead to significant impairments in recognition ($F_{(2,12)}=8.98$, $p<0.005$, repeated measures ANOVA), whereas the same permutations of songs composed of novel motifs had no effect ($F_{(2,12)}=1.63$, NS; Fig. 4). The mean (±sem) d' values for the permutations of songs composed of familiar and unfamiliar motifs were 0.97 ± 0.18 and 0.47 ± 0.11, respectively, and both of these means were significantly greater than chance ($d'=0$) (one tailed t-test, $p\leqslant0.0008$, both cases; Fig. 4). Collapsing across permutation type (i.e., motif level or song level), the overall classification for permuted versions of the familiar-motif songs was significantly better than that for the permuted versions of unfamiliar-motif songs ($F_{(1,6)}=11.70$, $p<0.05$, repeated measures ANOVA; Fig. 4).

Surprisingly, subjects classified the permuted songs with similar proficiency regardless of whether the submotif features were shuffled over the range of their original motif or over the whole song sample. For songs composed of familiar motifs, the mean (±sem) d' values associated with motif- and song-shuffled permutations were 0.99 ± 0.29 and 0.96 ± 0.23, respectively, and the difference between these means was not significant. For songs composed of unfamiliar motifs, the mean (±sem) d' values associated with motif- and song-shuffled permutations were 0.57 ± 0.11 and 0.38 ± 0.19, respectively, and the difference between these means was not significant. The mean d' values of all of the permutations, except the song-level permutations with unfamiliar motifs, were above chance ($d'=0$) ($p<0.05$, all cases; Fig. 4).

The similar levels of recognition observed for both motif- and song-shuffled permutations of songs with familiar motifs suggests that any violation of motif-segment sequencing leads to similar deficits. This may reflect a deficit in object recognition or in the recognition of explicit sequence of submotif features. To examine these hypotheses, we asked how the severity of a given submotif permutation, measured with R distance (see Sec. II), was related to recognition. If subjects had learned the explicit sequence of submotif segments, then more severe permutations (i.e., those that broke most of the segment transitions) should be harder to recognize than those that were less severe. In fact, we observed the opposite. For songs composed of familiar motifs, recognition of the motif-shuffled songs was positively correlated with R distance ($r=0.219$, $p<0.005$). That is, as the motif-level permutations in these songs became more severe, recognition improved. For the song-shuffled versions of familiar-motif songs and both types of permutations of the unfamiliar songs there was no correlation between R distance and the recognizability of a given song ($r<0.05$, all cases, NS). The mean (±sem) R distance for familiar and novel-motif songs combined was 0.82 ± 0.003 and 0.98 ± 0.001, for permutations within motifs and over the whole song, respectively.

Given that all of the test songs were recognized at levels above chance, it is helpful to consider the degree of acoustic similarity between the various stimulus sets. DTW provides a tool for assessing similarities between complex waveforms. To assess the motif-segment similarity across songs, we parsed the baseline stimuli into constituent submotif segments using the same procedures as for the test stimuli (see Sec. II), and then found the best DTW match. We expressed the best match as the minimum path length or "distance," $D_{\min}$, (a unitless number) between each segment in the baseline songs and all the segments in the naturally ordered test songs. Smaller values for $D_{\min}$ denote greater similarity. As expected the baseline submotif segments more closely matched the segments that made up the familiar-motif test songs than the unfamiliar-motif test songs from the same singer [Fig. 5(a)]. The mean (±sem) minimum distance ($D_{\min}$) between baseline and familiar-motif songs was 37.11 ± 0.52 and that between baseline and unfamiliar-motif songs was 48.75 ± 0.64. These values are significantly different ($p<0.0001$, $t=-28.38$, paired t-test). Even though the "unfamiliar" test songs had no motifs in common with the training stimuli, the submotif segments from these songs were significantly more similar to those from the training songs of the same singer than to the segments from the training songs of the opposing singer in the training set [Fig. 5(b); see Sec. II]. The mean (±sem) minimum distance between baseline and unfamiliar-motif songs from the opposing singer was 53.14 ± 0.64, and this is significantly larger than the best matches between baseline and unfamiliar-motif segments from the same singer ($p<0.0001$, $t=-14.97$, paired t-test). The data for the individual stimulus sets are shown in Fig. 5(b).

IV. CONCLUSIONS

The song recognition system of European starlings has emerged as a valuable model for auditory processing at behavioral and physiological levels (Leppelsack, 1974; Leppelsack and Vogt, 1976; Leppelsack, 1983; Hausberger *et al.*, 2000; Gentner *et al.*, 2001; Gentner and Margoliash, 2002; George *et al.*, 2003; Gentner, 2004; Cousillas *et al.*, 2005;

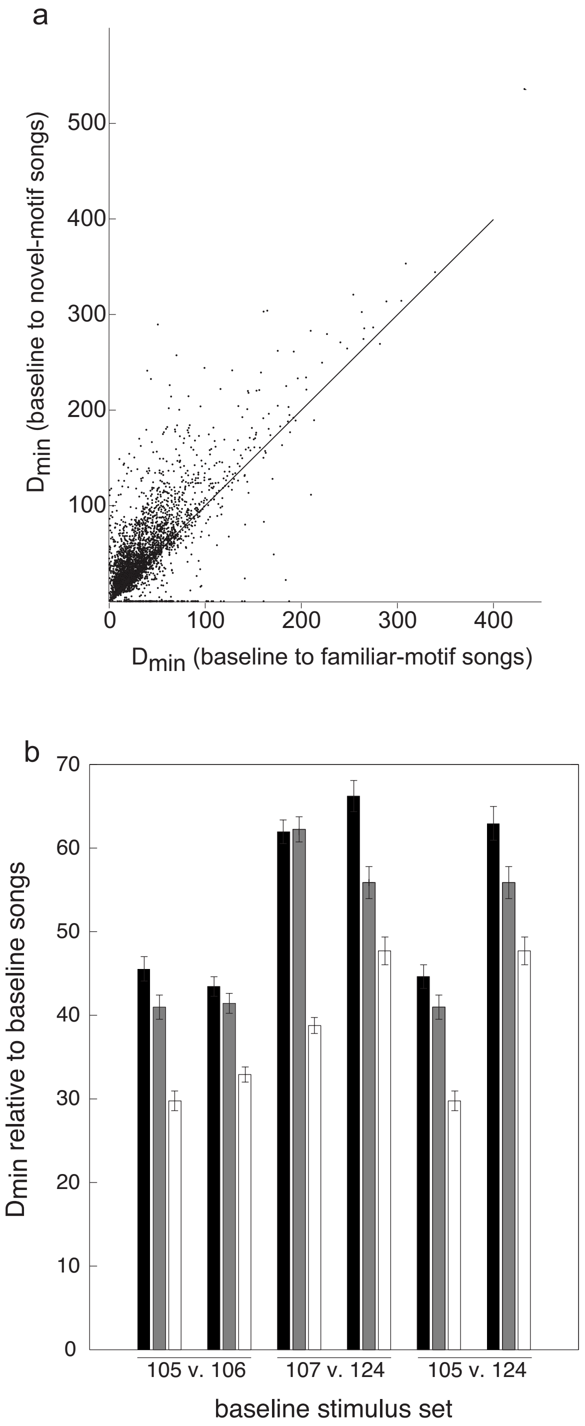

FIG. 5. Submotif feature similarity. Similarity expressed as the minimum DTW path (D_{min}; see methods) between submotif features. (a) Similarity between each submotif feature in the baseline training songs compared to those in the novel-motif test songs and to those in the familiar-motif test songs. The displacement of the distribution above the unity line reflects the fact that the features in the familiar-motif songs are more similar to the baseline submotif features, than are those in the novel-motif songs. (b) Mean similarity ($D_{min} \pm$ sd) among songs within each of the three different versions of the training and test stimuli. For a given stimulus set (e.g., "105 vs 106") the two numbers refer to the singers from which the songs were drawn and thus that the subjects learned to recognize (see methods). The three bars for each singer show the similarity of the training (baseline) songs from that bird relative to the training songs from its paired singer (black), relative to the same singer's novel-motif songs (gray), and relative to the same singer's familiar-motif songs (white). Note that the gray and white bars for a given singer are always the same, but are replotted to facilitate easy comparisons within each stimulus set.

George *et al.*, 2005; Gentner *et al.*, 2006). To date, however, studies of song recognition have relied on relatively long timescale manipulations of motif sequences, over several hundreds to thousands of milliseconds, to control recognition. The present results address how temporal patterning at the submotif level, in the range of tens to hundreds of milliseconds, effects the recognition of individual starling songs. We show that song recognition suffers significant impairments when the temporal sequencing of submotif level acoustic features is permuted, and that the effects of altering the sequencing of submotif features are restricted to familiar, i.e., learned, motifs. These results support two main conclusions. First, learned temporal patterns of submotif features, i.e., motifs, are perceptually salient auditory objects that emerge without explicit reinforcement, and which are positioned within a hierarchy of acoustic patterns. Second, starlings can readily access information at multiple levels within the acoustic pattern hierarchy for individual vocal recognition.

A. Auditory objects and song

The concept of an auditory object has received substantial research attention in recent years (Griffiths and Warren, 2004 for review), but remains somewhat controversial as a framework for understanding auditory perception. Often times, the term "auditory object" is used in the context of scene analysis (Bregman, 1990) to refer to the mental representation of an environmental sound source rather than the source itself or the sound it produces (e.g. Alain and Arnott, 2000). More generalized examples of auditory objects are common in the words and syllables that make up human language (Warren and Bashford, 1993; Warren, 1999). Thus, the concept of an auditory object has broad intuitive appeal. Indeed, by analogy to vision where the notion of an object is more intuitive (and dominant), it may be that consideration of natural auditory perception in the absence of objects is unrealistic. Pragmatically, the empirical questions concern the features of the acoustic signal that guide object structure, and the conditions under which different sets of these structural constraints hold. For example, the same acoustic communication signal may contain information about a range of relevant external events including the location of the sound source, its species, sex, individual identity, and more specific semantic content (e.g., food quality). As in vision, auditory objects are likely to exist at multiple scales spanning the relevant physical dimensions of the signal, i.e., time and spectral frequency, and features that define object boundaries in one context may be irrelevant in another. This line of reasoning suggests that a discussion of auditory objects requires explicit links to well-defined goals of the perceptual/cognitive system under investigation.

From this position, we define an auditory object as a set of coincident, or closely coincident, acoustic events that can be perceived as a whole, and that carries with it behaviorally relevant information. The results of the present study support the idea that starling song motifs form salient auditory objects of this sort. When starlings are trained to classify large sets of songs according to singer, they are significantly better at recognizing novel songs from the training singers in which the natural sequences of submotif level features is preserved, compared to songs in which the natural sequence of submotif features is permuted. If the starlings had learned to recognize songs by attending only to salient submotif features, that is by learning sets of notes rather than sets of motifs, then permuting the sequence of notes should not have impaired recognition. Instead, the effects of the note level permutations suggest that motifs constitute auditory objects that convey a significant portion of the acoustic information required for individual song recognition. We also show that the effects of permuting the sequence submotif features are restricted to familiar, i.e., learned, motifs. Recognition of songs comprising unfamiliar motifs from the training singers was not significantly affected by the same permutations. Importantly, although the perception of these auditory objects appears dependent on learning, reinforcement was never explicitly tied to an object-based solution strategy. Thus the perceptual sensitivity to these auditory objects is an emergent product, rather than target, of learning. This emergence may reflect a parsimonious solution to the recognition of spectrotemporally complex signals and/or the effects of segmentation and phrasing constraints imposed by song production mechanisms.

The pattern of errors associated with different permutations of the learned submotif feature sequences is also consistent with an object-level perception of motifs. We found no significant differences in the submotif feature permutations that operated over the whole song and those that were confined to the boundaries of each segment's original motif. The similar level of recognition observed for both kinds of permutations—those over the motif and those over the song—indicates that the knowledge of temporal patterning within a motif is very precise, but not necessarily explicit. Had the starlings learned each motif as a perceptually explicit sequence of submotif features, one would expect local (within motif boundary) permutations to be less disruptive to recognition than permutations over the entire song stimulus. Instead, any violation of the learned temporal sequence of submotif features, at least down the to the level tested here, appears to be disruptive for recognition. Likewise, the observation of a *positive* correlation between *R* distance (our measure for permutation severity) and recognition of familiar-motif-shuffled songs runs contrary to the notion that starlings learned explicit sequences of submotif features. Had they done so, one would expect to see a negative relationship between *R* distance and performance, because stronger violations of the training sequence should be, if anything, harder to recognize. Instead, the smallest changes to the submotif pattern structure exact the greatest toll on recognition.

Therefore, we conclude that submotif feature sequences form "temporal compounds," or global auditory patterns, of the sort described for human audition (Warren and Bashford, 1993; Warren, 1999).

Exploring the precise lower bound for the length of component features within a temporal sequence is topic for future research. At some temporal scale, one might predict that sequence permutations would simply abolish all recognition. In any case, the present results demonstrate that sensitivity to temporal pattering within the boundaries of a motif extends well down into the range of tens of milliseconds. This approaches the thresholds for classic psychophysical tests of temporal integration of roughly 2 ms (Klump and Maier, 1989). Interestingly, a 10 ms lower bound to temporal order sensitivity is consistent with the optimal time window found for discrimination of patterned spike trains elicited by song in the avian analog to primary auditory cortex (Narayan *et al.*, 2006), and integration windows at higher regions in the sensory hierarchy more closely reflect timescales close to mean segment and motif duration, ~200 and 1000 ms, respectively (Thompson and Gentner, 2007). In general, the neural bases that underlie object level representation of temporally patterned acoustic sequences are unexplored. Single neuron extracellular recordings from caudo-medial mesopallium (CMM), a region in the songbird forebrain analogous to mammal auditory cortex show strong responses to motifs in the songs that birds have learned to recognize (Gentner and Margoliash, 2003). This work provides a phenomenological correlate to such objects, and suggests that starlings and other songbirds will be useful organisms within which to explore these issues.

Although the strongest recognition is reserved for songs comprising naturally ordered renditions of familiar motifs, motif percepts do not appear to be holistic in the strictest sense of that term (see Warren, 1999). In the extreme, a holistic object percept is one that cannot be recognized by any one of (or any sequentially permuted subset of) its constituent parts, and the present pattern of results fails to meet this definition. Subjects suffered significant impairment when the sequencing of submotif features in familiar-motifs songs was permuted, but recognition of these permuted songs remained reliably above chance (Fig. 4). The only explanation for above chance recognition of the permuted familiar-motif songs is that all of the task-relevant diagnostic information was not abolished by the submotif permutation. Here again, as for motif level recognition, the effects of learning are evident as even the permuted versions of familiar-motif songs were more easily recognized than any of the unfamiliar-motif songs, including those with naturally ordered submotif features. Thus, the submotif features themselves, as well as their temporal patterning, are learned, and starlings have access to both levels of acoustic information when making classification judgments that reflect individual singer identity. It is unclear whether one or both temporal scales can carry information beyond individual identity (e.g. Hausberger *et al.*, 1995; Hausberger, 1997). The "classic" view on talker recognition in humans holds that individual identity information is coded separately in non-linguistic components of the vocal signal (Bricker and Pruzansky, 1976), but phonetic components can also contribute to talker recognition (Sheffert *et al.*, 2002). It is interesting to speculate that the presence of perceptually separable temporal scales in a non-human vocal communication signal, as we have shown here, may set the stage for increasingly complex forms of encoding observed in human vocal communication.

B. Assessing auditory object similarity

Taken together, the results of these and earlier behavioral studies (Gentner, 2004) suggest that when starlings learn to recognize conspecific songs from different singers, they memorize large numbers of unique motifs corresponding to each individual singer. However, although the temporal organization of submotif features plays a clear role in guiding song recognition, information coded at the level of the motif cannot explain song recognition entirely. Even when motif identity and the corresponding submotif temporal structure are abolished, recognition is above chance, albeit only modestly, at a severely impaired level (Fig. 4). This result conflicts slightly with previous results showing that recognition falls to chance when starlings are presented with song composed entirely of novel motifs sung by otherwise familiar singers (Gentner *et al.*, 2000). One possible explanation for this difference in recognition of novel motif songs is methodological. Previous studies used a slightly different training protocol in which the subject's task was to give one response to songs of a single male and another response to songs from four other males. Compared to the one-versus-one design used in the present study, the one-versus-many version of the recognition task necessarily involves lower motif overlap within the set of "many" songs. The lower motif overlap, in turn, likely makes the previous task more difficult which may mask the very subtle recognition of unfamiliar-motif songs observed in the present study. The different studies also drew subjects from different populations of starlings, and it is theoretically possible that perceptual sensitivity to the features that permit the modest recognition of novel motifs from familiar singers varies across populations.

Regardless of the source, starlings are able to recognize songs comprising unfamiliar motifs from otherwise familiar singers. The fact that natural motif organization provided no measurable improvement to the recognition of these unfamiliar-motif songs indicates that the information used to achieve this marginal recognition is carried by (and only by) the submotif features. This information may reflect true "voice characteristics," i.e., singer-invariant acoustic properties, imparted to all or a subset of the motifs sung by a single individual. To examine the features that guide the very subtle recognition of novel motifs one can look to the similarity measures employed in the present study. Unique motifs may result from individually specific groupings of submotif features that are shared among all starlings or from submotif features that are themselves specific to each individual. If submotif features are shared between birds, then two sets of novel motifs from one singer should be approximately "as similar" to a third set of motifs from another singer as the features making up all three sets of motifs are drawn from a common pool. If, however, submotif features are specific to each individual, then two distinct sets of motifs from the same singer may be more similar than two sets of motifs from different singers, as a bird may be more likely to use the same submotif feature in more than one motif. We found that for each of the three different stimulus sets, disjoint sets of motifs from the same singer were significantly more similar than sets of motifs from different singers. This suggests that submotif features are specific to each individual bird, and may be taken as evidence for a weak form of "voice characteristic" imparted to a least a subset of the notes that a bird produces. The use of voice characteristics (e.g., vocal timbre, the frequency of glottal pulsation, and spectral contours imparted by laryngeal morphology) is well documented for individual talker recognition in humans (Bricker and Pruzansky, 1976). A more precise understanding of the information that starlings used to recognize the unfamiliar-motif songs awaits further investigation.

Although starling song recognition is guided largely by the memorization of unique notes and their temporal patterns, the presence of and perceptual sensitivity to voice characteristics in any nonhuman animal, however subtle, is important theoretically. For humans, the rich semantic content of our language necessitates that most words are shared among speakers, and so precludes a "repertoire memorization" strategy for individual recognition. Instead, the voice characteristics used in speaker recognition appear to be coded in acoustic parameters of the signal that are predominantly non-linguistic (Remez *et al.*, 1997). Our results indicate that an independent communication channel exists in at least one other species. It is interesting to speculate that its subtle role in songbird vocal recognition may represent an unexploited, and unnecessary, capacity of vocal communication signals that lack a rich combinatorial semantic structure.

ACKNOWLEDGMENTS

Grant No. R01DC008358-01A1 from NIH NIDCD supported this research. The author thanks Satoru Fukushima for assistance in programming a portion of the DTW algorithm.

Adret-Hausberger, M., and Jenkins, P. F. (**1988**). "Complex organization of the warbling song in starlings," Behaviour **107**, 138–156.

Alain, C., and Arnott, S. R. (**2000**). "Selectively attending to auditory objects," Front. Biosci. **5**, D202–212.

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

Bregman, A. S. (**1990**). "The auditory scene," in *Auditory Scence Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge).

Bricker, P. D., and Pruzansky, S. (**1976**). "Speaker recognition," in *Contemporary Issues in Experimental Phonetics*, edited by N. J. Lass (Academic, New York), pp. 295–326.

Chaiken, M., Böhner, J., and Marler, P. (**1993**). "Song Acquisition in European Starlings, Sturnus vulgaris: a comparison of the songs of live-tutored, tape-tutored, untutored, and wild-caught males," Anim. Behav. **46**, 1079–1090.

Cousillas, H., Leppelsack, H. J., Leppelsack, E., Richard, J. P., Mathelier, M., and Hausberger, M. (**2005**). "Functional organization of the forebrain auditory centres of the European starling: a study based on natural sounds," Hear. Res. **207**, 10–21.

Eens, M. (**1992**). "Organization and functin of the song in the European starling *Sturnus vulgaris*," University of Antwerp, Thesis.

Eens, M. (**1997**). "Understanding the complex song of the European starling: An integrated approach," Advances in the Study of Behavior **26**, 355–434.

Eens, M., Pinxten, M., and Verheyen, R. F. (**1989**). "Temporal and sequential organization of song bouts in the European starling," Ardea **77**, 75–86.

Eens, M., Pinxten, R., and Verheyen, R. F. (**1991**). "Organization of Song in the European Starling—Species—Specificity and Individual-Differences," Belg. J. Zoolog. **121**, 257–278.

Gentner, T. Q. (**2004**). "Neural systems for individual song recognition in adult birds," Ann. N.Y. Acad. Sci. **1016**, 282–302.

Gentner, T. Q., Fenn, K. M., Margoliash, D., and Nusbaum, H. C. (**2006**). "Recursive syntactic pattern learning by songbirds," Nature (London) **440**, 1204–1207.

Gentner, T. Q., and Hulse, S. H. (**1998**). "Perceptual mechanisms for individual vocal recognition in European starlings, Sturnus vulgaris," Anim. Behav. **56**, 579–594.

Gentner, T. Q., and Hulse, S. H. (**2000**). "Perceptual classification based on the component structure of song in European starlings," J. Acoust. Soc. Am. **107**, 3369–3381.

Gentner, T. Q., Hulse, S. H., Bentley, G. E., and Ball, G. F. (**2000**). "Individual vocal recognition and the effect of partial lesions to HVc on discrimination, learning, and categorization of conspecific song in adult songbirds," J. Neurophysiol. **42**, 117–133.

Gentner, T. Q., Hulse, S. H., Duffy, D., and Ball, G. F. (**2001**). "Response biases in auditory forebrain regions of female songbirds following exposure to sexually relevant variation in male song," J. Neurophysiol. **46**, 48–58.

Gentner, T. Q., and Margoliash, D. (**2002**). *The Neuroethology of Vocal Communication: Perception and Cognition* (Springer, Berlin).

Gentner, T. Q., and Margoliash, D. (**2003**). "Neuronal populations and single cells representing learned auditory objects," Nature (London) **424**, 669–674.

George, I., Cousillas, H., Richard, J. P., and Hausberger, M. (**2003**). "A new extensive approach to single unit responses using multisite recording electrodes: application to the songbird brain," J. Neurosci. Methods **125**, 65–71.

George, I., Cousillas, H., Richard, J. P., and Hausberger, M. (**2005**). "State-dependent hemispheric specialization in the songbird brain," J. Comp. Neurol. **488**, 48–60.

Griffiths, T. D., and Warren, J. D. (**2004**). "What is an auditory object?," Nat. Rev. Neurosci. **5**, 887–892.

Hausberger, M. (**1997**). "Social influences on song acquisition and sharing in the European starling (*Sturnus vulgaris*)," in *Social Influences on Vocal Development*, edited by C. Snowden and M. Hausberger (Cambridge University Press, Cambridge), pp. 128–156.

Hausberger, M., and Cousillas, H. (**1995**). "Categorization in birdsong: From behavioural to neuronal responses," Behav. Processes **35**, 83–91.

Hausberger, M., Leppelsack, E., Richard, J., and Leppelsack, H. J. (**2000**). "Neuronal bases of categorization in starling song," Behav. Brain Res. **114**, 89–95.

Hausberger, M., Richard-Yris, M.-A., Henry, L., Lepage, L., and Schmidt, I. (**1995**). "Song sharing reflects the social organization in a captive group of European starlings (*Sturnis vulgaris*)," J. Comp. Psychol. **109**, 222–241.

Klump, G. M., and Maier, E. H. (**1989**). "Gap detection in the starling (*Sturnus vulgaris*): I Psychophysical thresholds," J. Comp. Physiol., A **164**, 531–538.

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.

Kroodsma, D. E. (**1989**). "Pseudoreplication external validity and the design of playback experiments," Anim. Behav. **38**, 715–719.

Leppelsack, H.-J. (**1974**). "Functional Properties of the Acoustic Pathway in the Field L of the Neostratum caudale of the Starling," J. Comp. Physiol. **88**, 271–320.

Leppelsack, H.-J. (**1983**). "Analysis of song in the auditory pathway of song birds," in *Advances in Vertebrate Neuroethology*, edited by J. P. Evert, B. R. Capranica, and D. J. Ingle (Plenum, New York), pp. 783–799.

Leppelsack, H. J., and Vogt, M. (**1976**). "Response to auditory neurons in the forebrainof a song bird to stimulation with species-specific sounds," J. Comp. Physiol. **107**, 263–274.

Narayan, R., Grana, G., and Sen, K. (**2006**). "Distinct time scales in cortical discrimination of natural sounds in songbirds," J. Neurophysiol. **96**, 252–258.

Remez, R. E., Fellowes, J. M., and Rubin, P. E. (**1997**). "Talker identification based on phonetic information," J. Exp. Psychol. Hum. Percept. Perform. **23**, 651–666.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., and Remez, R. E. (**2002**). "Learning to recognize talkers from natural, sinewave, and reversed speech samples," J. Exp. Psychol. Hum. Percept. Perform. **28**, 1447–1469.

Thompson, J. V., and Gentner, T. Q. (**2007**). "Temporal- and rate-coding schemes, and the emergence of recognition for complex acoustic communication signals," in *Society for Neuroscience* (Society of Neuroscience Abstracts, San Diego, CA).

Warren, R. M. (**1999**). *Auditory Perception: A New Analysis and Synthesis* (Cambridge University Press, New York).

Warren, R. M., and Bashford, J. A. (**1993**). "When acoustic sequences are not perceptual sequences: The global perception of auditory patterns," Percept. Psychophys. **54**, 121–1261.