Distributed recognition of natural songs by European starlings

Daniel Knudsen[a], Jason V. Thompson[a], Timothy Q. Gentner[a,b,*]

[a] Neuroscience Graduate Program, University of California, San Diego, CA, United States
[b] Department of Psychology, University of California, San Diego, CA, United States

ARTICLE INFO

ABSTRACT

Individual vocal recognition behaviors in songbirds provide an excellent framework for the investigation of comparative psychological and neurobiological mechanisms that support the perception and cognition of complex acoustic communication signals. To this end, the complex songs of European starlings have been studied extensively. Yet, several basic parameters of starling individual vocal recognition have not been assessed. Here we investigate the temporal extent of song information acquired by starlings during vocal recognition learning. We trained two groups of starlings using standard operant conditioning techniques to recognize several songs from two conspecific male singers. In the first experiment we tested their ability to maintain accurate recognition when presented with (1) random sequences of 1–12 motifs (stereotyped song components) drawn from the training songs, and (2) 0.1–12-s excerpts of continuous song drawn from the training songs. We found that song recognition improved monotonically as more vocal material is provided. In the second experiment, we systematically substituted continuous, varying length regions of white noise for portions of the training songs and again examined recognition accuracy. Recognition remained above chance levels for all noise substitutions tested (up to 91% of the training stimulus) although all but the smallest substitutions led to some decrement in song recognition. Overall, above chance recognition could be obtained with surprisingly few motifs, short excerpts of song, and in the absence

* Corresponding author at: University of California, San Diego, Department of Psychology, 9500 Gilman Dr. MC0109, La Jolla, CA 92093, United states.
E-mail address: tgentner@ucsd.edu (T.Q. Gentner).

of large portions of the training songs. These results suggest that starlings acquire a representation of song during individual vocal recognition learning that is robust to perturbations and distributed broadly over large portions of these complex acoustic sequences.

Vocal recognition is a widespread phenomenon among songbirds (reviewed by Stoddard, 1996) and serves two broad goals: species recognition, and recognition of individual conspecifics. Myriad acoustic features may guide vocal recognition, but it follows that selection pressures will emphasize the features that most reliably distinguish one's own species' songs from those of other sympatric species (see Nelson & Marler, 1990), and those that differentiate individual conspecifics. Among field sparrows, for example, frequency appears to be the most important cue for species recognition (Nelson, 1989). Likewise, laboratory studies demonstrate that songbirds are more adept at discriminations involving conspecific songs as compared to discriminations involving heterospecific songs (Cynx & Nottebohm, 1992; Dooling, Brown, Klump, & Okanoya, 1992; Heinz, Sinnott, & Sachs, 1980). Recognition systems appear to be tuned to the particular acoustic space of one's own species' vocalizations (Nelson, 1988) but also maintain more general capacities to distinguish among many different types of song. Within a species, particularly those with acoustically diverse songs, many acoustic features may be relevant for classification of song components (Sturdy, Phillmore, & Weisman, 2000). Recent laboratory studies of European starlings have addressed these questions directly by determining more precisely the form of the acoustic signal controlling individual recognition.

Male starlings present their songs in long episodes of continuous singing referred to as bouts. Song bouts, in turn, are composed of much smaller acoustic units referred to as motifs (Adret-Hausberger & Jenkins, 1988; Eens, Pinxten, & Verheyen, 1991), which in turn are composed of still smaller units called notes. Notes can be broadly classified by the presence of continuous energy in their spectrographic representations, and although several notes may occur in a given motif, their pattern is usually highly stereotyped between successive renditions of the same motif (a sample of song is shown in Fig. 1). One can thus consider starling song as a sequence of motifs, where each motif is an acoustically complex event. The number of unique motifs that a male starling can sing (i.e. his repertoire size) can be large, and different song bouts from the same male do not necessarily comprise the same set of motifs. The broad acoustic variation in their song provides several potential cues that starlings might use when learning to recognize the songs of an individual conspecific and while maintaining that recognition over time. One straightforward recognition mechanism is the association of specific motifs with specific singers.

Classifying songs according to their component (motif) structure represents an efficient strategy for representing these complex acoustic signals (see Fig. 1). Because individual starlings tend to possess unique motif repertoires (Adret-Hausberger & Jenkins, 1988; Eens, Pinxten, & Verheyen, 1989; Eens et al., 1991), disjoint sets of motifs will generally correspond to individual identity. Attending to the motif structure, therefore, captures a significant portion of the individual variation in the signal, albeit at the expense of requiring a large memory capacity. From a human perspective, this might seem a suboptimal strategy for individual vocal recognition, because the total set of vocal signals that can be produced by a single individual is theoretically infinite. Individual talker recognition in humans relies on voice characteristics such as timbre, glottal pulsation frequency, and spectral contours imparted by laryngeal morphology (Bricker & Pruzansky, 1976) that are largely independent of the linguistic content (Remez, Fellowes, & Rubin, 1997). The incorporation of individual identity information into the phonetics of the vocal signal, as appears to be the case for starlings, may be a plesiomorphic (evolutionarily older) state for human vocal communication. As the size of the lexicon and sharing of signal elements increases between individuals, however, the exclusive use of this 'object-dependent' strategy for individual vocal recognition becomes increasingly inefficient. It is likely that starlings, along with other organisms that rely on acoustically complex vocal signals to mediate adaptive behaviors, posses multiple mechanisms for extracting behaviorally relevant information from communication signals.

Although recognition of the complex songs of European starlings has been studied extensively, several basic parameters of starling individual vocal recognition have not been assessed. Recent results
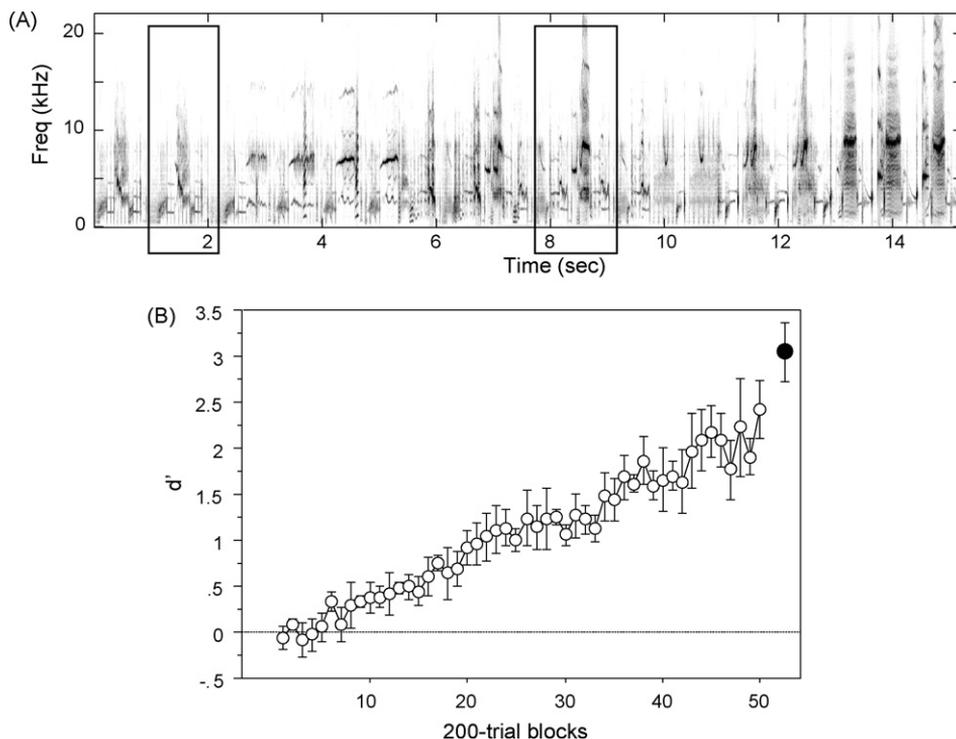
Fig. 1. Baseline stimuli and acquisition for Experiment 1. (A) Example spectrogram of one baseline training song. Note the diverse acoustic structure of the song. Example motifs are outlined by the boxes just before 2 and 8 s. (B) Mean (±sem) d' (open circles) for acquisition of the baseline song recognition task plotted over the first 50 200-trial blocks. The single point (open square) on the right shows the recognition accuracy averaged over the last five 200-trial blocks prior to probe testing.

indicate that some voice characteristics are present in starling song, and that perceptual knowledge spans multiple hierarchical levels of song organization (Gentner, 2008). Here, we investigate the temporal extent of song-knowledge that starlings acquire during canonical vocal recognition tasks.

General methods

Subjects

Ten adult European starlings, *Sturnus vulgaris*, served as subjects in this study. All subjects were wild 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 & Hulse, 2000), and the sex of subjects in this study was not controlled.

Apparatus

Starlings learned to classify the training stimuli using a custom-built operant apparatus, housed in a 61 cm × 81 cm × 56 cm ID sound attenuation chamber (Acoustic Systems). Inside the chamber, a subject was held in a weld-wire cage (41 cm × 41 cm × 35 cm) that permitted access to a 30 cm × 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 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 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.

Stimuli

Song recording

Recordings of eight 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 & Hulse, 1998). Briefly, a minimum of 0.5 h of song was recorded from each male while 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). The multiple songs of each bird were parsed into roughly 10 to 15-s exemplars of continuous singing taken from the beginning, middle, or end of a typically much longer song bout. Human observers labeled the motifs in each exemplar. These same stimuli have been used to explore the role of motif familiarity in the recognition of individual songs in several studies (Gentner & Hulse, 1998, 2000; Gentner, Hulse, Bentley, & Ball, 2000; Gentner & Margoliash, 2003). None of the males whose songs were used to generate the stimuli for the present study served as a subject in the operant testing described here.

Procedure

Shaping

Subjects learned to work the apparatus through a series of successive shaping procedures. Upon initially entering the operant chamber, the subject was allowed unrestricted access to the food hopper on a variable interval schedule, and then taught through auto-shaping to peck at a flashing LED in the center port to gain access to the food hopper. Once the subject pecked reliably at the center port to obtain food, pecks to the center port caused the center LED to cease flashing and the LED in either the left or right response port to begin flashing (side selected at random with equal probability). The subject was then required to peck at the port with the flashing LED to raise the food hopper. Once behavior to obtain food was again reliable (typically 100 trials), the center LED ceased flashing, while the requirement to peck at the center port to trigger the left or right LED to flash remained in effect. Shortly thereafter, pecks to the center port initiated the presentation of a song stimulus, and the trial proceeded as described below. In all cases, initial shaping occurred within 1–2 days, and was followed immediately by the start of song recognition training.

Song recognition training

Subjects learned to classify the two sets of song exemplars using a two-alternative choice (2AC) procedure. Subjects initiated a trial by pecking the center response button to trigger immediately the presentation of a single song exemplar. The specific baseline training stimuli are described in the methods for each experiment. Following completion of the song exemplar, the subject was required to peck either the left or the right response port within 2 s. Half of the song exemplars were associated with the left response port and the other half with the right port. The training songs were divided according to individual identity, such that a single individual sang the songs in each set. Thus, the

baseline training corresponds to basic individual vocal recognition (Gentner & Hulse, 1998, 2000). Correct responses (i.e. pecking the left/right port after hearing a song associated with that port) were rewarded with access to the food hopper for 3 s. Incorrect responses (i.e. pecking the left/right port after hearing a song associated with the opposite port) were punished by extinguishing the house light for 2–10 s and withholding food. Responses prior to completion of the stimulus were ignored. The trial ended when either the food hopper retracted following a correct response, or the house light re-illuminated following an incorrect response. The inter-trial interval was 2 s. Incorrect responses, and any trial in which the subject failed to peck either the left or the right key within 2 s of stimulus completion initiated a correction-trial sequence, during which the initiating stimulus was repeated on all subsequent trials until the animal responded correctly. Correction-trial data are not included in the analysis. The song exemplar presented on any given trial was selected randomly (with replacement) from the pool of all stimuli the animal was learning to recognize. Water was always available. Subjects were on a closed economy during training, with daily sessions lasting from sunrise until sunset. Food intake was monitored daily to insure wellbeing.

Test procedure

During later sessions, we presented test stimuli to gauge the basis for recognition abilities that each subject acquired during baseline training. Prior to initiation of the first test session, the rate of food reinforcement for correct responses and "punishment" (dimmed house lights) for incorrect responses was reduced from 100% where it had been during baseline training. After performance stabilized at reduced reinforcement rates, typically within one or two sessions, we began presenting test stimuli on a subset of the trials. The specific test stimuli, their presentation rates, and the reinforcement rates for probe sessions are given in the methods for each experiment. The test stimulus for a given trial was selected randomly from the set of all possible test stimuli for that subject. We reinforced responses to test stimuli non-differentially regardless of accuracy, such that each response to a test stimulus had an equal, non-zero chance of eliciting a food reward, the same chance of eliciting punishment (timeout without food), and some comparable (though not necessarily equal) chance of eliciting no operant consequence. Because reinforcement of the test stimuli was random and non-differential with respect to response outcome, subjects had no opportunity to learn to associate a given test stimulus with a left or right response. Thus, the correct classification of test stimuli can be taken as strong evidence for vocal recognition rather than rote learning of specific training exemplars (Hulse, 1995). If there is no recognition, classification accuracy will be at chance.

Analysis

We used *d*-prime (*d′*) to estimate the sensitivity for classification of baseline training song stimuli, and the various test stimuli as given by the equation:

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

where *H* gives the proportion of correct responses to one stimulus class (hit rate), *F* gives the proportion of incorrect responses to the opposing stimulus (false alarm rate), 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 ANOVA, and where appropriate used post hoc analyses to quantify the significance of specific differences between mean *d′* measures. We take *d′* = 0 to be chance performance.

Confidence interval simulations

When considering the data for a single subject (Figs. 2, 4 and 8), we used simulations to generate confidence intervals around either chance or a given bird's baseline performance over a number of trials. For confidence intervals around chance, we simulated 1000 blocks of trials in which the number of trials and overall response rates in a simulated block matched the real data (in 200-trial blocks for
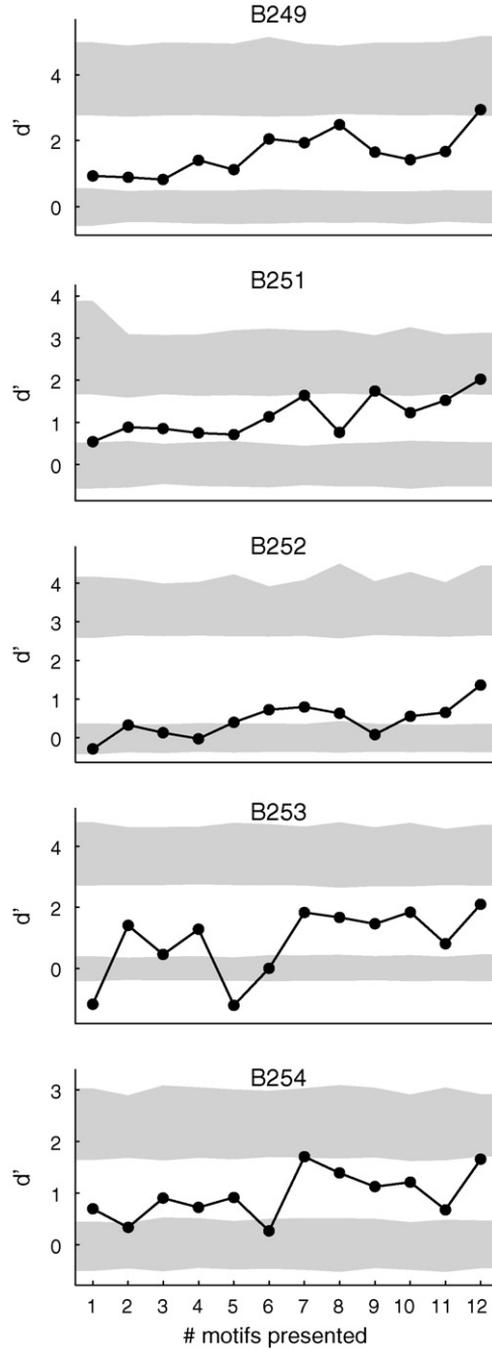
Fig. 2. Individual data for motif-probes. Each panel shows the behavior of an individual subject on motif-probe test trials. Circles indicate performance (d') as a function of the number of motifs presented. The upper region of gray shading in each figure is the 95% confidence interval around baseline performance, while the lower shaded region at the bottom of each figure describes the 95% confidence interval around chance. Individual subject identities are at the top of each panel.

training, or in blocks with the same number of trials as the experimental block of interest), but for which the response selection (e.g. left or right) was randomized. For confidence intervals around a bird's baseline performance, we again created 1000 simulated blocks of trials. However, this time we used a bootstrap technique in which each simulated block was created by randomly selecting, from among all the baseline trials in a session, a number of baseline trials equal to the number of test trials in the experimental block of interest. To calculate the confidence interval, we calculated d' for each of the 1000 simulated blocks. From this distribution of d' values expected by either chance or the bird's baseline performance, we took the 95th percentile as the critical statistical bound.

Comparisons across subjects

The absolute range of d' varies with trial number and response rate. To compensate for this in statistical analyses involving comparisons between subjects, we verified all reported effects using a normalized d' in which we express the value for a given class of test stimuli as a proportion of baseline performance: norm-$d' = (d'$ test$)/(d'$ baseline$)$. Analyses conducted on normalized d' values yielded results qualitatively identical to those reported for the raw d' measures. For ease of interpretation, and unless otherwise noted, we show the raw d' values and report the corresponding statistics.

Experiment 1

Starlings readily learn to classify the songs of individual conspecific singers at very high levels of accuracy (Gentner & Hulse, 1998, 2000), and this ability is tied closely to the motif structure of songs (e.g. Gentner, 2008). However, the large numbers of motifs within each song and their broad acoustic variability are likely to yield many redundant cues to aid song classification. In Experiment 1, we trained birds to recognize songs from two singers, then measured classification accuracy for subsections of the training songs by presenting (1) strings of 1–12 randomly ordered motifs from the training songs or (2) continuous 0.1–12 s excerpts of a training song.

Methods

Subjects

Five starlings participated in Experiment 1. At the start of training, all birds were naïve to the operant apparatus and the stimuli used in the experiment.

Baseline training stimuli

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 song exemplars and the assignment of those exemplars to either the left or right response port 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 in the general methods. 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 & Jenkins, 1988; Eens et al., 1991). Those sampled at later time points in a bout comprised only 'warble' motifs. Previous data indicate that recognition is easily learned with this length of a song exemplar (Gentner, 2008) 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 & Hulse, 1998).

Testing procedure and stimuli

After subjects had reached a stable, accurate level of performance on the baseline song classification task, we lowered the rate of reinforcement and punishment from 100% to 60%. Thus, on 40% of the trials there was no consequence for an operant response. We presented two types of test stimuli in separate sessions: (1) a random assortment of 1–12 motifs drawn from the songs used during baseline training, referred to as a "motif-probe", and (2) 0.1–12.0 s excerpt of continuous song drawn from one of the baseline training songs, referred to as an "excerpt-probe".

To construct the motif-probe stimulus on each trial, we chose randomly, with replacement and equal probability, from the set of all motifs that made up the eight songs in one of the two baseline training classes for a given subject. Motifs from the two baseline training classes were never combined in a single test exemplar. The same motif could appear multiple times within a given test sequence, but never in two or more consecutive positions (i.e. there were no consecutive motif repeats). Preliminary, unpublished data suggest that neither motif repetition nor motif selection from a common training song affects recognition under these conditions. The motif-probe test stimuli were presented on 10% of trials, chosen at random. Subjects were presented an average of 146 ± 4 (range: 99–197) motif-probe stimuli with a given number of motifs (1–12). We presented a maximum of 12 motifs because that was the smallest number contained in any of the baseline training stimuli.

To construct the excerpt-probe stimulus on a given trial we randomly chose a continuous portion of song from a randomly selected baseline training song such that the excerpt had a given duration and offset from the start of the original training exemplar. The duration and offset of each excerpt-probe was drawn from a uniform distribution of values ranging between 0.100–12.000 s and 0–12.000 s, respectively. The magnitude of the offset chosen on any given trial necessarily constrains the maximum duration of the stimulus that can be excerpted from the song. As the baseline training stimuli were all roughly 15-s long, the values used here allow us to examine recognition of excerpts from 0.1 to 3 s long taken at all offset values between 1 and 12 s. For excerpts longer than 3 s, the maximum offset we could examine decreased proportionally. The exact value chosen for the duration and offset on any given trial varied at millisecond resolution. Because the excerpt algorithm was agnostic to natural motif boundaries, we added a 10-ms linear ramp to the start and end of the excised waveform to eliminate any potential onset or offset transients induced by the signal manipulation. Subjects were presented an average of 642 ± 69 (range: 57–1906) excerpt-probe stimuli in each of 12 classes (defined by their duration, see below). Shorter duration excerpts could appear with a greater range of possible offsets, and so appeared more frequently than longer duration excerpts.

Analysis

We computed d' for all relevant classes of the stimuli. For the motif-probes, we binned stimuli into classes according to the number of motifs in each test stimulus. For the excerpt-probes, we binned the stimuli into 12 1-s bins: 0.000–1.000 s, 1.001–2.000 s, 2.001–3.000 s, etc., and computed the corresponding d' for stimuli in each bin. For all probe stimuli in a given bin, the 'hit rate' was taken to be the proportion of correct responses to test stimuli derived from one baseline singer, and the 'false alarm rate' is the proportion of incorrect responses to test stimuli derived from the opposing singer. The explicit choice of the singer tied to 'hit' or 'false alarm' proportions does not affect the value of d'.

Results

All subjects learned to recognize the 16 baseline training songs very quickly and ultimately reached a high degree of accuracy in classifying the songs of the two singers. Fig. 1B shows the mean acquisition curve for the five birds exposed to the baseline training in Experiment 1. Recognition performance improved significantly over the course of learning ($F_{(4, 49)} = 13.8$, $p < 0.0001$, rmANOVA, main effect of training), and performance, averaged across subjects, was consistently above chance levels by the ninth 200-trial block (Fig. 1B). On average subjects required 3699 ± 369 trials (range: 2714–4905) to reach a stable performance criterion where d' was greater than 1.0 for three consecutive 200-trial blocks. The mean ($\pm$sem) d' immediately prior to the start of the first test session was 3.05 (± 0.32), demonstrating very accurate recognition of the songs in each class.

Once subjects had reached a stable asymptotic level of performance on the baseline training songs, we lowered the rate of reinforcement and shortly thereafter began the test sessions with the motif-probe songs (methods). Each motif-probe stimulus was a random sequence of 1–12 motifs drawn from the baseline training songs with the constraint that motifs from different singers never combined into the same motif-probe stimulus. Thus, the individual identity of the original singer was preserved in the large set of motif-probe stimuli.

In general, the more motifs that appeared in each test stimulus, the easier it was for each subject to recognize the correct singer (Fig. 2). Although there was considerable variability across subjects,
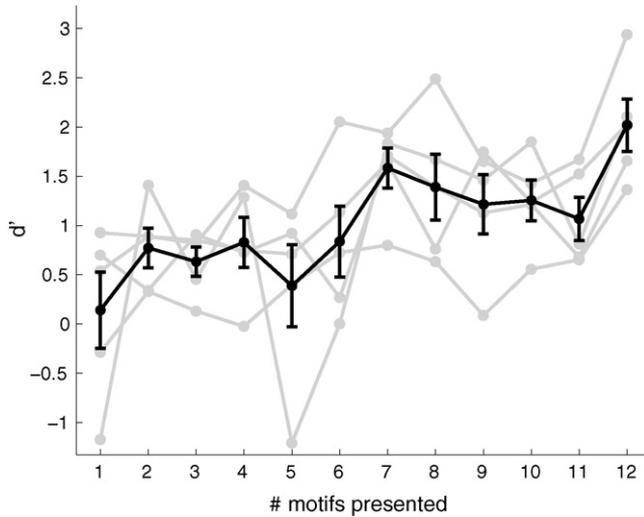
Fig. 3. Mean performance on motif-probes. Mean d' for individuals (light gray lines) and averaged across subjects (black line, ±sem) as a function of the number of motifs in a test song.

increasing the number of motifs in the test stimulus from 1 to 12 had a significant impact on average recognition accuracy ($F_{(4, 11)} = 5.39$, $p < 0.0001$, rmANOVA with all probe classes, Fig. 3). With only a single motif, mean performance ($d' = 0.14 \pm 0.39$) was not above chance ($t = 0.36$, $df = 4$, NS, t-test). With as few as two randomly chosen motifs, mean recognition performance was still modest ($d' = 0.77 \pm 0.20$), but nonetheless significantly above chance ($t = 3.82$, $df = 4$, $p = 0.018$). For 12-motif sequences, recognition performance, average across subjects, was very high ($d' = 2.02 \pm 0.27$, Fig. 3) and was significantly above chance ($t = 7.59$, $df = 4$, $p = 0.0016$). Even when the probe stimulus contained 12 motifs, however, mean recognition was still significantly lower than that for the baseline songs presented during the same sessions (d' for baseline stimuli = 2.86 ± 0.26; $t = 2.98$, $df = 4$, $p = 0.04$ paired t-test), and the mean normalized d' (a ratio of probe:baseline performance) was significantly less than 1 (ratio = 0.69 ± 0.08; $t = 3.765$, $df = 4$, $p = 0.019$).

Although the increase in performance with more motifs was roughly monotonic, it is not linear. The mean response curve (Fig. 3) shows abrupt steps between 1- and 2-motif sequences, another between 6 and 7-motif sequences, where average performance moves well-above chance, and a third large shift between performance on 11- and 12-motif sequences. Examining the data more closely confirms these intuitions. Among the motif-probes containing 2–6 motifs, there is no significant effect of increasing the number of motifs in a given test stimulus ($F_{(4, 4)} = 0.50$, NS). Likewise, among motif-probes containing 7–11 motifs, there is no significant effect of increasing the number of motifs in a given exemplar ($F_{(4, 4)} = 1.42$, NS). Thus, most of the improvement in average performance across the range of motif-probes comes from the increases between 1–2, 6–7, and 11–12 motifs. We note, however, that there is significant variability across subjects (Fig. 2).

After responding to a sufficient number of motif-probe stimuli from each of the 12 classes, we returned subjects to the baseline training songs only. We confirmed their stable performance on the baseline songs over several sessions, and then began the test sessions with the excerpt-probe songs (see methods). Recall, that each excerpt-probe stimulus was a random, 0.1–12.0 s continuous sub-section of song drawn from the baseline training songs. We measured performance as the ability to maintain accurate recognition of the individual identity of the singer from whose song the excerpt was drawn. Due to a programming error one subject produced no useable data.

As with the motif-probes, the longer the duration of the song excerpt in each probe stimulus, the easier it was for subjects to recognize the correct singer (Fig. 4). To analyze the results of these probe sessions, we collapsed the different length stimuli into 12 1-s bins, ranging from 0.01–1.00 to
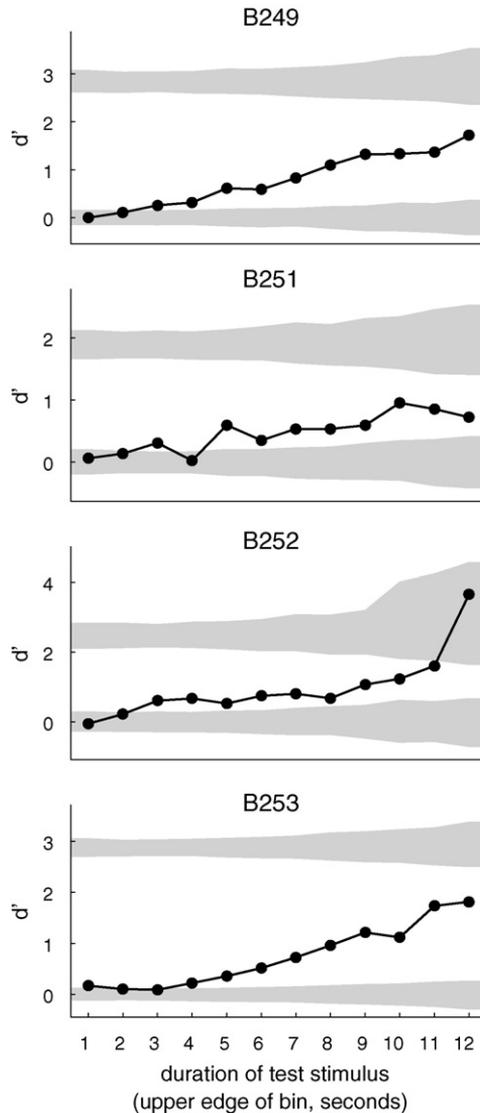
Fig. 4. Individual data for excerpt-probes. Each panel shows the behavior of an individual subject on excerpt-probe test trials. Circles indicate performance (d') as a function of the duration of the excerpt-probe, with confidence intervals and labels as in Fig. 2. Values on the x-axis are the upper edge of the 1-s test stimulus bin, e.g. '3' represents test stimuli 2.001–3.000 s in duration.

11.01–12.00 s (methods) and computed the d' value (for each bird) for all the stimuli that fell into each bin. Initially, we ignored any potential effects due to differences in the offset of a given excerpt from the start of the source song by collapsing across all excerpts of a given length. Increasing the duration of the test stimulus from 1 to 12 s had a significant impact on recognition accuracy ($F_{(3, 11)} = 12.07$, $p < 0.0001$, rmANOVA with all probe classes; Fig. 5). As shown in Fig. 5, the overall change in performance with increasing excerpt duration was monotonic and smooth. When presented exemplars 0.1–1.0 s long, recognition accuracy ($d' = 0.04 \pm 0.06$) was not significantly above chance ($t = 0.76$, $df = 3$, NS). Lengthening the probe stimuli slightly, to between 1 and 2 s, produced a rise in performance ($d' = 0.13 \pm 0.02$) to a level that was still modest, but significantly greater than chance ($t = 3.68$, $df = 3$,
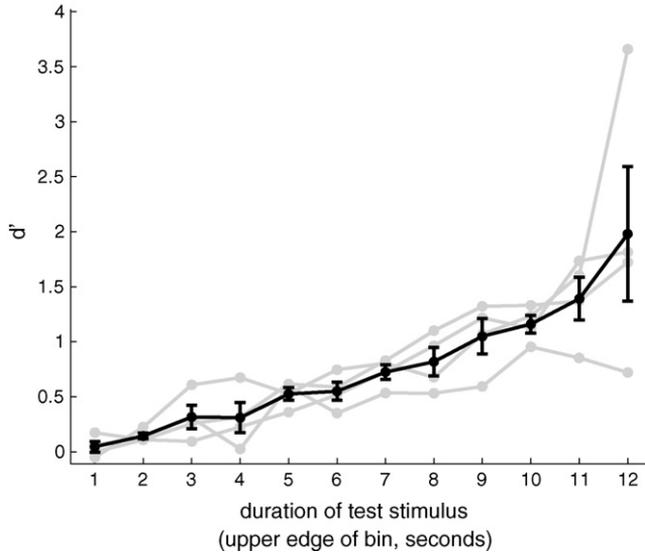
Fig. 5. Mean performance for excerpt-probes. Mean d' for individuals (light gray lines) and averaged across subjects (black line, ±sem) as a function of the duration of the test stimulus. The x-axis is labeled as in Fig. 4.

$p = 0.034$). For excerpt-probe stimuli 11–12 s long, mean performance ($d' = 1.42 \pm 0.37$) was significantly above chance ($t = 3.79$, $df = 3$, $p = 0.032$), but still significantly below baseline performance (d' for baseline $= 2.61 \pm 0.23$; $t = 4.711$, $df = 3$, p = 0.018, paired t-test) and the mean normalized d' was significantly less than 1 (ratio $= 0.54 \pm 0.13$; $t = 3.660$, $df = 3$, $p = 0.035$).

In addition to examining the effect of stimulus duration on recognition performance, the excerpt-probe stimuli also provide a means for examining how different regions of the training stimuli may contribute, on average, to recognition performance. For example it could be that either the initial or terminal portions of songs hold particular relevance in guiding performance. To test for such effects, we examined performance on 1-, 2- and 3-s excerpt-probe stimuli as function of each probe's offset relative to start of the source training song. As probe duration increases, the range of possible offsets
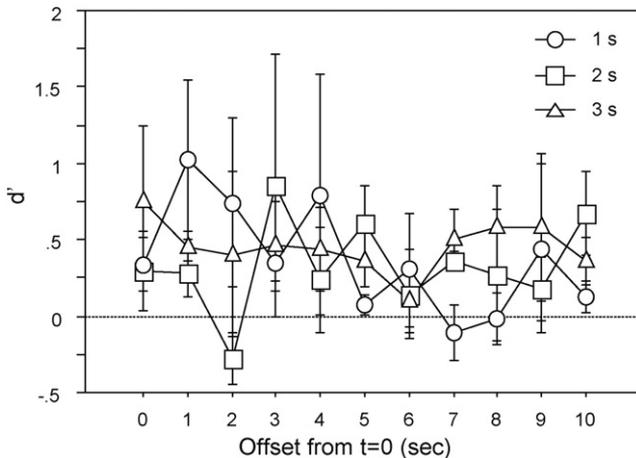

Fig. 6. Stimulus offset. Mean (±sem) d', across subjects, plotted as a function of test song offset (in seconds) relative to the start of the stimulus. Data for excerpt-probe stimuli of 1, 2 and 3 s are plotted separately.
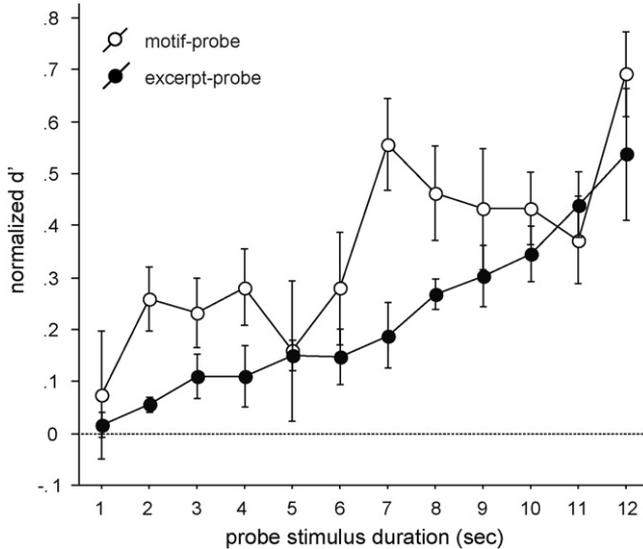
Fig. 7. Comparison performance for Experiment 1. Mean (±sem) normalized d' for motif-probe (open circles) and excerpt-probe (filled circles) stimuli plotted as a function of the approximate duration of each test stimulus (see text).

decreases. Thus, we restrict our analysis to these three durations because responses to them can be observed over a wide range of offsets (0–10 s). We observed no effect of relative offset for any of the three excerpt-probe durations ($F_{(3, 10)} = 0.77$, $F_{(3, 10)} = 0.27$, $F_{(3, 10)} = 0.91$, for 1, 2, and 3-s excerpt-probes, respectively, $p > 0.5$ all cases). Fig. 6 shows the similarity in performance across all of these probe conditions, and in the mean performance across subjects.

It is also helpful to consider performance in the motif- and excerpt-probe conditions together. Rigorous comparisons between specific stimuli in these conditions is hampered somewhat by differences in how the two types of probe stimuli are binned for analysis. On average, however, motifs are roughly 1 s long and so direct comparisons across excerpt-probe time-bins and numbers of motifs is reasonable (Fig. 7). Comparing between the two probe conditions in this way, we observe roughly similar patterns of response for similar duration test stimuli. Overall, subjects' tended to perform slightly better on the motif-probe stimuli than excerpt-probes of comparable length, but the advantage was not statistically significant ($F_{(33, 11)} = 1.32$, NS, interaction of probe length with condition).

Conclusions

The results of Experiment 1 confirm the intuition that as more vocal material is added to the test songs, in both conditions, recognition performance improves. With very little vocal material, either 1 motif or less than 1 s of song, performance is at chance levels. With 12 motifs or 12 s of song, performance is markedly better and well-above chance. On average, performance over the range of intermediate stimuli, from small to large amounts of song material, showed a monotonic increase in performance correlated with the amount of song in each test stimulus. The transition from short to long excerpts was much smoother, however, than the transition from few to many motifs. Although not statistically significant, it was somewhat surprising to observe that performance on the motif-probe stimuli tended to slightly lead performance on the excerpt-probe stimuli. Given that the excerpt-probes preserve the relative temporal relationships between motifs that subjects experience during baseline training, one might predict performance on these stimuli to lead that on the motif-probes where all these relative temporal relationships are abolished. The poorer performance may reflect the fact that excerpts could violate the natural start and stop points for single motifs, thereby rendering the first and last motifs in each excerpt more difficult to recognize.

Most subjects were generally above chance even for very small sets of motifs (2–6 motifs) but only marginally so. The inter-subject variability observed in the pattern of responding to the motif-probes may reflect different solution strategies and/or a reliance on different sized sets of motif for recognition. Thus, for any given bird, estimating the minimum number of motifs necessary for accurate recognition is difficult. We can say, however, that above chance recognition is possible in principle, and on average, with as few as two randomly selected motifs. The most reliable gains in recognition (across subjects) do not come until more than seven motifs are strung together. It is not clear what significant change in information is achieved when more than seven motifs are selected. Presumably this threshold is related to the overall size of (i.e. number of motifs in) the training sets, and the corresponding subset of motifs that subjects use to solve the task. Whatever the underlying cause, the non-linear nature of results from motif-probes contradicts a strategy where subjects learn only a single "diagnostic" motif (or sub-motif acoustic feature). Such a strategy would predict that as the number of motifs in each probe sequence increased, the likelihood of choosing the diagnostic motif would increase linearly. Similar arguments apply to larger sets of diagnostic motifs, although the probability of selecting all such motifs by chance decreases exponentially as the number of putatively diagnostic motifs increases. Based on these results, we conclude that the random selection of motifs from training songs is a reasonable proxy, but nonetheless incomplete dimension, along which to model the features that control individual song recognition.

Consistent with the foregoing ideas, there appear to be no strong effects tied to the position of song material within the source stimulus. Excerpt-probe stimuli drawn from the start or middle of a training song were just as recognizable as those drawn from the terminal portions. We note, however, that variance was high in the offset-probe tests, and that this may obscure more subtle effects tied to the position of exemplars within a stimulus. Likewise, individual birds may possess idiosyncratic strategies tied to specific training exemplars that are washed out in the kinds of averaging required to address the role of offset in the present case.

Overall, the results support the conclusion that starlings extract numerous cues to individual vocal recognition from each training song, that these cues are distributed over temporally broad regions of each song, and that adding more cues improves recognition in a graded fashion. No single feature/motif or set of features/motifs can be seen as both necessary and sufficient for individual vocal recognition in starlings.

Experiment 2

In Experiment 2, we trained birds to classify conspecific song and then tested their classification accuracy on test stimuli that had portions of their acoustic material replaced by continuous regions of white noise. This is a complementary manipulation to that of Experiment 1, and we were therefore able to assess different aspects of the birds' classification of conspecific song based on degraded versions of training stimuli. Whereas Experiment 1 presented test material that could start and stop at any region of the training stimulus, and could be out of order with respect to the training stimuli in the case of the motif-probe stimuli, the test stimuli in Experiment 2 maintained the relative and absolute timing information present in the training stimuli while still limiting the amount of acoustic material available to the bird.

Methods

Subjects

Five starlings participated in Experiment 2. At the start of training, all birds were naïve to the operant apparatus and the stimuli used in the experiment.

Baseline training stimuli

For each bird, two roughly 10-s segments of starling song from one singer were assigned to the left response port and two segments of song from a different singer were assigned to the right response port. The specific songs used during training and their left or right key assignments were varied across subjects. In total, eight different stimuli were used across the five birds tested (duration, mean ± sd:

9.860 ± 0.657 s]. We calculated d' over 200-trial blocks to assess performance. Birds were maintained on the baseline training task until completing at least three consecutive 200-trial blocks with d' significantly above chance (see general methods).

Testing procedure and stimuli

Once each bird reached behavioral criterion on the baseline training, we decreased the rate of reinforcement to 40%, so that on 60% of the trials, there was no consequence for an operant response. We then began introducing test stimuli. During testing, each trial initiated by the bird had a 50% chance of being a baseline trial and a 50% chance of being a test trial, here referred to as a "noise-probe" stimulus. Baseline trial stimuli were the same as the training stimuli, and had the same behavioral contingencies. For noise-probe stimuli, a baseline stimulus was chosen and then a random offset was selected uniformly from the beginning of the stimulus at a 1-ms resolution (range: 1–9858 ms, mean 3163 ms, sd 2230 ms). Starting at this offset, a random duration was then selected uniformly from the remaining portion of the stimulus, again at a 1-ms resolution (range: 2–9586 ms, mean 3112 ms, sd 2232 ms). The region of song beginning at the selected offset and lasting for the selected duration was then replaced with broadband white noise. Note that the range of the possible durations are constrained by the particular offset selected for a given trial, so that noise-probe stimuli that had later offsets necessarily had fewer possible duration values that could be selected. The mean amplitude of the white noise was the same in each noise-probe so that the noise itself carried no information about the identity of the acoustic information it was replacing. Noise-probe stimuli were reinforced non-differentially, meaning that the operant response of the bird had an equal chance of eliciting a reward or a punishment (see general methods).

For analysis, we binned the noise-probe stimuli into 9 1-s bins: 0.000–1.000 s, 1.001–2.000 s, 2.001–3.000 s, etc., and computed the corresponding d' for stimuli in each bin as described in Experiment 1.

Results

Subjects learned to accurately recognize and classify the four baseline training songs ultimately reaching a high level of performance. As in Experiment 1, recognition performance improved significantly over the course of learning ($F_{(4, 43)} = 3.699$, $p < 0.0001$, rmANOVA, main effect of training). The mean (±sem) d' immediately prior to the start of the first test session was 2.62 (±0.45), demonstrating accurate recognition of the songs from each singer.

Responses to the noise-probe stimuli showed a pattern consistent with the results of Experiment 1. Recognition performance for all birds became degraded as the duration of white noise in each noise-probe stimulus increased (Fig. 8). To assess this change quantitatively, we collapsed test stimuli into nine bins based on the duration of noise in each (methods), and computed the d' value for all of the test stimuli in each bin (Fig. 9). Increasing the duration of the white noise in each probe stimulus (thereby decreasing the amount of song) significantly affected the birds' classification performance across these nine bins ($F_{(4, 8)} = 14.935$, $p < 0.0001$, rmANOVA main effect of noise duration).

Across all birds, mean performance on noise-probes was indistinguishable from baseline when noise was substituted for 0–1 s of song (d' for noise-probes = 2.20 ± 0.31; d' for baseline = 2.20 ± 0.26; $t = 0.047$, $df = 4$, $p = 0.517$, paired t-test). For all the other noise-probe stimuli, with more than 1 s of noise substituted for songs, recognition was significantly below baseline performance ($p < 0.05$, $df = 4$, all cases, paired t-test), and decreased monotonically with increasing noise durations (Fig. 9).

Although performance fell below baseline values when more than 1-s of noise was substituted for song, recognition accuracy remained above chance for all durations of noise examined. Even when presented with test stimuli that included 8–9 s of noise, the most severe stimulus degradation tested, birds were able to perform the classification task significantly above chance levels ($d' = 0.783 ± 0.26$; $t = 3.075$, $df = 4$, $p = 0.019$). Thus, birds could still recognize the correct singer of a song when presented with as little as 9–11% of a baseline stimulus, on average.

The results are consistent with those of Experiment 1 where increasing the total amount of song available to the bird improved recognition. To compare the results of Experiments 1 and 2 directly, we classified excerpt-probe and noise-probe stimuli as the proportion of song remaining (relative to the
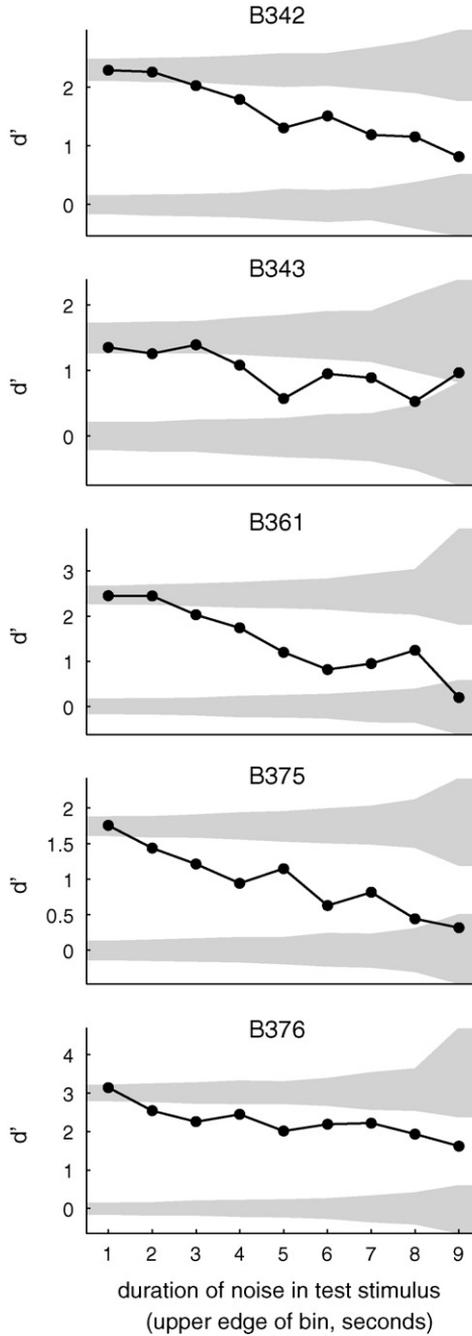
Fig. 8. Individual data for noise-probes. Each panel shows the behavior of an individual subject on noise-probe test trials. Circles indicate performance (d') as a function of the duration of the noise in each probe stimulus, with confidence intervals and labels as in Fig. 2. Values on the x-axis are the upper edge of the 1-s test stimulus bin (e.g. '3' represents test stimuli with noise 2.001–3.000 s in duration).
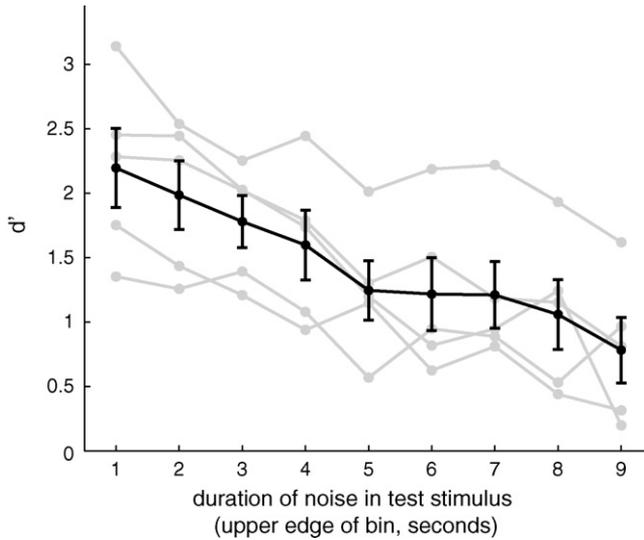
Fig. 9. Mean performance for noise-probes. Mean d' as in Fig. 3, with x-axis labeled as in Fig. 8.

original) in a given test stimulus, binned the probes into 10 equal-proportion bins, and calculated a d' value for each bin. Because of the way the stimuli were constructed, we had insufficient data to estimate performance for bins with a proportion of 0.8–1.0 for the excerpt-probes and 0.9–1.0 for noise-probes. The results of this comparison are shown in Fig. 10. Both conditions showed similar improvement in recognition over comparable ranges of probe stimuli, but recognition of the noise-probe stimuli was clearly better than that of the excerpt-probes. To assess the rate at which behavior improved as a function of available training song, we fit lines to both datasets. The results of the linear regression for
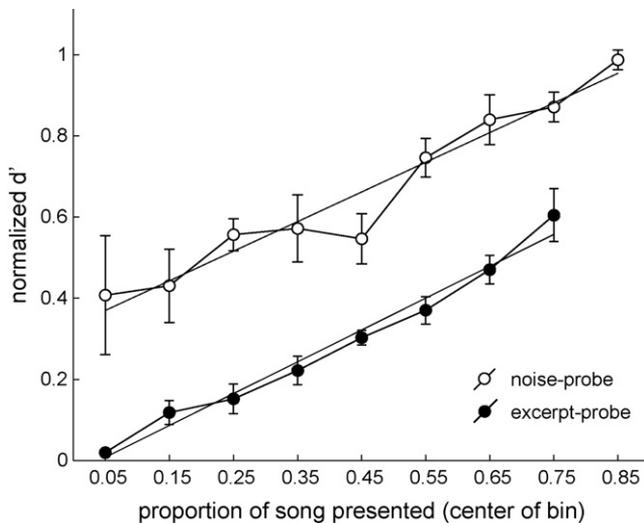

Fig. 10. Comparison performance on excerpt-probe and noise-probe test stimuli. Mean ($\pm$sem) normalized d' for excerpt-probe (solid line, filled circles) and noise-probe (dashed line, open circles) stimuli plotted against the proportion of baseline song remaining in the test stimulus. Thin lines show the linear regression for data in each condition. Excerpt-probe: $y = 0.79 \times x - 0.03$, $R^2 = 0.98$; noise-probe: $y = 0.73 \times x + 0.33$, $R^2 = 0.94$.

the excerpt-probe stimuli (Y-intercept = −0.031, slope = 0.785, R^2 = 0.980) and the noise-probe stimuli (Y-intercept = 0.333, slope = 0.730, R^2 = 0.944), produced a very similar slope, suggesting that the gain in performance seen with increasing baseline stimulus material in Experiment 1 was similar to the gain seen in Experiment 2. The difference in the Y-intercept values indicates that recognition of the excerpt-probe stimuli in Experiment 1 was, overall, more difficult than recognition of the probe stimuli in Experiment 2.

Conclusions

The results of Experiment 2 demonstrate that substituting noise for continuous regions of a familiar song can systematically impair song recognition. Recognition performance on noise-probe stimuli is indistinguishable from baseline levels until more than 1 s of the song is masked, after which performance decreases monotonically with increasing noise duration. Importantly, however, accurate classification performance was maintained at all noise durations tested. The ability of birds to correctly recognize the individual singer based on as little as 9–11% of a learned song segment supports the conclusion that the representation of a given conspecific song is distributed across a large portion of the acoustic material available to the bird, and not simply in a few salient auditory features.

In isolation, the results of Experiment 2 could be explained by a generalization decrement. That is, substituting noise for increasingly larger portions of a training song may simply render each probe stimulus more dissimilar to its noise-free version. This dissimilarity, rather than impaired recognition, may account for the observed change in performance. We take up this question in the general discussion and argue that recognition provides a more parsimonious explanation consistent with all of the present results and those from previous studies of song perception in starlings.

Although not directly assessed in the same subjects, the normalized performance values were higher on average in Experiment 2 than in Experiment 1, even when accounting for the level of performance on the training stimuli. This difference may be tied to the availability of absolute timing information present in the noise-probe stimuli, but absent in the motif-probe and excerpt-probe stimuli. It is also possible that differences in performance across the two experiments reflect disparities in the training sets used for each condition. There was much more material for each bird to recognize in Experiment 1 than in Experiment 2 (Experiment 1: 16 songs per subject, mean duration 14.97 s; Experiment 2: 4 songs per subject, mean duration 9.86 s). Dissociating these potential effects requires additional experimentation.

General discussion

The results of Experiments 1 and 2 provide converging evidence in support of the conclusion that knowledge of the songs acquired during individual vocal recognition is distributed across significant portions of conspecific song. Recognition is not restricted to a single motif, or a single region of song. Furthermore, recognition is redundant in that no single acoustic feature of song is necessary, and multiple features (or motifs or regions) are sufficient to enable accurate recognition.

Observations from the present studies provide much greater detail in our understanding of the features that guide song recognition. Across both probe conditions in Experiment 1, performance was consistently below baseline recognition. Even when provided with very large amounts of material, either 12 motifs in the first probe condition or 12 s of song in the second, performance never reached parity with the baseline accuracy. Subjects always performed better on the baseline songs than on the probe stimuli. In both probe conditions, the maximum number of motifs or test stimulus duration never reached the average number of motif/duration for the baseline songs. Baseline training songs had, on average, about 14 motifs and lasted about 15 s. Therefore, we cannot rule out the possibility that adding more motifs (or increasing the duration) of the test stimuli might have brought performance in line with that seen for the baseline. One would expect this to be the case for increasing the duration of the excerpt-probe stimuli. However, some deficit in recognition for motif-probe stimuli, even for long runs of randomly sequenced motifs, is consistent with previous work showing that temporal patterning of motifs in the training songs can affect recognition (Gentner, 2008; Gentner & Hulse, 1998).

Multiple lines of converging evidence support the conclusion that motifs function as an important natural perceptual unit for song recognition in starlings. Synthetic songs that follow the natural motif sequence patterns of each individual are recognized without discernable impairment compared to natural songs (Gentner & Hulse, 1998). Recognition performance of starlings tested with chimeric songs composed of differing proportions of familiar motifs is predicted by the ratio of motifs from each singer (Gentner & Hulse, 2000). Abolishing the sub-motif structure in learned songs leads to significant deficits in song recognition, and sensitivity to motif-level organization emerges without explicit reinforcement (Gentner, 2008). Moreover, familiar motifs (presented in isolation) elicit robust extracellular responses from single auditory neurons in the starling forebrain (Gentner & Margoliash, 2003; Jeanne, Sharpee, & Gentner, 2008; Thompson & Gentner, 2008). Finally, the vocal production of single motifs is highly stereotyped across multiple iterations by the same singer, the natural repetition structure of song operates at the motif level, and where song sharing does occur, it is motifs that are shared. To this long list we can add the results of the present study showing that on average, presenting a randomly chosen, non-sequential, pair of familiar motifs is sufficient for recognition at above chance levels.

In all cases tested, full recognition performance is only maintained when motif structure is intact, but motifs are not the only temporal scale for song perception. Previous results suggest that starlings are also sensitive to features at temporal scales below the motif level (Gentner, 2008) and at temporal scales that span multiple motifs. Starlings can more easily recognize strings of familiar motifs that follow familiar motif transitions than those that have random transitions between motifs (Gentner & Hulse, 1998). Starlings can also learn to recognize patterns defined over disjoint sets of motifs, e.g. the patterns $a_i b_k a_j b_l$ and $a_i b_k b_l a_j$, where $a_{i,j}$ are elements of one set of familiar motifs and $b_{k,l}$ are elements of another set (Gentner, Fenn, Margoliash, & Nusbaum, 2006). Together, the data indicate that different tasks can focus attention on the song in differing ways. The fact that songs carry salient information at temporal scales longer and shorter than that of the motif should not be surprising. The same is true in humans. Human listeners trained with speech synthesizers to recognize either words or sentences show no generalization across context (Greenspan, Nusbaum, & Pisoni, 1988). Likewise, humans trained under specific conditions to extract relevant speech cues from sentences of natural speech generalize identification only to words within sentences, not to isolated words (Nygaard & Pisoni, 1998).

Our discussion of starlings' behavior in the kind of operant tasks presented here turns on the notion that birds learn to recognize sets of songs (or sets of motifs composing songs) associated with each response. This follows from the well-documented ability of individual vocal recognition in all songbird species studied (Stoddard, 1996). It is important, however, to consider simpler explanations that may fit the pattern of data. The results from Experiment 2, in particular, can be interpreted in terms of a generalization decrement, whereby performance falls as probe stimuli become increasingly dissimilar to the training stimuli when more noise is added. Several lines of evidence contradict the generalization decrement hypothesis. Following nearly identical baseline training, starlings can continue to classify novel songs very accurately, i.e. without any detectable response decrement, so long as the novel songs contain at least some number of familiar motifs. This is true even when the novel songs represent substantial permutations of the training songs, either because familiar motifs from multiple trainings songs are combined to make the test songs (Gentner & Hulse, 1998, 2000), or because substantial portions of the motifs in the test songs are completely unfamiliar (Gentner & Hulse, 1998; Gentner et al., 2000). A similar maintenance of recognition in the face of training song permutation can be seen in the results of Experiment 1, where probe songs are composed of motifs drawn (randomly) from multiple training songs and presented in a random order. In this case, similarity to any single training song is low, yet accuracy steadily improves as more motifs are added to the probe songs. Given that the baseline training for Experiments 1 and 2 was very similar, we contend that recognition, not stimulus generalization, is the most parsimonious description of the behavior during all the probe sessions described here (and in our earlier studies).

One interesting conclusion based on the present results is that part of the way singer identity is represented in songs is tied to the relative and absolute timing of motifs (or other acoustic features) within the song. The improvement in recognition for Experiment 2 over the excerpt-probes from Experiment 1 is consistent with this idea. Results demonstrating that star-

lings are sensitive to both relative and absolute motif position cues (unpublished) support this hypothesis.

Having a recognition system sensitive to distributed information, such as that implied by the current results would be adaptive in several ways. Under natural conditions, a singer's vocalizations may be degraded in multiple ways before reaching the receiver. The kind of recognition system described here would be resilient to masking by environmental sounds or perhaps even other singers, and to song bout interruptions on the part of the singer and partial reception by the receiver. The general finding that starlings need only a relatively small portion of acoustic material, taken from any number of points in a bout, to correctly recognize singers points to an auditory system that is adapted to these sorts of environmental challenges.

The finding that recognition increases (or decreases) gradually with the addition (removal) of vocal material in songs means that, at least under these kinds of training conditions, behavior cannot be considered as a dichotomous (either/or) dependent measure of song, but rather a more probabilistic judgment of singer identity. This may reflect the fact that vocal identity itself, as coded in the acoustics of the song, is more probabilistic than discrete, or that the underlying decision mechanisms that ultimately transform information coded in song to behavior are themselves probabilistic. Answering these kinds of questions is not a simple endeavor. It requires precise control over singer identity, song acoustics, recognition behavior, and access to dependent measures at multiple points in the neural processing chain. The latter requirement points to one of the great strengths of the songbird system in general, and of vocal recognition in starlings, more specifically, as this species provides one of the few organisms in which all these measure are feasible. The work described here sets the stage for further manipulations that will directly control recognition behavior under more invasive physiological procedures, such as awake, chronic electrophysiology (Knudsen & Gentner, 2009).

References

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

Bricker, P. D., & Pruzansky, S. (1976). Speaker recognition. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 295–326). New York: Academic Press.

Cynx, J., & Nottebohm, F. (1992). Testosterone facilitates some conspecific song discriminations in castrated zebra finches (*Taeniopygia guttata*). *Proclamations of the National Academy of Sciences USA, 89*, 1376–1378.

Dooling, R., Brown, S., Klump, G., & Okanoya, K. (1992). Auditory perception of conspecific and heterospecific vocalizations in birds: Evidence for special processes. *Journal of Comparative Psychology, 106*, 20–28.

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

Eens, M., Pinxten, R., & Verheyen, R. F. (1991). Organization of song in the European starling – Species-specificity and individual-differences. *Belgian Journal of Zoology, 121*, 257–278.

Gentner, T. Q. (2008). Temporal scales of auditory objects underlying birdsong vocal recognition. *Journal of The Acoustical Society of America, 124*, 1350–1359.

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

Gentner, T. Q., & Hulse, S. H. (1998). Perceptual mechanisms for individual vocal recognition in European starlings, *Sturnus vulgaris*. *Animal Behaviour, 56*, 579–594.

Gentner, T. Q., & Hulse, S. H. (2000). Perceptual classification based on the component structure of song in European starlings. *Journal of the Acoustical Society of America, 107*, 3369–3381.

Gentner, T. Q., Hulse, S. H., Bentley, G. E., & 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. *Journal of Neurobiology, 42*, 117–133.

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

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 421–433.

Heinz, R. D., Sinnott, J. M., & Sachs, M. B. (1980). Auditory intensity discrimination in blackbirds and pigeons. *Journal of Comparative and Physiological Psychology, 94*, 993–1002.

Hulse, S. H. (1995). The discrimination-transfer procedure for studying auditory perception and perceptual invariance in animals. In G. M. Klump, R. J. Dooling, R. R. Fay, & W. C. Stebbins (Eds.), *Methods in comparative psychoacoustics* (pp. 319–330). Basel, Switzerland: Birkhäuser Verlag.

Jeanne, J. M., Sharpee, T. O., & Gentner, T. Q. (2008). Song recognition learning drives experience-dependant responses in the auditory forebrain region CLM. *Society of Neuroscience Abstracts, 34*.

Knudsen, D., & Gentner, T. Q. (2009). Behavioral state modulation of responses to conspecific song in European starling auditory forebrain. *Society of Neuroscience Abstracts, 35*.

Nelson, D. A. (1988). Feature weighting in species song recognition by the field sparrow (*Spizella pusilla*). *Behaviour, 106*, 158–181.

Nelson, D. A. (1989). Song frequency as a cue for recognition of species and individuals in the field sparrow (*Spizella pusilla*). *Journal of Comparative Psychology, 103*, 171–176.

Nelson, D. A., & Marler, P. (1990). The perception of birdsong and an ecological concept of signal space. In W. C. Stebbins, & M. A. Berkely (Eds.), *Comaprative perception: Complex signals* (pp. 443–477). New York: John Wiley and Sons.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60*, 355–376.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 651–666.

Stoddard, P. K. (1996). Vocal recognition of neighbors by territorial passerines. In D. E. Kroodsma, & E. H. Miller (Eds.), *Ecology and evolution of acoustic communication in birds* (pp. 356–374). Ithaca: Comstock/Cornell.

Sturdy, C. B., Phillmore, L. S., & Weisman, R. G. (2000). Call-note discriminations in black-capped chickadees (*Poecile atricapillus*). *Journal of Comparative Psychology, 114*, 357–364.

Thompson, J. V., & Gentner, T. Q. (2008). Song recognition learning elicits differential suppression of neural responses to learned sons in NCM. Paper presented at the Society for Neuroscience Abstracts.