Context-dependent categorical perception in a songbird

Robert F. Lachlan[a,1,2] and Stephen Nowicki[a,b]

[a]Department of Biology and [b]Department of Neurobiology, Duke University Medical School, Duke University, Durham, NC 27708

Edited by W. Tecumseh Fitch, University of Vienna, Vienna, Austria, and accepted by the Editorial Board December 3, 2014 (received for review June 12, 2014)

Some of the psychological abilities that underlie human speech are shared with other species. One hallmark of speech is that linguistic context affects both how speech sounds are categorized into phonemes, and how different versions of phonemes are produced. We here confirm earlier findings that swamp sparrows categorically perceive the notes that constitute their learned songs and then investigate how categorical boundaries differ according to context. We clustered notes according to their acoustic structure, and found statistical evidence for clustering into 10 population-wide note types. Examining how three related types were perceived, we found, in both discrimination and labeling tests, that an "intermediate" note type is categorized with a "short" type when it occurs at the beginning of a song syllable, but with a "long" type at the end of a syllable. In sum, three produced note-type clusters appear to be underlain by two perceived categories. Thus, in birdsong, as in human speech, categorical perception is context-dependent, and as is the case for human phonology, there is a complex relationship between underlying categorical representations and surface forms. Our results therefore suggest that complex phonology can evolve even in the absence of rich linguistic components, like syntax and semantics.

bird song | categorical perception | phonology | speech perception

Although language as a whole is unique to humans, some of its components, particularly in the domains of phonology and phonetics, are shared with other species (1–4). This sharing has often been demonstrated in tests of animals' perception of human speech, but occasionally an animal's own communication system may reveal speech-like traits. One notable case is bird song, which is learned by imitation and the development of which shares several other features with speech (1, 2, 5). Because songs are constructed hierarchically from smaller units (Fig. 1*A*), birds may possess a "phonology" similar to that of human language (2–4).

In speech, words and other linguistic units are composed from smaller units called phonemes, which are themselves categories, each encompassing a range of acoustic variants. Phonemes typically exhibit a degree of categorical perception: that is, individuals label continuously varying stimuli as belonging to discrete categories with abrupt boundaries between them, and discriminate stimuli that span these boundaries more readily than stimuli within one category. Categorical boundaries between phonemes are learned early in life and shared within a speech community (6, 7). It has become clear, however, that entirely categorical perception (where individuals can perceive variation only between categories) rarely, if ever, applies in speech. One key departure from the ideal is that categorical perceptual boundaries between phonemes vary with linguistic context (8–10), such as the position of the sound (11, 12).

Categorization is essential for the linguistic functions of speech, allowing categorical distinctions between discrete words and forming the basis of phonology. As with perception, phonology is also highly context-sensitive: different variants of phonemes are used predictably in different contexts (13). Also as with perception, context-sensitivity is considered a central and ubiquitous facet of phonology, which provides links to higher aspects of language (9, 13). The two phenomena intersect in "partial phonemic overlapping" (14), in which identical sounds are used in different contexts and are perceived as different phonemes. In Danish, for example, the sound [d] is perceived as the phoneme /d/ in the initial position of syllables, and /t/ in the final position (15). We here test whether a similar phenomenon occurs in the natural communication system of a songbird.

Categorical perception is now thought to be a ubiquitous feature of perceptual systems, and has been demonstrated in the natural communication systems of several other animal species (16–20). In most cases, however, these have involved nonlearned signals and serve a communicative function by facilitating discrimination between conspecific and heterospecific stimuli. The degree to which natural communication systems of animals are organized like human phonology remains generally unstudied (13), in particular, the degree to which perceptual and phonological categories are sensitive to context. Nevertheless, tests of Japanese quails' perception of human speech suggest a widely shared ability to allow context to influence categorization (21, 22).

Learned bird song provides a particularly fruitful system for exploring similarities with speech, and perhaps the best evidence for a phoneme-like system has been found in a songbird species, the swamp sparrow (*Melospiza georgiana*). Swamp sparrow songs consist of a single repeated syllable of two to five notes (Fig. 1*A* and Fig. S1). Although there may be 60 syllable types in a local population (23), the constituent notes cluster into around 6–10 "note-type" categories, which were originally thought to be species-universal, but which have recently been found to vary in structure and number between populations (24, 25). Swamp

Significance

Song-learning birds share several of the cognitive traits that underlie human speech. Continuous variation in swamp sparrow song notes, for example, is perceived in a categorical manner, similar to human perception of phonemes, the smallest units of human speech. Although speakers and listeners are generally unaware of the fact, many phonemic categories in speech vary in their structure and in how they are perceived, depending on linguistic context. Herein, we test how categorical perception in swamp sparrows is influenced by context and demonstrate that one note type is categorized differently, depending on its position within a song syllable. To our knowledge, our results suggest for the first time that this central characteristic of human phonology is also found in a nonhuman communication system.

Author contributions: R.F.L. and S.N. designed research; R.F.L. performed research; R.F.L. analyzed data; and R.F.L. and S.N. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. W.T.F. is a guest editor invited by the Editorial Board.

See Commentary on page 1658.

[1]To whom correspondence should be addressed. Email: r.f.lachlan@qmul.ac.uk.

[2]Present address: Department of Psychology, School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1410844112/-/DCSupplemental.

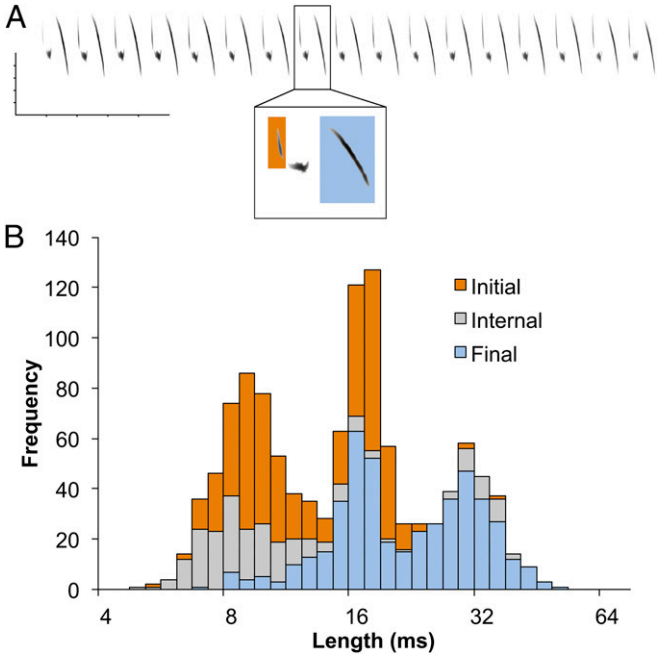

Fig. 1. Swamp sparrow song structure. (*A*) Spectrogram of a swamp sparrow song, demonstrating the typical organization of repeated groups of notes, called "syllables" (frequency scale bar, 5 kHz; time scale bar, 0.5 s); in this song type, the syllable includes three note types. A common class of notes are those that rapidly descend in frequency; in this song there are two such notes, occurring in the initial and final positions within the syllable. (*B*) In a population, there is continuous variation within this class of notes (n = 1,183 notes from 657 syllable from 206 male swamp sparrows), especially in note length (see also Fig. S2). Nevertheless, notes clearly fall into three separate clusters, or note types. These note types are not distributed randomly throughout syllables: long note types are mostly absent from the initial position, and short note types are mostly absent from the final position within syllables.

sparrow note-type categories serve to discriminate syllable types, and thus their perception plays a role in the assessment of songs by receivers: whether they have been precisely learned or whether they follow local population regularities (23). Both field (17, 26) and neurobiological (26) experiments have demonstrated that two of these note types are perceived categorically by other individuals.

In the first such experiment, Nelson and Marler (17) showed that swamp sparrows from a population in the Hudson Valley of New York perceive some notes categorically, based on duration, even though these notes vary continuously in length, with a perceptual boundary coinciding with the trough in a bimodal distribution of note durations (see, for example, Fig. S3*A*). Prather et al. (26) similarly found evidence for categorical perception for the same class of notes in a northwestern Pennsylvania population, but with the perceptual boundary occurring at a longer note length than in New York. This difference is consistent with the idea that note-type categories and categorical perception boundaries culturally evolve and differentiate between populations, much like human speech (27), but a detailed comparison of the distribution of note type structure between Pennsylvania and New York had not been carried out until now.

In this study, we first demonstrate that in the region of acoustic space where there are two note-type clusters in New York, there are three clusters in Pennsylvania. We show that these note types exhibit a phonotactic regularity, in that notes from one cluster generally do not occur at the start of syllables, whereas notes from another cluster are mostly absent at the end of syllables. We then use this distributional data as the basis to test whether categorical perception is context-dependent, examining reduced discrimination within, relative to between categories and category labeling, and explore the similarity between note-type perception and partial overlapping in human phonology and speech perception.

Results

Note-Type Structure in Swamp Sparrow Populations. We began by examining in detail how the difference in perceptual boundaries between the New York and Pennsylvania populations relates to differences in note-type structure. We analyzed the notes sung in the repertoires of 206 males from Pennsylvania and 101 males from New York, measuring frequency parameters and note lengths from spectrograms. Bayesian Gaussian mixture modeling cluster analysis found evidence for 10 note clusters in Pennsylvania and 8 clusters in New York (Fig. S2). In both populations there was a significant clustering tendency (Duda–Hart test, $P < 0.001$ for both populations).

In Pennsylvania, clusters diverged from those found in New York by our analysis in both the number of clusters and in the location of clusters in acoustic space (*SI Results and Discussion* and Fig. S2). Most pertinently, the class of notes that were the focus of previous perceptual experiments in both New York and Pennsylvania (that is, short notes that decrease rapidly in frequency) were assigned, as expected, to two clear clusters in New York (Fig. S3*A*) but to three clusters in Pennsylvania (Fig. S3*B*). Two of the Pennsylvania clusters broadly corresponded to the two types previously investigated in New York; the third cluster was of intermediate length and did not correspond to a cluster in New York. For the remainder of the paper, we focus only on these three note types in Pennsylvania, and we refer to them here as "short," "intermediate," and "long."

We observed a very clear phonotactic structure (28) in the distribution of these three note types (Fig. 1*B*). The note types mostly occurred in either the initial or final position within the syllable (84% of total occurrences of these types; 71% of syllables began and 75% of syllables ended with one of these types). The notes clustering with the short-type occur much more frequently in the initial than in the final position within the syllable [253 vs. 52 times, MCMCglmm (29) logistic regression, n = 173 individuals, $P < 0.0001$], whereas conversely, the notes in the long cluster occur much more frequently in the final than the initial position (258 vs. 11 times, n = 175, $P < 0.0001$). Notes assigned to the intermediate-length cluster, however, are found in both initial and final positions at approximately equal frequencies (191 vs. 175 times, n = 183, $P > 0.4$).

Discrimination of Note-Type Categories. In the previous study of this population (26), a categorical boundary was found between intermediate and long note types, but not between short and intermediate notes, even though these are clearly different note clusters. A possible explanation for this apparent mismatch between production and perception comes from the phonotactic structure found in the use of these note types (Fig. 1*B*). In both earlier field experiments (26), notes were only manipulated in the initial position. Therefore, the observed categorical perception boundary corresponded with the phonotactic distribution: short and intermediate notes are common in the initial position, whereas there are very few long notes. We therefore tested whether this phonotactic rule in song production is reflected in category perception by examining male swamp sparrows' responses to songs in which we had altered note length in either the initial or the final position in the syllable.

We used the habituation/dishabituation technique used in previous studies on swamp sparrows (17, 26), and also used to demonstrate categorical perception in infants (30), to test whether territorial males discriminate between note types. In this protocol, dishabituation in an aggressive visual territorial display ("wing-waving") indicates that males discriminate between two stimuli. We synthesized short, intermediate, and long notes and substituted

them into natural syllables. We played back one such modified stimulus song to a subject until it was habituated and then replaced it with another modified song, and observed whether the rate of wing-waving increased. We quantified the rate of wing-waving by counting the number of 10-s segments during playback in which subjects produced at least one display (maximum value: 18), and used the difference in wing-waving rate for the habituated stimulus and the second stimulus as a "dishabituation score."

We found the pattern of dishabituation to vary greatly depending on whether we substituted notes at the beginning or at the end of a syllable (Fig. 2*A*). Overall, the best-fitting generalized linear mixed-model (GLMM) statistical model was one that included both the length and position of the substituted notes, as well as the interaction between the two factors, and this model provided a significantly better fit to the data than a null model (Table 1). In the initial position of the syllable, we found significantly higher dishabituation scores between intermediate and long types than between short and intermediate types (estimate of difference in the linear function of dishabituation between intermediate vs. long – short vs. intermediate = 1.07, SE = 0.221, $n = 10$ sets of stimuli, $z = 4.82$, $P < 0.001$), replicating the results of Prather et al. (26). As in that study, the increased level of discrimination between intermediate and long notes indicates a categorical perception boundary at this location. Conversely, when we substituted notes in the final position, we found significantly higher dishabituation scores between short and intermediate types than between intermediate and long types (estimate of difference in the linear function of dishabituation between short vs. intermediate − intermediate vs. long = 0.644, SE = 0.191, $n = 10$, z = 3.37, $P < 0.005$) (Fig. 2*A*).

Labeling of Note-Type Categories. In addition to reduced discrimination within categories relative to between categories, a second hallmark of categorical perception is that items are labeled according to their category (31). In the case of categorical perception in animal communication, this often takes the form of reduced response intensity for stimuli lying outside population-typical categories (20). We therefore carried out a second experiment in which we measured the intensity of males' territorial response to songs in which note length had been manipulated.

The best-fitting GLMM model was one that included the interaction between note length and position, and this model provided a significantly better fit to the data than a null model (Table 2). In the initial position, the response to the long notes was significantly weaker than that to the short and intermediate notes (Fig. 2*B* and Table S1). In contrast, in the final position, the response to the short notes was significantly weaker than that to the intermediate and long notes (Table S1).

Syllable-Type Norms and Perception. Earlier findings of categorical perception in swamp sparrows have been interpreted as supporting the idea that perceptual categories align with the produced note-type clusters. An alternative possibility that is also consistent with the earlier studies is that swamp sparrows do not possess general, phoneme-like note-type perceptual categories at all, but instead form separate categories for the notes in different syllable types.

This hypothesis might be plausible because syllable types can be shared by many males within a population (23, 24), a consequence of vocal learning (32). For example, when we visually categorized the 657 syllables sung by the 206 males in our Pennsylvania sample, we found only 65 syllable types, the most common of which was sung by 62 males. Because syllables are clustered into types based on their structural similarity, it is not surprising that for a particular syllable type, each of its notes is typically restricted to only one note type, across all renditions of the syllable by different males in the population. For a given note, we therefore use the term "syllable-type norm" to refer to the most common note type used in that position across all versions of the same syllable type within the population. For example, syllable type "Y" from Pennsylvania (the syllable type illustrated in Fig. 1*A*) was sung by 18 males and the syllable type norm in the initial position was a short note type, and in the terminal position, the norm was a long note type (because all 18 males started the syllable with a short note and ended it with a long note). In fact, only 9.76% of notes in Pennsylvania were

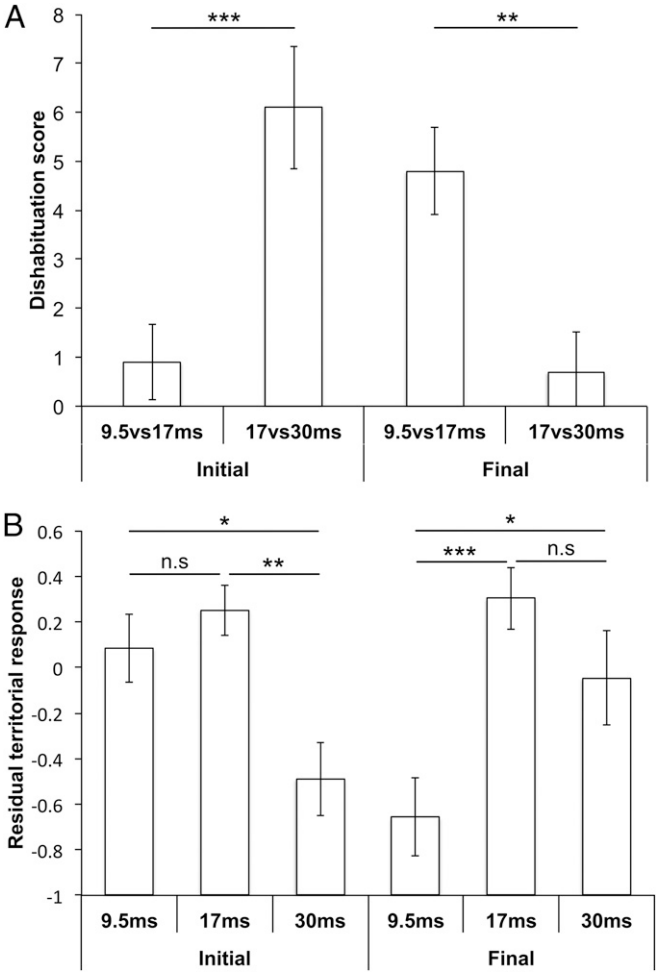

Fig. 2. Category boundaries vary according to positional context. In two experiments we compared male swamp sparrows' responses to songs in which notes had been substituted with synthetic notes of length 9.5 ms, 17 ms, or 30 ms. (*A*) Discrimination experiment. We measured dishabituation in a territorial display after switching stimuli from 9.5 to 17 ms or from 17 to 30 ms (or vice versa). The *y* axis shows the increase in the number of displays given after switching stimuli; large values (maximum possible value: 18) indicate that individuals discriminated between the stimuli ($n = 40$ males). (*B*) Labeling experiment. We played back songs to males inside their territories and used the first principal component of a suite of behavioral measures as an overall measure of the intensity of territorial response. The *y* axis shows the residual of a GLMM model that tested how subject, song-type, and stimulus order variation influenced response intensity. Larger values represent a more intense response. The *x* axis shows the length of the substituted note and the position of the substituted note ($n = 48$ males). Figures show means and error bars representing SEMs. Lines indicate post hoc tests of contrasts for complete GLMM models (see text). *$P < 0.05$; **$P < 0.005$; ***$P < 0.001$ (corrected for multiple comparisons); n.s., not significant. In both experiments, when substitutions were made in the initial position, a categorical response was shown between 9.5 ms and 17 ms but not between 17 ms and 30 ms. When substitutions were made in the final position, a categorical response was shown between 17 ms and 30 ms, but not between 9.5 ms and 17 ms.

Table 1. Summary of GLMM models in discrimination experiment

Model	df	AIC	Log-likelihood	χ^2 test	P (>χ^2)	P (bootstrap)
Model 1: Null (*Materials and Methods*)	2	80.5	−36.27	—	—	—
Model 2: Model 1 + Contrast	3	82.1	−36.04	0.470	0.493	0.528
Model 3: Model 2 + Position	4	84.1	−36.03	0.013	0.909	0.917
Model 4: Model 3 + Contrast x Position	**5**	**59.6**	**−22.79**	**26.5**	**0.0000003**	**<0.0001**
Model 5: Model 4 + Norm + Contrast x Norm	9	60.8	−19.4	6.73	0.151	0.184

The models are nested and increase in complexity from top to bottom row. The χ^2 tests relate to comparisons with the next simplest model in the table. Models were also compared using parametric bootstrapping. The best supported model according to the AIC criterion is in boldface. The sample size was 40 trials divided between 4 combinations of Contrast (either short vs. intermediate or intermediate vs. long) and Position (either initial or final), and with three note-type norms (short, intermediate, long). x represents an interaction term.

classified by the cluster analysis in a different cluster from the majority of notes in the same position in the same syllable type.

In our experiments, the stimuli that we selected varied in their syllable-type norms for the focal notes (in the initial position the norm was either short or intermediate, in the final position the norm was either intermediate or long). We therefore investigated whether category boundaries differed according to these syllable-type norms. We found no evidence, however, for such a difference in either experiment when we included syllable-type norm (Fig. S4); in fact, the Akaike Information Criterion (AIC) score of the GLMM model increased (by 1.2 in the discrimination experiment and 5.4 in the labeling experiment; see model 5 in Tables 1 and 2). This finding indicates that the syllable-type norm provides no additional explanatory power to explain swamp sparrows' responses to song.

We next reanalyzed the subset of our data in the discrimination experiment for which the syllable-type norm was an intermediate-length note. These notes, before we replaced them, were essentially identical in structure, but varied in their position in the syllable. Thus, if syllable-type norm was the principal factor underlying categorization, then we would predict that there should be no difference in the categorical perception boundaries of these notes. Nevertheless, we found exactly the same pattern of response as for the complete dataset: greater discrimination between intermediate and long notes in the initial position, and between short and intermediate notes in the final position (Fig. S4). The interaction between substituted note length and note position had a significant effect on dishabituation (parametric bootstrap $P < 0.00005$), just as it did for the complete data.

In summary, note perception does not appear to be affected by syllable-type norm at all. This finding suggests that categorical perception in swamp sparrows reflects broader, phoneme-like perceptual categories, rather than syllable-specific ones. Uniquely among animals, syllable types are learned categories whose constituent units are also learned categories. Our finding of context-dependent categorization, however, suggests that these note-level perceptual categories do not have a simple one-to-one relationship with produced note-type clusters.

Discussion

Previous studies investigating the swamp sparrows' perception of their note types found reduced discrimination either side of a categorical boundary (17) and categorical-labeling responses in neuronal activity (26). Our results support the idea that note types are perceived in a categorical fashion (even though we only tested for reduced discrimination on one side of the boundary). More significantly, our results go on to demonstrate that the categorization of note types is more complex than previously thought: just as in human phonology, context strongly influences how note types are assigned to perceptual categories. Although there are clearly three note-type clusters produced by birds from Pennsylvania (Fig. 1*B*), these clusters do not correspond to three perceptual categories. Instead, both discrimination and labeling tests reveal only two perceptual categories, with the perceptual boundary between categories depending on the position of the note in a syllable. A likely interpretation of our results is that the trimodal distribution of produced notes we examined reflects two overlapping perceptual categories, similar to the phenomenon of phonemic overlapping in speech (14, 15). Short and intermediate notes belong to one category that occurs in the initial position of a syllable, whereas intermediate and long notes belong to the second category, which occurs in the final position. Structurally identical intermediate notes are thus categorized differently when they are in the initial position of syllables from when they are in the final position.

We cannot completely rule out an alternative explanation suggested by phonological rules in human speech (13, 28): that only one perceptual category underlies all three note types, and a rule disallows short notes from occurring the final position and long notes from occurring in the initial position. By this interpretation, short- and long-note allophones are in complementary distribution: both belong to the same category but never occur in the same context. Although both one- and two-category hypotheses are possible, either case requires a complex relationship between underlying representation and surface form that mirrors human phonology.

Table 2. Fit of different GLMM models in labeling experiment

Model	df	AIC	Log-likelihood	χ^2 tests	P (>χ^2)	P (bootstrap)
Model 1 Null (see *Materials and Methods*)	5	316.5	−153.3	—	—	—
Model 2: Model 1 +Type	6	318.1	−153.1	0.407	0.524	0.532
Model 3: Model 2 + Position	7	319.8	−152.9	0.265	0.607	0.612
Model 4: Model 3 + Type x Position	**8**	**312.5**	**−148.2**	**9.378**	**0.0022**	**0.0031**
Model 5: Model 4 + Norm + Norm × Position	12	317.1	−146.5	3.397	0.494	0.535

The χ^2 tests relate to comparisons with the null model 1. The best-supported model is in boldface. The χ^2 tests were between models and the next-simplest model in the list. The sample size was 96 trials for 48 individuals, using 12 syllable types, with three note types (short, intermediate, long), two positions (initial, final), and three syllable-type norms (short, intermediate, long). x represents an interaction term.

The finding that three surface note-type clusters are underlain by a smaller number of perceptual categories raises the question of how these note-type clusters arose and how they are maintained. One possibility is suggested by the fact that intermediate note types are acceptable in both initial and final positions within the syllable. The beginning of a syllable is preceded and marked by a longer than normal internote gap (Fig. 1*A*), but this cue may be more or less ambiguous. We speculate that in ambiguous syllables, notes in the initial position may be erroneously perceived as being in the final position (and vice versa). In such a situation, only intermediate notes guarantee acceptability. We suggest that although the short and long clusters may directly reflect underlying categories, intermediate clusters are an emergent, culturally evolving response to this ambiguity. Analysis of phonological structure across a wider range of populations is required to test this hypothesis.

Our study demonstrates that the phonology and perception of swamp sparrow songs share even more features with human phonology and speech perception than previously suspected. As in speech, swamp sparrows develop population-wide note-type categories that vary between populations, that are perceived categorically, and that have a complex correspondence with produced note-type clusters. As in speech, categorization is influenced by context in a higher hierarchical level, suggesting a possible role for top-down influences on perception. Unlike human speech, however, swamp sparrows do not use this sophisticated phonological system to construct a communication system with symbolic meaning or complex grammar. Instead, our labeling experiment provides a possible clue for the communicative function this phonology might serve. We observed weak responses upon playback of notes that lay outside the population norms. Producing songs that obey phonological rules may signal to others an ability to learn accurately, or membership of the local population, both features that have been proposed to underlie the evolution of accurate learning in songbirds (33, 34). One conclusion that can be drawn from our work, therefore, is that mechanisms underpinning complex phonology, which are present in both human speech and swamp sparrow song, could have evolved before the linguistic traits of semantics and syntax, and instead may have their origins in the logic of assessment signals that govern much of animal communication.

Materials and Methods

Swamp Sparrow Phonology: Distribution of Produced Notes. We recorded and analyzed the song repertoires of 307 male swamp sparrows from two populations. See *Supporting Information* for details.

Discrimination Experiment. We conducted field playback experiments using the protocol of previous experiments that examined note perception in swamp sparrows (17, 26). First, a habituation song stimulus was played in 3-min blocks at six songs per minute. In each block, the intensity of response was tallied as *W*, the number of 10-s subblocks in which at least one wing-wave was observed. There was a 3-min gap between blocks. The block with the maximum response intensity was noted, and the habituation procedure continued until in two consecutive blocks the response was 25% or less than the maximum. At this point, after the 3-min gap, the stimulus was switched to a dishabituation song that was played as a final block. The dishabituation score was calculated as $W_{dishab} - W_{hab}$, the difference in response intensity between the final block and the penultimate block.

Experiments were carried out by one of us (R.F.L.) using a custom program written for the Android system and run on a Motorola Droid smartphone to record wing-wave events and calculate response intensities. The code for this program is available from R.F.L. Stimuli were coded such that the experimenter was blind to the type of trial carried out. Overall, our dishabituation scores were very similar to those in Prather et al. (26), a testament to the unambiguity of the wing-waving display as a response measure. We balanced our design for the order of trials, and for the direction of change in the dishabituation block (i.e., from a shorter to a longer note, or vice versa).

Trials were initiated between 6:00 AM and 11:00 AM and carried out between May 25th and June 19th, 2010. We played songs from an Altec Lansing imt620 speaker, mounted on a tripod at a height of ∼1 m, using song stimuli stored and played through an Apple iPod. Stimulus blocks, concatenated with the 3-min gaps between them, were encoded as wav files with 44.1-kHz sampling rate and 16-bit depth. The blocks were constructed using the Audacity audio editor (v1.3b), and adjusted in amplitude to 80 dB SPL at 1 m. The iPod was controlled with the remote control of the speaker. The experimenter stood at a distance of ∼10 m from the speaker.

Stimuli songs were drawn from the sample of Pennsylvania songs recorded as described in the above section. We only used songs that we judged to be of high recording quality as stimuli. We applied a high-pass filter in Audacity at 1 kHz. All stimuli (including habituation and dishabituation stimuli) were constructed by: (*i*) copying a syllable located near the middle of the recorded song, (*ii*) substituting the focal note for a synthetic note within that syllable, and (*iii*) repeating the modified syllable to make a song at least 2-s long. Three sets of stimuli were constructed for each stimulus song, with synthetic notes of 9.5 ms, 17 ms, and 30 ms in length, chosen to be equally spaced on a logarithmic scale and to coincide with the peaks of note-type distributions (we ensured that the three modified songs had the same number of syllables). The synthetic notes were generated in Praat (www.praat.org), using a custom script. Similar synthetic notes had previously been used in neurobiological studies of swamp sparrows by Prather et al. (26), and allowed us to control both note length and frequency range in light of our finding that both varied between clusters (Figs. S2 and S3*C*). We adjusted start frequency to match that of the note in the original recording (start frequency varied little between the focal note types in the Pennsylvania population). End frequency was set so that bandwidth was typical for the note cluster (Fig. S3*C*), and so that it increased in a linear way between clusters (9.5-ms stimuli: 2.3 kHZ; 17-ms stimuli: 3.3 kHZ; 30-ms stimuli: 4.3 kHZ). Synthesized notes were adjusted in Audacity to match the amplitude of the notes they were replacing, and were placed in the stimulus with a 13-ms gap separating them from the subsequent (for initial position substitutions) or previous (for final position substitutions) note. Intersyllable gaps were set at 35 ms. These gap lengths between notes were chosen to match average values in swamp sparrow songs, and matched those used by Nelson and Marler (17) and Prather et al. (26).

For stimuli, we used syllables that had a note in one of the three identified clusters (Fig. S3) in either the initial or final position. We used 20 syllable types; in 10 of these, we substituted a note in the initial position, and in the other 10 we substituted a note in the terminal position. The original note-type category of the substituted note for eight of the syllables was the intermediate note type; for six it was the short cluster (in the initial position only); and for the remainder it was the long cluster (in the final position only).

Each trial consisted of one of two contrasts between two note lengths, either short vs. intermediate or intermediate vs. long. For each syllable type, we carried out one trial for each of the two contrasts, making 40 trials (with 40 subjects) in total. In half of the trials (randomly selected), the 17-ms stimulus served as the habituation stimulus; in the other half, the 17-ms stimulus was the dishabituation stimulus.

We analyzed playback results with GLMM using lmer (35) in R (36) specifying a Poisson family, and specifying an observation-level random factor to control for overdispersion. We constructed a set of nested models, with the simplest, null model formula being $W_{\text{dishab}} \sim W_{\text{hab}} + (1|\text{Type}) + (1|\text{Trial})$; in other words, we explained the number of displays in the dishabituation block by the number of displays in the habituation block and by two random factors, syllable type, and trial (the latter as the observation-level factor). We compared models using parametric bootstrapping using the pbkrtest package (with 10,000 resamples) and the ANOVA function of lme4, and selected models on the basis of AIC scores. We tested for significance of specific contrasts using the multcomp package (37) with Tukey contrasts.

Labeling Experiment. We played a subset of 12 song types from the stimuli in the discrimination experiment to 48 male swamp sparrows between May 29th and June 20th, 2012. Each trial consisted of 2 min of playback (at a rate of six songs per minute) of one of the stimuli. Each male was given two trials, 3 d apart. In the two trials, different versions of the same song type were played. In one trial, the song contained an intermediate note type, whereas in the second trial either a short or a long note type was substituted instead. Half of the males heard songs with notes substituted in the initial position and half heard songs with notes substituted in the final position. Half of the males heard songs with the intermediate note substitution in their first trial and half of them heard the intermediate note substitution in their second trial.

We played songs from a Bose SoundLink speaker, mounted on a tripod at a height of ∼1 m, using song stimuli stored and played through an Apple iPod. The amplitude of stimuli was adjusted to be 80 dB SPL at 1 m. The playback speaker was placed near the center of a male's territory. Each trial consisted of 2 min of silence followed by 2 min of playback and then

a further 2 min of silence. Male behavior was recorded from the start of playback until the end of the trial by speaking into a voice recorder at a distance of ∼10 m to the speaker. Stimuli were coded such that the experimenter was blind to the type of trial carried out.

During the experiment we recorded four aspects of the subject's response to playback: (*i*) approximate closeness of the subject to the speaker, estimated based on markers placed at 2 m, 4 m, and 8 m on both sides of the speaker before the trial, averaged across 5-s intervals within the trial as described in ref. 38; (*ii*) number of flights past the speaker; (*iii*) number of songs; and (*iv*) number of wing-waving displays. We used a Box-Cox transform ($\lambda_1 = 0$, $\lambda_2 = 1$) on measures 2, 3, and 4. We then carried out a principal components (PC) analysis (using the correlation matrix method) of the responses. We found that the first principal component explained 45% of the total variation in responses, and was positively loaded with all four measures (closeness to the speaker: 0.350; number of flights: 0.592; number of songs: 0.662; and number of wing-wave displays: 0.297). We used this first component as a proxy for overall response intensity.

We analyzed playback results with GLMM using lmer (35) in R (36) specifying a Gaussian family. We constructed a set of nested models, with the simplest, null model specified as PC1∼Trial+(1|Type)+(1|Subject), in other words, explaining variation in PC1 (the first principal component of the playback response) by Trial (whether it was the first or second playback to a subject), and two random factors: Type (the syllable type) and Subject. We compared models using parametric bootstrapping using the pbkrtest package (with 10,000 resamples) and the ANOVA function of lme4, and selected models on the basis of AIC scores. We tested the significance of specific contrasts using the multcomp package (37), which adjusts for multiple comparisons, using Tukey contrasts.

ACKNOWLEDGMENTS. We thank Edna Andrews, Gabriël Beckers, Elliott Moreton, and three anonymous reviewers for providing helpful comments on the manuscript. This research was funded by Duke University and an Arthur and Barbara Pape Award. This is a contribution of the Pymatuning Laboratory of Ecology, which provided logistic support for the field work.

1. Marler P (1970) Birdsong and speech development: Could there be parallels? *Am Sci* 58(6):669–673.
2. Doupe AJ, Kuhl PK (1999) Birdsong and human speech: Common themes and mechanisms. *Annu Rev Neurosci* 22:567–631.
3. Fitch WT (2010) *The Evolution of Language* (Cambridge Univ Press, Cambridge, UK).
4. Berwick RC, Okanoya K, Beckers GJL, Bolhuis JJ (2011) Songs to syntax: The linguistics of birdsong. *Trends Cogn Sci* 15(3):113–121.
5. Bolhuis JJ, Okanoya K, Scharff C (2010) Twitter evolution: Converging mechanisms in birdsong and human speech. *Nat Rev Neurosci* 11(11):747–759.
6. Pegg JE, Werker JF (1997) Adult and infant perception of two English phones. *J Acoust Soc Am* 102(6):3742–3753.
7. Kuhl PK (2004) Early language acquisition: Cracking the speech code. *Nat Rev Neurosci* 5(11):831–843.
8. Repp BH, Liberman AM (1987) *Categorical Perception: The Groundwork of Cognition*, ed Harnad SN (Cambridge Univ Press, Cambridge, UK), pp 89–122.
9. Diehl RL, Lotto AJ, Holt LL (2004) Speech perception. *Annu Rev Psychol* 55:149–179.
10. Samuel AG (2011) Speech perception. *Annu Rev Psychol* 62:49–72.
11. Mann VA, Repp BH (1980) Influence of vocalic context on perception of the [zh]-[s] distinction. *Percept Psychophys* 28(3):213–228.
12. Mann VA (1980) Influence of preceding liquid on stop-consonant perception. *Percept Psychophys* 28(5):407–412.
13. Yip MJ (2006) The search for phonology in other species. *Trends Cogn Sci* 10(10): 442–446.
14. Bloch B (1941) Phonemic overlapping. *Am Speech* 16(4):278–284.
15. Jakobson R, Fant CGM, Halle M (1963) *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates* (MIT Press, Cambridge, MA).
16. Ehret G, Haack B (1981) Categorical perception of mouse pup ultrasound by lactating females. *Naturwissenschaften* 68(4):208–209.
17. Nelson DA, Marler P (1989) Categorical perception of a natural stimulus continuum: Birdsong. *Science* 244(4907):976–978.
18. May B, Moody DB, Stebbins WC (1989) Categorical perception of conspecific communication sounds by Japanese macaques, *Macaca fuscata*. *J Acoust Soc Am* 85(2): 837–847.
19. Wyttenbach RA, May ML, Hoy RR (1996) Categorical perception of sound frequency by crickets. *Science* 273(5281):1542–1544.
20. Baugh AT, Akre KL, Ryan MJ (2008) Categorical perception of a natural, multivariate signal: Mating call recognition in túngara frogs. *Proc Natl Acad Sci USA* 105(26): 8985–8988.
21. Kluender KR, Diehl RL, Killeen PR (1987) Japanese quail can learn phonetic categories. *Science* 237(4819):1195–1197.
22. Lotto AJ, Kluender KR, Holt LL (1997) Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *J Acoust Soc Am* 102(2 Pt 1):1134–1140.
23. Lachlan RF, Anderson RC, Peters S, Searcy WA, Nowicki S (2014) Typical versions of learned swamp sparrow song types are more effective signals than are less typical versions. *Proc Biol Sci* 281(1785):20140252–20140252.
24. Marler P, Pickert R (1984) Species-universal microstructure in the learned song of the swamp sparrow (*Melospiza georgiana*). *Anim Behav* 32(3):673–689.
25. Lachlan RF, Verhagen L, Peters S, ten Cate C (2010) Are there species-universal categories in bird song phonology and syntax? A comparative study of chaffinches (*Fringilla coelebs*), zebra finches (*Taenopygia guttata*), and swamp sparrows (*Melospiza georgiana*). *J Comp Psychol* 124(1):92–108.
26. Prather JF, Nowicki S, Anderson RC, Peters S, Mooney R (2009) Neural correlates of categorical perception in learned vocal communication. *Nat Neurosci* 12(2):221–228.
27. Abramson AS, Lisker L (1970) Discriminability along the voicing continuum: Cross-language tests. *Proceedings of the Sixth International Congress of Phonetic Sciences* (Academia Publishing House of the Czechoslovak Academy of Sciences, Prague) pp 569–573.
28. Clark J, Yallop C, Fletcher J (2007) *An Introduction to Phonetics and Phonology* (Blackwell, Oxford), 3rd Ed.
29. Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *J Stat Softw* 33(2):1–22.
30. Eimas PD, Siqueland ER, Jusczyk P, Vigorito J (1971) Speech perception in infants. *Science* 171(3968):303–306.
31. Harnad S (1987) *Categorical Perception: The Groundwork of Cognition*, ed Harnad SN (Cambridge Univ Press, Cambridge, UK), pp 1–52.
32. Marler P, Peters S (1981) Sparrows learn adult song and more from memory. *Science* 213(4509):780–782.
33. Nottebohm F (1970) Ontogeny of bird song. *Science* 167(3920):950–956.
34. Lachlan RF, Nowicki S (2012) How reliable is song learning accuracy as a signal of male early condition? *Am Nat* 180(6):751–761.
35. Bates D, Maechler M, Bolker BM (2013) lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. Available at: CRAN.R-project.org/package=lme4. Accessed March 3, 2012.
36. R Core Team (2013) *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna). Available at: www.R-project.org. Accessed June 2, 2012.
37. Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. *Biom J* 50(3):346–363.
38. Peters SS, Searcy WA, Marler P (1980) Species song discrimination in choice experiments with territorial male swamp and song sparrows. *Anim Behav* 28(2):393–404.