SUPPLEMENTARY METHODS

Test datasets

Within long-billed hermits, only males in this species sing and crystallized songs appear after birds are about 6 months old. All included songs were crystallized songs. In the study population, relatedness is low even between individuals from the same lek (Araya-Salas et al. 2019) and is likely to be low among all individuals in our sample. We did not control for variation among individuals producing the same song type, as this made the dataset more realistic and provided a good test for our method.

We collected recordings of live budgerigars in a controlled laboratory environment. The individuals used for this study were acquired from a large breeding population and are assumed to have low relatedness. In order to promote calling during recording sessions, we played recordings of unfamiliar budgerigar vocalizations at low amplitudes and also ensured that isolated individuals were in visual contact with the flock mates. Calls were recorded during 30 min sessions that occurred twice per week using an Audio-Technica Pro 37 microphone input to a Dell DHMPC running Syrinx 2.6 (Burt 2006, www.syrinxpc.com) with a 22.05 kHz sampling rate. Calls were automatically partitioned and saved to separate wav files by Syrinx.

To illustrate the manner in which songs and calls are used by budgerigars and long-billed hermits in natural settings, we have included spectrograms of vocal displays recorded from wild long-billed hermits and budgerigars that contain multiple elements (i.e., multiple songs or calls) in Fig. S1. Broadly speaking, long-billed hermit songs span a wider range of frequencies and have higher harmonic content than budgerigar calls, and the differences in the fundamental frequency contours of these signals are readily apparent when viewing spectrograms of recordings.

Synthetic data creation

We varied the duration of synthetic sounds based on the distribution of durations of natural vocalizations in each species (Fig. S2). The natural vocalizations used as templates have very little harmonic content. Hence, harmonic content was simulated arbitrarily as frequency contours an octave (twice the frequency) and a fifth (2.5 times) above the dominant frequency contour. Variation in background noise was generated by adding normally distributed noise (i.e., white noise) to each signal. Allowing for different levels of harmonic content made it possible to simulate recordings with low levels of signal attenuation, such as those collected at close range, as well as recordings with high levels of attenuation, which could be caused by environmental factors such as habitat type, as well as recording conditions. Sample spectrograms of synthesized signals are shown in Fig. S3. Based on visual assessment of spectrograms of synthetic signals and their strong resemblance to spectrograms of the live bird recordings used for this study, as well as the parameters with which we designed the synthetic signals, we are confident that the synthetic signals closely resemble those of live birds. Thus, we expect that signals resembling the synthetic songs might be heard in nature.

By using simulated data with known classes, we were able to make better predictions about which signal characteristics or recording conditions are likely to affect performance while also avoiding the time-consuming collection of data from live animals. This approach of using synthetic data with known variation and class labels for every signal types is analogous to data augmentation in supervised machine learning. Data augmentation is a process in which labeled training data is slightly altered or modified in order to create additional annotated examples for training an algorithm, and is often employed when labeled data is scarce (Krizhevsky et al. 2012). Data augmentation has been shown to enhance performance of deep learning models in the classification of acoustic data (McFee et al. 2015, Salmon and Bello 2017). This approach may be particularly valuable when developing tools to help bioacoustics researchers in the analysis of field recordings because environmental conditions can alter acoustic structure in distinct ways through scattering, frequency-dependent attenuation and introduction of noise. Previous work has shown that creating synthetic datasets can improve performance of unsupervised random forests (Dalleau et al. 2018). In addition, this approach provides test sets with known attributes to evaluate performance of new methods. However, to our knowledge this technique has not been used to evaluate classification methods of animal vocalizations.
The code for data synthesis used in this study is included in the appendix.

Feature measurements

Acoustic features were selected with the aim of creating a general, reproducible framework using accessible, commonly used acoustic measurements. This included mel frequency cepstral coefficients (MFCCs; Lyon and Ordubadi 1982), which have been applied widely applied in the analysis of bioacoustic signals (reviewed in Stowell and Plumbley 2010). For each spectrogram, we calculated 25 MFCCs and their derivatives and extracted descriptive statistics (e.g., mean, median, and variance) from these values *sensu* Salamon et al. (2014), producing 179 MFCC measurements for each audio clip. We measured acoustic parameters using the specan function from R package warbler (Araya-Salas and Smith-Vidaurre 2017), which includes many of the same metrics provided by the seewave R package (Sueur et al. 2008) and Raven Pro (Center for Conservation Bioacoustics, 2019), and included peak frequency (i.e., frequency at which the highest power is present), bandwidth (i.e., frequency range of a signal), signal duration, and robust measurements based on energy distributions within the spectrogram. Recently, researchers have shown that incorporating parameters extracted directly from spectrogram images may facilitate high levels of classification accuracy for avian signals (Smith-Vidaurre et al. 2019); however, to ensure our approach is widely generalizable, we do not include such features here.

Supervised random forest analyses

When using a supervised random forest approach, individual decision trees are constructed by splitting data into two classes at each node using a randomly selected feature measurement, with the goal of optimizing the split between labeled classes. Out-of-bag error is a metric that is commonly used to assess the ability of random forest models to distinguish between distinct classes. Out-of-bag error is calculated by iteratively removing a single sample and building a random forest model with the remaining data, and then testing whether that sample is classified to the same category as other samples from the same class.

Unsupervised random forest analyses

Contrasting supervised random forest models, an unsupervised random forest uses unlabeled samples to create a collection of decision trees by optimally splitting the distribution of values for a randomly selected feature measurement at each node. This process enables unsupervised random forests to find groupings among similar samples and allows for measuring the degree of dissimilarity among all data points (Breiman 2001). This is possible with unlabeled data because decision trees assign all samples to end nodes, i.e., different classes, and one can then calculate the pairwise distance between samples within a data set as the proportion of times a pair of samples is classified in the same end node.

SUPPLEMENTARY RESULTS

We evaluated the ability of our models to correctly classify similar elements together by comparing our results to the classification rates that would be expected by random chance. We calculated random chance of correct assignment as $1/c$, where c is the number of different classes. Note that to find statistical significance of observed correct classification rates versus those theoretically expected by chance one must adjust for a finite number of test data points (see Combrisson and Jerbi 2015). However, we use this value only as a point of reference for assessing supervised random forest performance. To evaluate the performance of our unsupervised method we use rigorous statistical testing, including calculating the acoustic space occupied by all signals in a dataset as well as the adjusted Rand index.

Figure S1. Spectrograms showing examples of vocal displays recorded from wild birds. a) long-billed hermit songs produced by the same individual, b) budgerigar calls produced by the same individual. Sounds were obtained from Xeno Canto (www.xeno-canto.org).

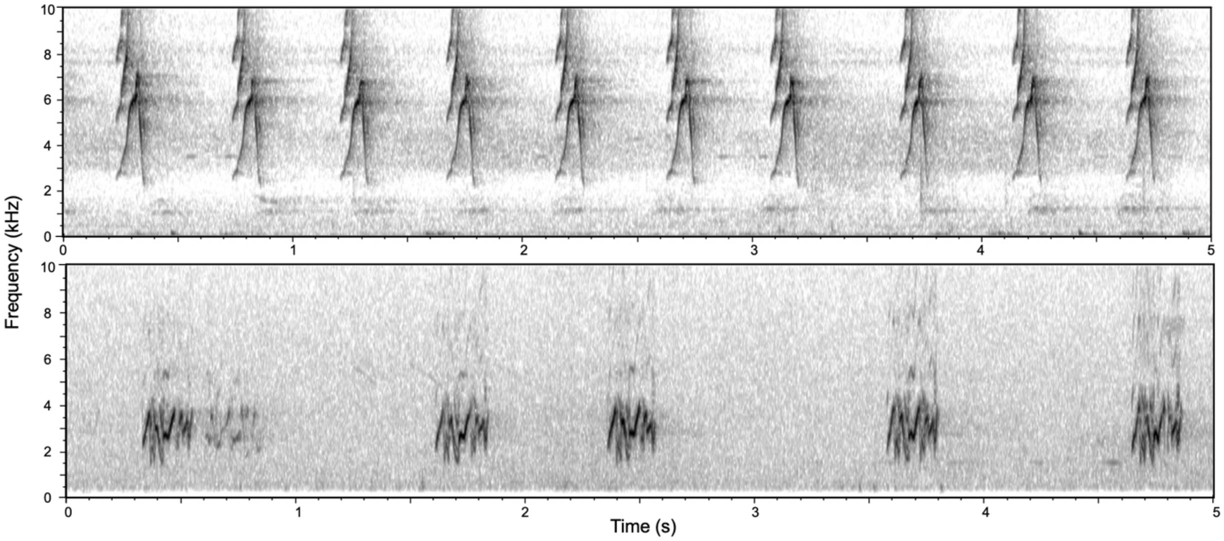


Figure S2. Histograms showing durations of a) field-recorded long-billed hermit songs, and b) lab-recorded budgerigar calls. Distributions of durations from live bird recordings were used to create synthetic datasets.

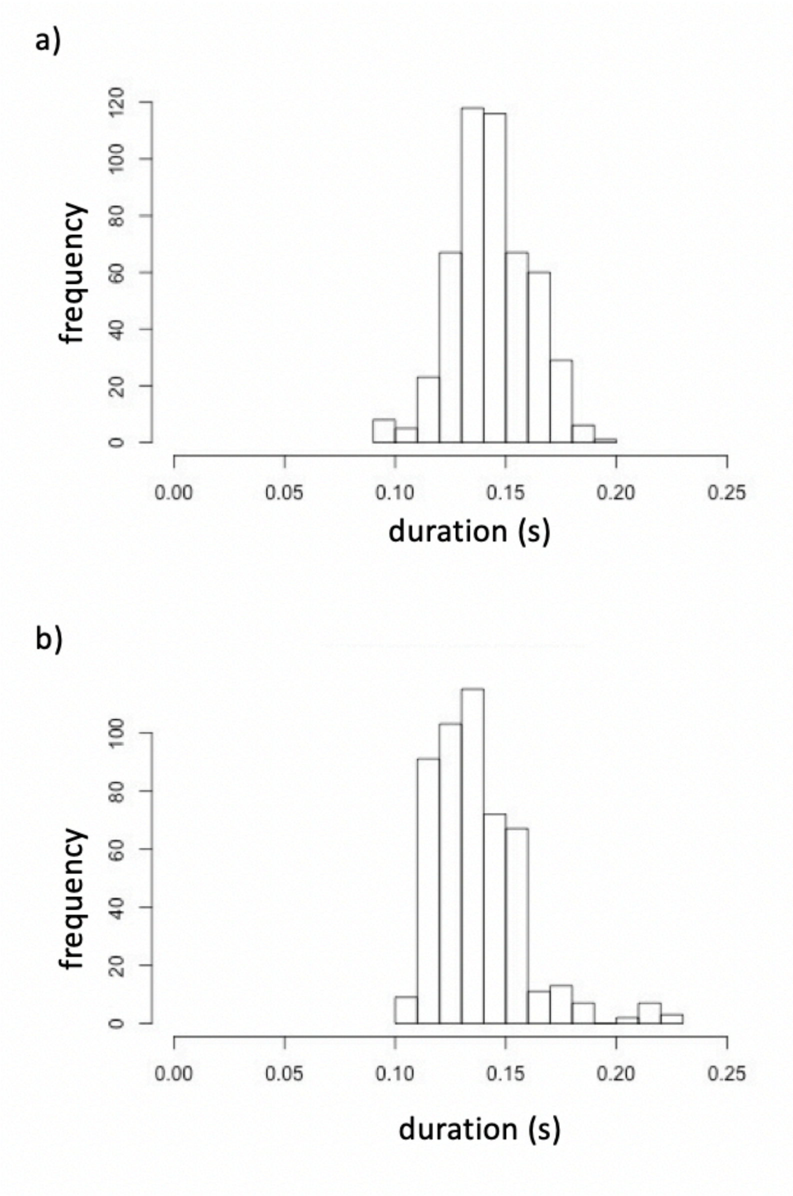
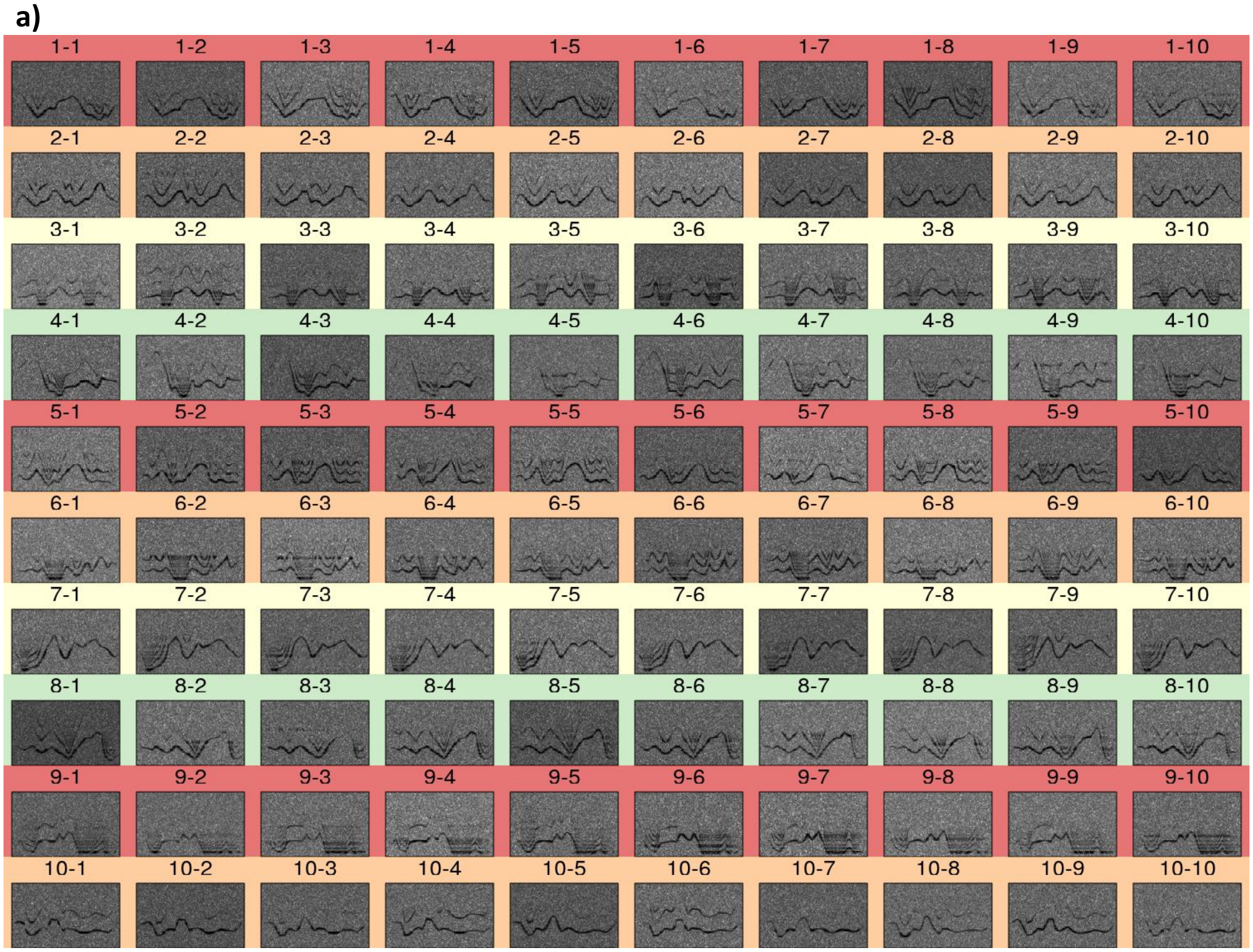


Figure S3. Spectrograms showing examples of signals in test datasets. a) synthetic budgerigar calls, b) synthetic long-billed hermit songs. Spectrograms in the same row show different synthetic signals that are considered to be the same element type.



b)

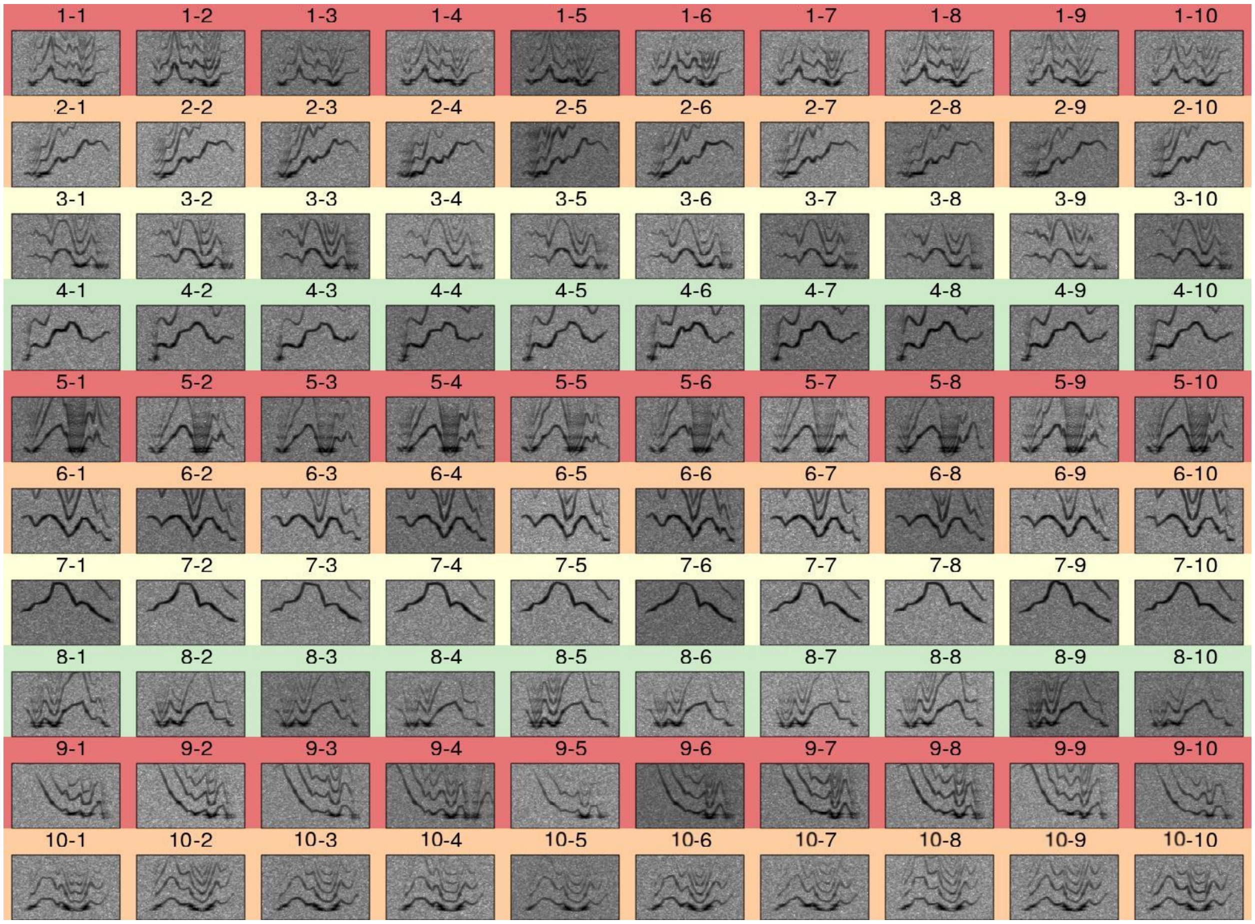

1-1
1-2
1-3
1-4
1-5
1-6
1-7
1-8
1-9
1-10
2-1
2-2
2-3
2-4
2-5
2-6
2-7
2-8
2-9
2-10
3-1
3-2
3-3
3-4
3-5
3-6
3-7
3-8
3-9
3-10
4-1
4-2
4-3
4-4
4-5
4-6
4-7
4-8
4-9
4-10
5-1
5-2
5-3
5-4
5-5
5-6
5-7
5-8
5-9
5-10
6-1
6-2
6-3
6-4
6-5
6-6
6-7
6-8
6-9
6-10
7-1
7-2
7-3
7-4
7-5
7-6
7-7
7-8
7-9
7-10
8-1
8-2
8-3
8-4
8-5
8-6
8-7
8-8
8-9
8-10
9-1
9-2
9-3
9-4
9-5
9-6
9-7
9-8
9-9
9-10
10-1
10-2
10-3
10-4
10-5
10-6
10-7
10-8
10-9
10-10

Figure S4. Sample plots showing silhouette widths for different numbers of clusters applied to distance matrices obtained from unsupervised random forest models. a) synthetic budgerigar dataset with 20 unique element types, b) synthetic long-billed hermit dataset with 20 unique element types, c) synthetic budgerigar dataset with 50 unique element types, d) synthetic long-billed hermit dataset with 50 unique element types, e) synthetic budgerigar dataset with 100 unique element types, f) synthetic long-billed hermit dataset with 100 unique element types, g) lab-recorded budgerigar dataset with 15 unique element types, h) field-recorded long-billed hermit dataset with 50 unique element types. Silhouette width is calculated as the ratio mean distance between points within a cluster to mean distance between clusters; the optimal number of clusters is typically determined to be that which results in the largest silhouette width. Blue and red dashed lines indicate true and estimated number of discrete elements in datasets, respectively. For synthetic datasets with large numbers of discrete elements, silhouette width is often high around the true number of discrete elements, and then increases again when high numbers of clusters are used. High silhouette values with very large numbers of clusters are likely due to overfitting, e.g., clusters created around very few data points.

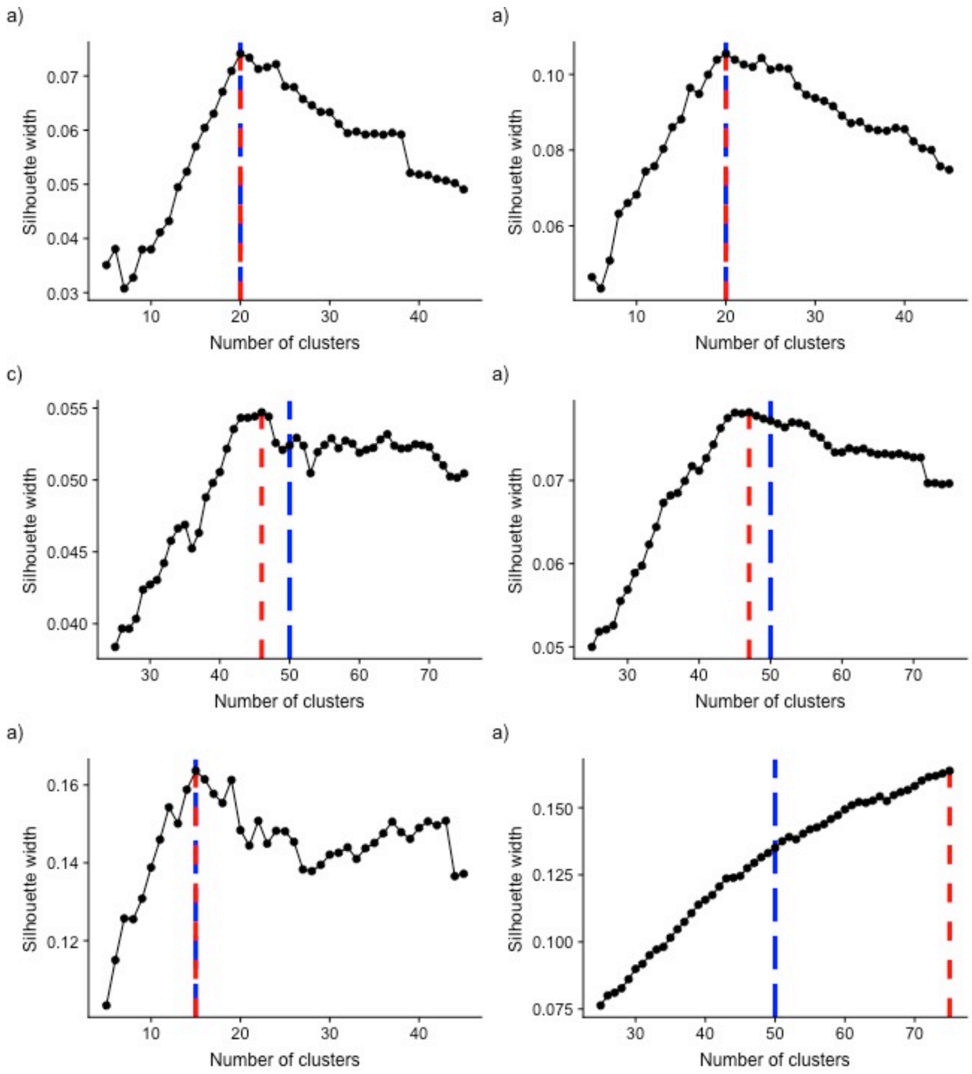


Table S1. Variable importance rankings indicating which feature measurements were most useful in splitting data into distinct classes were different for each of the four dataset types used for testing. Variable rankings were produced by the separate unsupervised random forest models created for each data set. Rankings shown for synthetic data were randomly selected from random forest models created for synthetic budgerigar and long-billed hermit data sets with 100 unique elements. Variable names are listed as they are referred to by the R packages warbleR and seewave and correspond to the feature measurements listed in the main text. The number of variables used varies between datasets because highly correlated measurements were removed before random forest models were created.

Variable ranking	**Field-recorded long- billed hermit songs**	**Lab-recorded budgerigar calls**	**Synthetic long billed hermit songs**	**Synthetic budgerigar calls**
1	var.cc23	max.cc1	min.cc12	max.cc13
2	var.cc16	xc.dim.1	median.cc9	mean.cc24
3	var.cc24	xc.dim.3	kurt.cc21	kurt.cc25
4	var.cc15	xc.dim.2	kurt.cc16	var.cc25
5	var.cc22	xc.dim.4	var.cc4	var.cc8
6	median.cc4	xc.dim.5	max.cc23	var.cc4
7	var.cc14	median.cc2	max.cc22	var.cc22
8	var.cc13	time.ent	median.cc8	skew.cc2
9	var.cc11	freq.Q25	mean.cc23	skew.cc22
10	sfm	freq.median	skew.cc8	kurt.cc20
11	entropy	median.cc16	kurt.cc1	kurt.cc23
12	median.cc3	time.Q75	kurt.cc19	kurt.cc21
13	median.cc5	sfm	var.cc22	skew.cc20
14	dtw.dim.1	min.cc2	skew.cc7	freq.IQR
15	var.cc9	kurt.cc7	max.cc21	time.Q25
16	kurt.cc15	var.cc1	min.cc9	maxdom
17	min.cc15	median.cc3	time.median	xc.dim.4
18	skew.cc4	var.cc7	dtw.dim.3	var.cc6
19	var.cc10	median.cc7	max.cc19	var.cc18
20	mean.cc6	dtw.dim.1	kurt.cc11	var.cc19
21	kurt.cc14	var.cc5	skew.cc22	var.cc5
22	mean.cc15	var.cc4	kurt.cc2	mean.cc22
23	freq.IQR	time.IQR	var.cc3	median.cc15
24	max.cc14	time.median	median.cc13	skew.cc19
25	skew.cc15	var.cc10	var.cc8	skew.cc25
26	var.cc25	entropy	kurt.cc20	var.cc15
27	max.cc13	freq.Q75	mean.cc18	max.cc15
28	mean.cc14	skew.cc1	var.cc25	max.cc20

29	max.cc16	sd	max.cc24	median.cc13
30	skew.cc14	median.cc6	max.cc8	var.cc23
31	min.cc14	var.cc6	max.cc14	var.cc17
32	max.cc15	time.Q25	kurt.cc17	kurt.cc8
33	var.cc21	max.cc3	skew.cc6	var.cc20
34	min.cc10	dtw.dim.4	skew.cc1	var.cc14
35	max.cc11	dtw.dim.5	skew.cc10	mean.cc11
36	modindx	skew.cc2	var.cc16	median.cc7
37	skew.cc18	median.cc5	median.cc24	max.cc25
38	min.cc3	var.cc9	var.cc23	max.cc11
39	min.cc6	var.cc3	var.cc13	max.cc3
40	median.cc7	median.cc15	max.cc20	min.cc11
41	skew.cc10	skew.cc6	max.cc9	min.cc24
42	skew.cc16	max.cc5	max.cc15	min.cc15
43	kurt.cc16	skew.cc3	xc.dim.4	max.cc4
44	max.cc3	mean.cc5	min.cc10	max.cc5
45	max.cc5	median.cc9	max.cc3	min.cc19
46	time.ent	skew.cc5	min.cc15	min.cc25
47	var.cc19	skew.cc8	xc.dim.2	min.cc22
48	skew.cc13	min.cc3	sfm	min.cc18
49	skew.cc9	kurt.cc6	dtw.dim.1	min.cc1
50	time.Q75	kurt.cc4	min.cc19	xc.dim.2
51	var.cc17	var.cc8	sp.ent	xc.dim.3
52	time.median	median.cc11	mindom	min.cc13
53	max.cc19	kurt	min.cc8	var.cc10
54	min.cc16	kurt.cc1	max.cc10	mean.cc23
55	var.cc8	median.cc17	median.cc2	kurt.cc19
56	mean.cc10	var.cc12	mean.cc6	skew.cc13
57	max.cc10	skew.cc4	var.cc24	var.cc13
58	var.cc18	kurt.cc3	median.cc22	median.cc19
59	mean.cc9	min.cc4	skew.cc20	median.cc9
60	kurt.cc11	min.cc6	skew.cc18	median.cc10
61	max.cc23	median.cc10	var.cc9	max.cc23
62	var.cc6	skew.cc7	skew.cc11	median.cc12
63	max.cc18	max.cc9	skew.cc3	median.cc8
64	kurt.cc13	var.cc11	max.cc25	max.cc16
65	kurt.cc7	median.cc18	max.cc16	max.cc17
66	kurt.cc4	var.cc2	median.cc5	min.cc12
67	median.cc15	min.cc7	min.cc25	min.cc9

68	min.cc9	kurt.cc8	min.cc23	min.cc7
69	var.cc12	median.cc20	min.cc21	min.cc21
70	skew.cc11	meanpeakf	min.cc11	min.cc23
71	median.cc14	var.cc24	min.cc4	max.cc1
72	skew.cc3	kurt.cc5	dfrange	max.cc2
73	median.cc11	modindx	modindx	min.cc8
74	var.cc7	median.cc19	xc.dim.5	min.cc2
75	median.cc13	max.cc2	min.cc6	dtw.dim.1
76	min.cc4	var.cc13	dtw.dim.5	dtw.dim.3
77	min.cc24	max.cc7	min.cc2	median.cc21
78	median.cc2	min.cc8	max.cc7	skew.cc10
79	min.cc5	var.cc25	var.cc14	kurt.cc17
80	median.cc18	mean.cc12	kurt.cc14	kurt.cc16
81	skew.cc19	median.cc8	mean.d2.cc	skew.cc5
82	max.cc8	median.cc4	kurt.cc25	var.cc16
83	skew.cc17	median.cc21	kurt.cc3	mean.cc16
84	xc.dim.2	dtw.dim.2	kurt.cc4	median.cc18
85	mean.cc8	var.cc16	skew.cc17	median.cc6
86	var.cc20	min.cc1	var.cc17	max.cc12
87	skew.cc5	max.cc4	var.cc6	var.cc12
88	min.cc22	min.cc16	mean.cc11	median.cc17
89	skew.cc7	min.cc11	var.cc12	max.cc19
90	var.cc3	min.cc5	skew.cc14	median.cc3
91	kurt.cc10	median.cc13	skew.cc5	skew.cc6
92	max.cc4	freq.IQR	skew.cc24	var.cc2
93	min.cc18	dtw.dim.3	kurt.cc24	median.cc5
94	dtw.dim.2	max.cc12	kurt.cc23	skew.cc4
95	min.cc2	kurt.cc2	kurt.cc22	max.cc22
96	min.cc12	max.cc15	skew.cc23	max.cc18
97	median.cc21	max.cc10	skew.cc21	var.cc1
98	time.Q25	kurt.cc9	kurt.cc13	max.cc14
99	meanpeakf	max.cc6	skew.cc25	var.cc7
100	mindom	max.cc21	kurt.cc18	skew.cc3
101	kurt.cc3	min.cc20	kurt.cc9	kurt.cc13
102	median.cc23	max.cc8	kurt.cc15	kurt.cc24
103	min.cc23	var.cc15	kurt.cc7	kurt.cc22
104	var.cc1	var.cc22	kurt.cc5	skew.cc15
105	startdom	skew.cc19	kurt.cc12	var.cc21
106	kurt.cc23	max.cc19	skew.cc15	skew.cc7

107	min.cc7	skew.cc9	kurt.cc10	skew.cc12
108	kurt.cc8	var.cc14	kurt.cc8	skew.cc11
109	max.cc24	median.cc14	skew.cc2	skew.cc23
110	kurt.cc22	var.cc23	skew.cc12	kurt.cc6
111	kurt.cc9	maxdom	var.cc11	kurt.cc2
112	max.cc17	median.cc22	mean.cc10	skew.cc21
113	median.cc16	skew.cc20	var.cc5	kurt.cc1
114	max.cc9	var.cc20	mean.cc4	skew.cc16
115	max.cc21	max.cc22	median.cc16	kurt.cc10
116	min.cc13	max.cc14	median.cc12	kurt.cc4
117	var.cc5	median.cc23	median.cc17	kurt.cc15
118	min.cc20	kurt.cc10	var.cc21	kurt.cc14
119	skew.cc6	kurt.cc11	var.cc15	kurt.cc5
120	median.cc12	min.cc9	var.cc18	skew.cc14
121	var.cc4	var.cc17	var.cc7	skew.cc18
122	dtw.dim.5	skew.cc10	mean.cc7	skew.cc24
123	max.cc22	kurt.cc25	median.cc25	kurt.cc3
124	max.cc1	min.cc13	var.cc2	kurt.cc12
125	max.cc2	var.cc21	mean.cc2	kurt.cc9
126	skew.cc2	kurt.cc12	mean.cc15	kurt.cc7
127	median.cc24	skew.cc11	median.cc21	kurt.cc18
128	skew.cc22	max.cc16	var.cc20	skew.cc17
129	min.cc21	min.cc19	skew.cc4	skew.cc1
130	skew.cc21	mean.cc24	skew.cc9	var.cc9
131	mean.cc25	min.cc10	skew.cc16	mean.cc15
132	var.cc2	max.cc11	kurt.cc6	max.cc24
133	maxdom	max.cc13	skew.cc19	max.cc9
134	xc.dim.4	min.cc18	skew.cc13	min.cc17
135	kurt	max.cc23	var.cc10	min.cc6
136	min.cc17	var.cc18	mean.cc19	meanpeakf
137	max.cc7	var.cc19	median.cc14	startdom
138	kurt.cc21	dfrange	max.cc17	meandom
139	max.cc6	min.cc12	max.cc12	sfm
140	dfrange	min.cc21	max.cc1	time.Q75
141	min.cc19	skew.cc12	min.cc14	mindom
142	min.cc11	min.cc22	meanpeakf	time.ent
143	median.cc22	max.cc20	kurt	kurt
144	dtw.dim.3	max.cc17	freq.IQR	time.median
145	median.cc19	skew.cc18	freq.Q75	freq.Q25

146	skew.cc20	startdom	duration	freq.median
147	median.cc20	skew.cc21	sd	freq.Q75
148	dfslope	kurt.cc13	time.ent	xc.dim.1
149	kurt.cc6	min.cc14	time.Q25	min.cc20
150	skew.cc1	kurt.cc19	min.cc5	median.cc20
151	xc.dim.1	kurt.cc20	min.cc7	var.cc24
152	max.cc25	kurt.cc14	min.cc17	skew.cc8
153	kurt.cc24	skew.cc22	max.cc5	var.cc11
154	skew.cc12	min.cc24	min.cc22	median.cc14
155	min.cc8	min.cc15	max.cc11	median.cc4
156	dtw.dim.4	kurt.cc22	max.cc4	max.cc21
157	skew.cc8	max.cc18	max.cc6	max.cc10
158	skew.cc23	skew.cc17	max.cc13	max.cc8
159	max.cc20	skew.cc15	min.cc20	max.cc7
160	kurt.cc12	kurt.cc23	min.cc13	max.cc6
161	kurt.cc18	min.cc25	min.cc18	min.cc10
162	median.cc17	dfslope	startdom	min.cc4
163	skew.cc24	median.cc25	xc.dim.3	xc.dim.5
164	xc.dim.3	skew.cc16	dtw.dim.4	dtw.dim.4
165	kurt.cc1	skew.cc14	xc.dim.1	dtw.dim.5
166	kurt.cc17	skew.cc13	time.Q75	min.cc5
167	kurt.cc5	max.cc24	maxdom	min.cc3
168	kurt.cc19	min.cc23	enddom	dtw.dim.2
169	max.cc12	kurt.cc16	time.IQR	dfslope
170	min.cc25	min.cc17	dfslope	modindx
171	xc.dim.5	max.cc25	dtw.dim.2	time.IQR
172	kurt.cc2	skew.cc23	min.cc16	enddom
173	skew.cc25	kurt.cc17	max.cc18	min.cc14
174	enddom	kurt.cc24	min.cc24	median.cc25
175	kurt.cc20	kurt.cc15	var.cc19	min.cc16
176	kurt.cc25	kurt.cc18	median.cc20	skew.cc9
177		skew.cc25		kurt.cc11
178		kurt.cc21		var.cc3
179		mindom		elm.type
180		skew.cc24		sd
181		enddom		duration

WORKS CITED

Araya-Salas, M., & Smith-Vidaurre, G. (2017). warbleR: an R package to streamline analysis of animal acoustic signals. *Methods in Ecology and Evolution*, *8*(2), 184-191.

Araya-Salas M, Smith-Vidaurre G, Mennill DJ, González-Gómez PL, Cahill J, Wright TF. (2019). Social group signatures in hummingbird displays provide evidence of co-occurrence of vocal and visual learning. Proceedings of the Royal Society B: Biological Sciences 286:20190666.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*, 250: 126-136.

Dalleau, K., Couceiro, M., & Smaïl-Tabbone, M. (2018). Unsupervised extremely randomized trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Cham, 2018.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097-1105.

Lyon, R. H., & Ordubadi, A. (1982). Use of cepstra in acoustical signal analysis.

McFee, B., E. Humphrey, and J. Bello. (2015). A software framework for musical data augmentation. In *16th International Society for Music Information Retrieval Conf*erence, pp. 248–254.

Raven Pro 1.6.1. Center for Conservation Bioacoustics. (2019). Raven Pro: Interactive Sound Analysis Software (Version 1.6.1) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology. Available from http://ravensoundsoftware.com/.

Salamon, J., Rocha, B., & Gómez, E. (2012, March). Musical genre classification using melody features extracted from polyphonic music signals. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 81-84). IEEE.

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24: 279-283.

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave, a free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*(2), 213-226.

Smith-Vidaurre, G., Araya-Salas, M., & Wright, T. F. (2019). Individual signatures outweigh social group identity in contact calls of a communally nesting parrot. *Behavioral Ecology*.

Stowell, D., & Plumbley, M. D. (2010). Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers. *Centre for Digital Music, Queen Mary University of London, Tech. Rep. C4DM-TR-09-12*.