SUPPLEMENTARY MATERIALS

Analyses reported can be reproduced using the data provided by Smith-Vidaurre et. al. (2019). Code will also be made available on GitHub: https://github.com/gsvidaurre/strong-individual-signatures.

SUPPLEMENTARY METHODS

1. Sound Analysis with Monk Parakeet Contact Calls

1.1 Contact Call Selection, Preliminary Assessment of Acoustic Similarity by Visual Inspection and Quality Control Processing

We selected contact calls from original recordings using Raven version 1.5 (Cornell Laboratory of Ornithology, Ithaca, NY, USA). Although monk parakeets have a diverse vocal repertoire, contact calls have a distinct structure and frequency range that allowed us to distinguish them from other call types (Martella and Bucher 1990). We collected metadata on local call context in a separate spreadsheet as we selected calls, including group size and identity for calls recorded for higher social scales. We imported Raven selection tables containing temporal coordinates of selected calls into R (R Core Team 2018) with the package Rraven version 1.0.4 (Araya-Salas 2017). We used the package warbleR version 1.1.15 (Araya-Salas and Smith-Vidaurre 2017) to optimize

parameters for Fourier transformation and generate spectrograms in R. We proceeded with the following parameters: Hanning window, window length of 378, overlap of 90, minimum color level of -53, bandpass filter of 0.5 to 9 kHz, and amplitude threshold of 15 (% relative to background). For spectrograms, we used a margin of 0.01 seconds around each signal, frequency limits of 0 – 10 kHz and a resolution of 300 ppi for .jpeg image files. Unless specified otherwise, we used these same parameters when generating spectrograms, as well as for measurements of acoustic similarity or acoustic and spectrogram image parameters, throughout our analyses.

We made catalogs of spectrograms at the individual and site social scales using warbleR. We used these catalogs for quality control processing, as well as a preliminary assessment of acoustic similarity by visual inspection. We used catalogs to visually score calls by quality (High, Medium, Low), depending on the visible ratio of signal amplitude to background noise or visible patterns of amplitude saturation. We also scored calls by whether or not there was overlap within the bandpass filter limits with other acoustic signals (e.g. conspecifics or heterospecifics calling nearby or the recordist's narration). We entered this information into the selection table for selected calls. A single observer performed visual inspection for quality control processing and assessing acoustic similarity. Our preliminary results with catalogs of repeatedly sampled individuals indicated relatively high consistency at the individual scale, such that repeated calls from individuals could be identified by consistent frequency modulation patterns that

were distinct from other individuals. However, catalogs at the site social scale generally showed very high variability within sites, indicating that individuals were not converging on shared calls within sites.

We also used catalogs to address potential repeated sampling of the same individual(s) at higher social scales. At higher social scales (pairs, flocks, sites), we were limited to using a single contact call per individual, as birds produced few contact calls at these scales. We selected one contact call per individual and assumed that each contact call represented a unique individual. However, as we were recording unmarked birds, it was possible that some calls represented repeated sampling of the same individual, which could lead to inflated call homogeneity at higher social scales. As our preliminary visual inspection results indicated that repeated calls from individuals could be identified by frequency modulation patterns, one observer visually assessed patterns of individual consistency and distinctiveness in site call catalogs to identify repeated sampling of individuals in the data set used for higher social scales. As the pair and flock social scales were nested within the site scale, the site catalogs contained calls used for the former social scales. We identified 19 calls at this step that represented such probable repeated sampling of individuals based on visual similarity; these were distributed among 11 sites across the transect and represented a small fraction of the total calls in the data set. Each of these 11 sites had a mean of 1.73 calls and a range of 1 – 6 calls flagged as potential repeated individuals. A single site, BCAR, had 6 calls attributed to

repeated individuals. BCAR was a site where we recorded birds making short and frequent flights for stick-collecting, which increased the likelihood of repeatedly sampling the same individuals. Although we did not expect this low level of potential repeated sampling to bias our results, we nonetheless removed these calls from subsequent analyses.

After visual quality scoring and addressing potential repeated sampling of individuals at higher social scales, we imported the selection table of calls across social scales back into R. We used metadata on call quality to remove calls across social scales that had either low-quality scores and/or overlapping signals within the bandpass filter limits (regardless of quality), or calls for higher social scales that had been attributed to potential repeated sampling of individuals. We used the *warbleR* package to continue quality-control processing. We removed calls from duplicate recording sessions, retaining unique recording sessions per recording site that had the most high-quality contact calls. We tailored temporal coordinates of calls across all social scales to reflect consistent selection of start and end times per call. We calculated signal-to-noise (SNR) ratio and removed calls that had SNR < 7 across all social scales. Finally, we calculated sample sizes for unique social groups across social scales. We retained repeatedly sampled individuals with 4 or more contact calls. We retained pairs for which we had a call per individual, and retained flocks for which we had 2 or more calls. We retained sites with 5 or more calls. See Supplementary Table 1 for information on sample sizes across sites.

89

90 Supplementary Table 1: Recording Sites in Uruguay Retained for Sound Analysis

	Site	Site Name	Department	Latitude	Longitude	n_{Calls}	Date
1	PIED	Piedra de los Indios	Colonia	-34.413	-57.849	21	25 - Oct
2	* CHAC	La Chacra de los Olivos	Colonia	-34.413	-57.843	12	21 - Aug
3	LENA	Las Leñas	Colonia	-34.411	-57.838	19	23 - Oct
4	PFER-01	Parque Ferrando - 01	Colonia	-34.468	-57.831	34	19 - Jun
5	PFER-03	Parque Ferrando - 03	Colonia	-34.465	-57.827	19	21 - Jun
6	INES-08	INIA La Estanzuela - 08	Colonia	-34.345	-57.733	27	13 - Jul
7	* EMBR	Embarcadero de Riachuelo	Colonia	-34.444	-57.729	22	21 - Jul
8	INES-01	INIA La Estanzuela - 01	Colonia	-34.349	-57.727	12	03 - Jul

9	INES-07	INIA La Estanzuela - 07	Colonia	-34.346	-57.71	9	13 - Jul
10	INES-06	INIA La Estanzuela - 06	Colonia	-34.344	-57.708	6	13 - Jul
11	RIAC-02	Riachuelo - 02	Colonia	-34.436	-57.706	8	28 - Jun
12	RIAC-01	Riachuelo - 01	Colonia	-34.437	-57.706	17	28 - Jun
13	INES-05	INIA La Estanzuela - 05	Colonia	-34.34	-57.69	6	15 - Jul
14	SEMI	Semillero	Colonia	-34.326	-57.68	11	25 - Jul
15	INES-03	INIA La Estanzuela - 03	Colonia	-34.336	-57.668	15	11 - Jul
16	INES-04	INIA La Estanzuela - 04	Colonia	-34.335	-57.668	9	11 - Jul
17	ARAP	Las Termas del Arapey	Salto	-30.946	-57.52	12	07 - May
18	* 1145	Ruta 1 km 145	Colonia	-34.376	-57.502	13	26 - Jul

19	ROSA	Rosario	Colonia	-34.338	-57.336	15	27 - Jul
20	ECIL	Ecilda Paullier	San José	-34.361	-57.06	17	28 - Jul
21	PAVO	Arroyo Pavón	San José	-34.442	-56.967	25	17 - Oct
22	ARAZ	Balneario de Arazati	San José	-34.535	-56.812	15	03 - Nov
23	KIYU	Balneario de Kiyú	San José	-34.607	-56.715	8	03 - Nov
24	BAGU	La Baguala	Montevideo	-34.848	-56.384	20	09 - Oct
25	INBR	INIA Las Brujas	Canelones	-34.668	-56.33	19	03 - Sep
26	PEIX	Camino Peixoto	Montevideo	-34.765	-56.279	19	06 - Oct
27	BCAR	Bodegas Carrau	Montevideo	-34.788	-56.224	13	20 - Oct
28	FAGR	Facultad de Agronomía	Montevideo	-34.838	-56.219	7	05 - Sep

29	CEME	Cementerio Central	Montevideo	-34.913	-56.187	6	18 - Oct
30	GOLF	Club de Golf	Montevideo	-34.923	-56.164	22	20 - Nov
31	PROO	Parque Roosevelt	Montevideo	-34.855	-56.022	12	14 - Sep
32	PLVE	Plaza Venus, Piriápolis	Maldonado	-34.87	-55.264	11	21 - May
33	QUEB	Quebrada del Castillo	Maldonado	-34.834	-55.26	16	13 - Sep
34	CISN	La Laguna de los Cisnes	Maldonado	-34.861	-55.15	28	13 - Sep
35	SAUC	La Laguna del Sauce	Maldonado	-34.857	-55.041	6	12 - Sep
36	HIPE	Centro de Entrenamiento Hípico Punta del Este	Maldonado	-34.826	-55.01	5	12 - Sep

37	ELTE	El Tesoro	Maldonado	-34.889	-54.863	23	13 - Sep
38	VALI	Barra de Valizas	Rocha	-34.334	-53.803	23	16 - Nov
39	OJOS	Ojos de Agua	Rocha	-33.804	-53.506	23	16 - Nov

All calls were recorded in 2017. The dataset for higher social scales encompassed 605 contact total calls across 39 sites. The three sites where we recorded repeatedly sampled individuals for the individual scale are marked with asterisks (**CHAC**: n_{Calls} = 7, $n_{Individuals}$ = 1; **EMBR**: n_{Calls} = 12, $n_{Individuals}$ = 2; **1145**: n_{Calls} = 78, $n_{Individuals}$ = 5). We obtained recordings for each repeatedly sampled individual over a single day. Contact calls were repeatedly sampled from a single unmarked individual at site CHAC on 21-August, and from 2 marked individuals at site EMBR on 17-June and 21-June. We repeatedly sampled 4 unmarked individuals at site 1145 on 24-June (2 unmarked birds), 26-June, 28-June and one marked individual on 29-June.

1.2 Visual Inspection of Contact Call Acoustic Similarity Across Multiple Observers

We collected classifications of monk parakeet contact calls across social scales by multiple observers. We employed Shiny, which is a flexible framework for building dynamic and interactive graphics in R (Chang et al. 2018). We modified original code provided by Dr. Geovany Ramirez to render spectrograms as drag and drop elements in Shiny (hosted on GitHub: **https://github.com/geoabi/shinyDragAndDrop**), as well as code modified from a multi-page Shiny example created by Jaehyeon Kim (hosted on GitHub: **https://github.com/jaehyeon-kim/shiny-multipage**). We set up our Shiny app to present a 4-class problem per each of the 4 social scales (individual, pair, flock, site) to observers. For the individual scale, we selected 4 calls per each of the 4 repeatedly sampled individuals (16 calls total), used for random forests model validation (Supplementary Methods 1.7). We randomly selected calls from 3 of these individuals that had more than 4 calls, and selected all calls from the individual for which we had sampled only 4 calls (16 calls total). We used these 4 individuals in order to provide a direct comparison to acoustic similarity generated during random forests validation (Supplementary Methods 3.3). For the higher social scales, we selected 4 social groups among all social groups available per social scale. For the pair scale, we randomly selected 4 pairs among the 44 pairs in our data set (8 calls total). We subset the 29 flocks in our data set to retain flocks with 3 calls. We randomly selected 4 flocks out of the remaining 10 flocks (12 calls total). For the site scale, we calculated mean SNR by site and retained the first half of sites with the highest SNR. We randomly

selected 4 sites from this subset of sites. We randomly selected 4 calls from all calls available for these 4 sites (16 calls total). We generated blinded spectrograms after selecting calls across social scales, by removing titles or any other textual information that could give away the social group or geographic location. We set up the app such that each social scale was a separate page, and social scales were presented randomly to each observer. On each page, spectrograms for the given social scale were randomly ordered and then presented together as drag and drop elements. Observers were prompted to evaluate visible patterns of acoustic similarity, and drag each spectrogram into one of 4 separate classes (Classes A through D) based on such perceived patterns of shared call structure. Observers were also informed that number of calls per class was the same across classes. Each time an observer clicked "Next", the app collected the observer's classifications in a .csv file (one per observer).

1.3 Measuring Acoustic Similarity by Spectrographic Cross-Correlation (SPCC)

We used the warbleR package to measure SPCC acoustic similarity (Araya-Salas and Smith-Vidaurre 2017). SPCC slides two spectrograms over each other in sliding time steps and correlates amplitude values at each step. This method yields a pairwise matrix of peak correlation values between acoustic signals. We used Pearson's correlation method to calculate pairwise acoustic similarity among calls. We also used Fourier transformation and other sound analysis

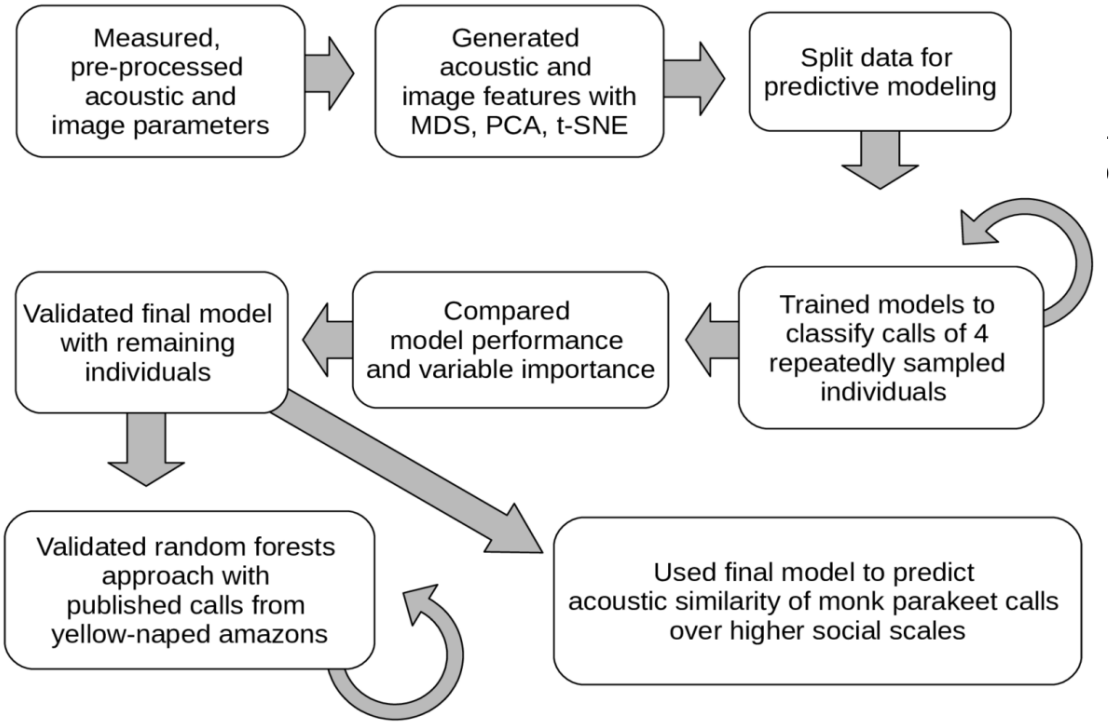
parameters as described above (Supplementary Methods 1.1). We saved the resulting pairwise matrix containing peak correlation values for subsequent analyses, including input into random forests.

1.4 Overview of Random Forests Approach to Measure Acoustic Similarity

Random forests is a machine learning approach used for prediction in classification or regression problems. A forest is composed of up to thousands of decision trees, and each tree splits data based on values of predictor variables. The decision trees generate a random forest by selecting a random subset of predictor variables at each split. The resulting forest of uncorrelated trees, when well-trained, can serve as a strong learner capable of accurate predictions (Valletta et al. 2017), including for avian acoustic signals (Keen et al. 2014; Humphries et al. 2018). We used random forests to measure acoustic similarity from a large set of acoustic and image features characterizing monk parakeet contact call structure.

We implemented random forests in a supervised approach to ensure that models would be biologically relevant. Our approach to supervised model-building and training was influenced by the complexity of these acoustic signals. Our visual assessments of site call catalogs confirmed that calls within sites were so variable that multi-observer scoring or classification (necessary to produce labels for supervised random forests) could be easily confounded. However, after visual assessment of catalogs for repeatedly sampled individuals, we found that

individuals produced sufficiently consistent calls to use individual identity as reliable classes. Thus, we trained models on calls from repeatedly sampled individuals, using individual identities as labels to assess classification performance. This approach allowed us to learn a single acoustic similarity metric for calls over higher social scales that was independent of pair, flock and site labels or geographic distance values. We built and trained three random forests models with half of the repeatedly sampled individuals, and selected among models with the highest classification performance during training. After model validation with the second half of repeatedly sampled individuals, we selected a final model to learn acoustic similarity for calls at higher social scales (e.g. test data set). We extracted the resulting proximity matrix to ask how acoustic similarity manifested across the pair, flock, and site social scales, as well as over geographic distance. See Supplementary Figure 1 for a general workflow of our approach.


Measured, pre-processed acoustic and image parameters
Generated acoustic and image features with MDS, PCA, t-SNE
Split data for predictive modeling
Trained models to classify calls of 4 repeatedly sampled individuals
Compared model performance and variable importance
Validated final model with remaining individuals
Validated random forests approach with published calls from yellow-naped amazons
Used final model to predict acoustic similarity of monk parakeet calls over higher social scales

1.5 Acoustic and Image Parameters Used to Generate Features for Random Forests Models

We used a large set of acoustic and image parameters to build random forests models. These parameters included a variety of measurements of acoustic similarity or acoustic structure: SPCC, dynamic time warping (DTW) on dominant frequency time series (dfDTW), DTW on spectral entropy time series (spentDTW), multivariate DTW on dominant frequency and spectral entropy time series (multiDTW), 27 acoustic parameters measured across the time, frequency and amplitude domains, 88 Mel-frequency cepstral coefficients and derivatives, as well as 2919 image parameters.

SPCC and DTW-based parameters were pairwise acoustic similarity measurements. In selecting 27 acoustic parameters across the 3 sound domains, we excluded all estimates of fundamental frequency from acoustic parameters. During preliminary analyses, we found that fundamental frequency estimates did not map on to fundamental frequency traces visible in spectrograms (using the trackfreqs function in warbleR), likely due to shifts in the relative energy of the fundamental frequency versus higher harmonics throughout the duration of each call. The 27 acoustic parameters we retained were: duration, mean frequency, standard deviation of frequency, median frequency, the first and third quartile frequencies, the interquartile frequency range, median time, first and third quartile times, the interquartile time range, skewness, kurtosis, spectral entropy, time entropy, entropy (product of spectral and time entropy), spectral flatness,

mean dominant frequency, minimum dominant frequency, maximum dominant frequency, dominant frequency range, modulation index, dominant frequency at the start and end of the signal, the slope of the dominant frequency, peak frequency, and mean peak frequency (see Araya-Salas & Smith-Vidaurre (2017) and specan function documentation in the warbleR package for more information on these acoustic parameters). We used warbleR version 1.1.15 to measure all acoustic parameters, including Mel-frequency cepstral coefficients (Araya-Salas and Smith-Vidaurre 2017), which in turn relies on the packages seewave version 2.1.0 (Sueur et al. 2008) and tuneR version 1.3.3 (Ligges et al. 2018). We measured spectrogram image parameters using the image-processing software WND-CHRM version 1.6 (Shamir et al. 2008), which has previously been used to measure and classify cetacean acoustic signals (Shamir et al. 2014). See supplementary code on GitHub for more details on how we generated spectrograms for WND-CHRM. WND-CHRM extracts thousands of image parameters, including Chebyshev statistics, Chebyshev-Fourier statistics, Gabor filters, edge statistics, and other parameters used for image processing. See Shamir et al. (2008) for an extensive list of the image processing parameters measured by WND-CHRM. We used the packages Rtsne version 0.13 (Krijthe 2015), caret version 6.0-80 (Wing and Kuhn 2018), randomForest version 4.6-14 (Liaw and Wiener 2002) and ranger version 0.10.1 (Wright & Ziegler, 2017) for machine learning approaches.

1.6 Extraction of Acoustic and Image Features to Build Random Forests Models

We compiled acoustic and image parameters across calls for feature extraction by complementary unsupervised machine learning methods: Multidimensional Scaling (MDS) and Principal Components Analysis (PCA), versus t-Distributed Stochastic Neighbor Embedding (t-SNE), a newer method for visualization and dimensionality reduction that can outperform PCA under some conditions (van der Maaten and Hinton 2008; van der Maaten 2009). These methods served to convert raw acoustic and image parameters to tabular features for random forests, while reducing the dimensionality and collinearity of the raw parameters. We optimized feature extraction with the repeatedly sampled individual calls, and retained all features derived by the complementary feature extraction methods (MDS and PCA versus t-SNE) for random forests models. We repeated the feature extraction routine for the site call data set, and built a final predictor set of MDS, PCA and t-SNE acoustic and image features for calls across social scales. We added 4 random variables to serve as built-in “noise” variables to ground-truth random forests variable importance. We removed highly collinear features (Pearson’s $r > 0.75$) from our predictor dataset prior to model training and checked that all remaining features were not highly correlated to signal-to-noise ratio (SNR) (Pearson’s $r < 0.75$).

1.7 Splitting Calls for Training, Validation and Testing

We chose 4 repeatedly sampled individuals for supervised model training (73 calls, 75.3% of repeatedly sampled individual calls), and set aside the remaining repeatedly sampled individuals (4 birds, 24 calls, 24.7% of repeatedly sampled individual calls) to validate model performance. We set aside calls at higher social scales for measuring acoustic similarity with the final validated model.

1.8 Training Model 1 with Different Random Forests Implementations

Previous work has shown that random forests implementations in different software (R, Python, SAS) yield different results, particularly related to variable importance (Soifua 2018). We could not find much information comparing the efficacy of different random forests implementations in R. As such, we proceeded by building our first model (Model 1) with two implementations from the ranger and randomForest packages. We retained all acoustic and image features that remained after filtering for high collinearity. We tuned *mtry* over 10 evenly spaced values from 2 to the total number of predictors. *mtry* is the number of random variables to be selected at each decision tree split, injecting randomness into the resulting forest. We iterated over varying numbers of trees (500, 1000, 1500, 2000, 2500). We trained models using 5 iterations of repeated 10-fold cross-validation via the caret package. We compared training performance and variable importance of Model 1 over values of *mtry*, total trees and the ranger and randomForest implementations. We used permuted variable importance for ranger, and Gini variable importance for randomForest.

1.9 Model 1 Training Performance and Variable Importance Results

ranger yielded higher Model 1 training accuracy than randomForest (Supplementary Table 2). We also found different variable importance results between implementations, similar to reported results among random forests implementations in R, Python and SAS (Soifua 2018). However, these differences in variable importance we identified could in part be due to using different variable importance metrics per implementation. Given the difference in performance between implementations, we proceeded with the ranger implementation.

Supplementary Table 2: Random Forests Model Training and Validation Performance For Monk Parakeet Contact Calls

Model	Implementation	Final Number of Trees	mtry	Training Accuracy (%)	Validation Accuracy (%) by Model-based Clustering
1	ranger	2500	33	99.18	-
	randomForest	2500	348	87.4	-
2	**ranger**	**500**	**2**	**100**	**95.8**
3	ranger	2500	2	100	95.8

Random model forests model training and validation for monk parakeets. Model 1 corresponds to the full set of acoustic and image features. Models 2 and 3 were built by either manual or automatic feature selection. The final number of trees is the total number of decision trees grown for each forest. mtry is the number of variables randomly selected at each split per tree. Training accuracy is reported as the percentage of correctly classified calls reported by random forests. Validation accuracy is reported as the percentage of correctly classified calls by model-based clustering on the proximity matrix. The final model we used for predicting acoustic similarity over higher social scales is in bold.

1.10 Training ranger *Models 2 and 3*

We built and trained two additional ranger models. We built Model 2 by manual feature selection, in which we removed variables with importance equal to or less than random variables in Model 1. We built Model 3 by automatic feature selection, using a built-in bagged trees caret function. We trained Models 2 and 3 by iterating over *mtry* and total trees as in Model 1 training.

1.11 Comparing Classification Performance and Variable Importance Across ranger Models

All three ranger models ranked several SPCC and Mel-frequency cepstral features among the top important variables. In preliminary results with repeatedly sampled individual calls, we found that SPCC and Mel-frequency cepstral

coefficients represented visible patterns of individual consistency and distinctiveness. Therefore, we considered variable importance of SPCC and Mel-frequency cepstral coefficients reliable indicators of models' biological relevance. Model 1 achieved 99.18% training classification accuracy, while Models 2 and 3 both achieved 100% training classification accuracy. We chose Models 2 and 3 (manual or automatic feature selection, respectively) for model validation.

1.12 Random Forests Model Validation with ranger Models 2 and 3

We performed model validation by predicting acoustic similarity of the repeatedly sampled individual validation dataset with Models 2 and 3. As this dataset encompassed different classes (e.g. different individuals altogether) than calls used for training, we ignored the random forests classification results and extracted the proximity matrix as the predicted acoustic similarity. We ran model-based clustering on the proximity matrix using mclust version 5.4.1 (Scrucca et al. 2017) to ask how well each random forests model predicted patterns of acoustic similarity with respect to individual identity. We allowed the clustering algorithm to choose a best number of clusters among 1 – 6 total clusters (2 beyond the true number of clusters, e.g. 4 repeatedly sampled individuals). The clustering approach identified 4 optimal clusters for both Models 2 and 3, matching the true number of individuals used for validation, and classified all but one call correctly for 95.8% classification accuracy (Figure 2C, Supplementary

Table 2). We chose Model 2 (manual feature selection) for final testing, although Model 3 (automatic feature selection) would have served just as well.

1.13 Additional Validation of Acoustic Similarity Predicted by ranger Model 2

We performed additional validation of acoustic similarity predicted by Model 2. We ruled out a role for SNR in driving patterns of acoustic similarity predicted by random forests. We identified centroid calls per cluster (see above). We used Spearman's correlation to determine whether distance to centroid for non-centroid calls was significantly correlated with SNR, and found no significant correlation (Spearman's *rho* = -0.04, *p* = 0.8562).

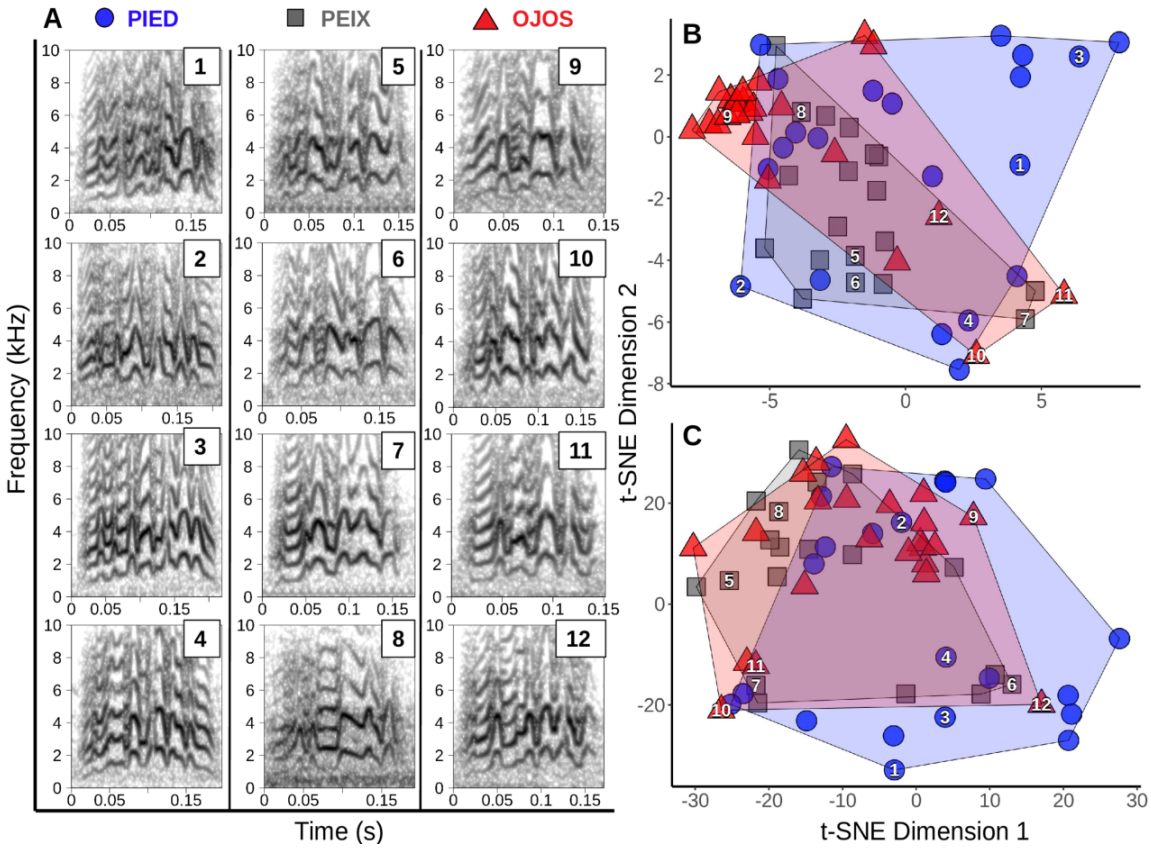
1.14 Predicting Acoustic Similarity at Higher Social Scales with ranger Model 2

Our validation results confirmed that random forests yielded biologically relevant acoustic similarity patterns. Indeed, random forests acoustic similarity reflected patterns of individual consistency and distinctiveness identified by SPCC (Figure 2B). We used the final, validated random forests model (ranger Model 2, Supplementary Table 2) to predict acoustic similarity of calls at higher social scales. We extracted the resulting proximity matrix for subsequent analyses to ask how acoustic similarity manifested across the pair, flock, and site social scales, as well as over geographic distance. We did not use the random forests proximity matrix for the repeatedly sampled individual validation dataset in

subsequent analyses at the individual scale, as we had used these individuals to train models.

1.15 Higher Acoustic Similarity over Closer Geographic Distances by Random Forests

Random forests, by relying on many quantitative features, picked up a significant signature of geographic distance (e.g. overdispersion among sites in acoustic space) missed by SPCC (Figure 3B,D, Supplementary Figure 2B,C, Table I). Among the features that displayed such signatures of geographic distance were dfDTW t-SNE features, MDS features of acoustic parameters measured across the time, frequency and amplitude domains, MDS and t-SNE features of Mel-frequency cepstral coefficients, and MDS and t-SNE features of image parameters.



Supplementary Figure 2: Monk parakeet contact calls exhibit low acoustic similarity within sites. A) Catalog with spectrograms of 4 randomly selected calls for 3 sites across the transect: PIED (westernmost), PEIX (middle) and OJOS (easternmost) (Supplementary Table 1). The legend indicates site identity. B) Distribution of calls in SPCC acoustic space. C) Distribution of contact calls in random forests acoustic space. We used t-SNE for dimensionality reduction of similarity matrices. Numbered symbols in B and C correspond to numbered spectrograms in A. Convex hull polygons in B and C delineate the acoustic space encompassed by each site's set of calls.

2. Using Matrix Regression to Evaluate Patterns of Acoustic Similarity over Social Scales and Geographic Distance

The Mantel test is a linear matrix regression method (Mantel 1967) that can be used to assess relationships between variables composed of non-independent data, including pairwise similarity measurements (Wright 1996). Another useful feature of Mantel tests is that they can accept matrices of binary or continuous values, such that acoustic similarity matrices can be correlated against matrices of binary group identity or geographic distance values (Wright 1996). As such, Mantel tests have often been used to assess patterns of SPCC acoustic similarity in parrot contact calls (Wright 1996; Guerra et al. 2008; Wright et al. 2008).

We used Mantel tests to ask if calls were more similar within social groups at each social scale, and if acoustic similarity decreased over geographic distance for monk parakeets. We performed Mantel tests on calls from 4 social scales and 2 geographic scales (regional: all sites across the transect ,and local: sites in the Colonia department), using SPCC and random forests acoustic similarity. We encoded individual identity or social group membership at each social scale by generating pairwise binary identity matrices (e.g. 1 = two calls from the same individual or social group, 0 = two calls from different individuals or social groups). We converted acoustic similarity and binary identity matrices to distance matrices by subtracting matrices from 1. We implemented Mantel tests using the R package vegan version 2.5-2 with 9999 permutations (Oksanen et al. 2018). We also used vegan to perform Mantel-based spatial autocorrelation with

999 permutations, to evaluate whether acoustic similarity decreased in a linear fashion over increasing geographic distance. We split calls into 25 distance classes of 2km or 52 classes of 10km at the local and regional geographic scales, respectively. The first distance class per geographic scale included calls recorded at the same site. We dropped distance classes with zero or too few observations. We generated correlograms using Holm's p-value correction for multiple testing (Holm 1979).

3. Validation of Our Analytical Approach and Findings with Monk Parakeets

3.1 Validation of Our Random Forests Approach with Another Parrot Species

3.1.1 Overview of Species Comparison

We validated our analytical approach measuring acoustic similarity of monk parakeet calls by SPCC and random forests. We asked whether SPCC and random forests could identify previously documented patterns of acoustic similarity in another parrot species. We compared random forests and SPCC acoustic similarity at the site social scale between monk parakeets and yellow-naped amazons (*Amazona auropalliata*), a species that exhibits hierarchical mapping over social scales and regional dialects on the Pacific coast of northern Costa Rica and southern Nicaragua (Wright 1996). We used contact calls recorded in a single year for each species, and published calls for yellow-naped

amazons (Wright, 1996). We measured SPCC similarity and built random forests models per species. For this section of the Supplementary Methods, we abbreviate monk parakeets as MNK and yellow-naped amazons as YNA.

3.1.2 Pre-processing MNK and YNA Contact Calls

YNA calls were contained within cuts of original recordings. We pre-processed YNA calls in a manner consistent with our previous pre-processing of MNK calls (Supplementary Methods 1.1). We removed YNA calls with visibly obvious background noise. We did not calculate SNR for YNA calls, as there was not sufficient time before and after selected calls in each cut to calculate noise levels. We standardized MNK calls to the same sampling rate as YNA calls (22050 Hz). Calls for both species were at 16 bit sampling depth. We extracted selected MNK calls as cuts of original recordings to mirror selection of YNA calls. We added 0.5 seconds of silence before and after calls of both species to facilitate SPCC measurements. We made selection tables to facilitate measuring acoustic and image parameters in R and WND-CHRM.

3.1.3 Fourier Transformation Parameters Used for Species Comparison

We identified Fourier transformation parameters for pre-processed MNK and YNA calls. For MNK, we settled on: Hanning window, window length of 288, overlap of 90, and minimum color level of -40. We kept all other parameters the same as in our previous analyses with the MNK calls (Supplementary Methods

1.1). For YNA, we used: Hanning window, window length of 378, overlap of 90, minimum color level of -40, 0 – 4 kHz bandpass filter, and amplitude threshold of 10 (% relative to background). Unless specified otherwise, we used these same parameters for all measurements of acoustic similarity or acoustic and image parameters.

3.1.4 Measuring Acoustic Similarity by SPCC

We measured SPCC acoustic similarity for both species. We reran SPCC for MNK calls using the parameters above, as these calls had been down-sampled. We used Pearson's correlation method and saved the resulting pairwise matrices containing peak correlation values for subsequent analyses.

3.1.5 Overview of Random Forests Modeling Approach

We measured acoustic similarity of contact calls by random forests for MNK and YNA. We used a similar workflow to measure parameters, extract features, build, train, and validate models as in our previous analysis with monk parakeet calls (Supplementary Methods 1.4 – 1.15). We used the resulting proximity matrices to evaluate whether random forests could identify previously documented patterns of acoustic similarity for YNA at the site social scale (Wright 1996).

3.1.6 Training Model 1 Between Species

As in our previous random forests analysis with MNK calls, we used both the ranger and randomForest implementations to build and train Model 1 per species. We used repeatedly sampled individuals and regional dialects for MNK and YNA model training, respectively. We trained MNK Model 1 using the same repeatedly sampled individuals as in our prior approach (73 calls across 4 individuals or 75.3% of MNK calls at the individual scale). We trained YNA Model 1 using 4 sites for each of the Northern and Southern regional dialects documented in northwestern Costa Rica in 1994 (Wright 1996). Each site had 23 – 40 calls, for a total of 274 calls, or 65.7% of the YNA calls. We iterated over *mtry* values and total number of trees as before.

3.1.7 Model 1 Training Performance Between Implementations

We found that ranger again outperformed randomForest in training classification accuracy. Variable importance metrics differed between the implementations. These differences in model training performance and variable importance held across species. We decided to proceed with the ranger implementation, as in our prior random forests modeling approach.

3.1.8 Building and Training Model 2

We built Model 2 by manually selecting the most important features from Model 1 using the mean importance of built-in random variables as a threshold. We

trained Model 2 on the same calls from repeatedly sampled individuals or regional dialects used for Model 1 training.

3.1.9 Model 2 Training Performance

We compared performance and variable importance metrics across the best performing model per species. We found that Model 2 per species yielded high classification accuracy during training (> 95%).

3.1.10 Model 2 Validation Performance

We proceeded with model validation with Model 2 per species (Supplementary Table 3). We did not perform model validation for MNK and YNA Model 1, as these models had lower training performance. The YNA validation data set was composed of 36 calls from 3 sites representing 2 dialects (2 Northern and 1 Southern (Wright 1996). Each site had 10 – 16 calls from 1 – 2 individuals. We extracted random forests proximity matrices per validation data set and performed model-based clustering. We restricting the clustering algorithm to the true number of clusters present in the validation data set, which served to assess the biological relevance of our models. Model-based clustering for MNK exhibited 91.7% classification accuracy with validation calls, with only 2 misclassified calls (Supplementary Table 3). The YNA model yielded 100% classification accuracy by random forests (this was possible to assess because the training and validation data sets contained the same class labels, e.g. Northern and Southern

dialects), and 100% classification accuracy by model-based clustering (Supplementary Table 3).

Supplementary Table 3: Random Forests Model Training and Validation Performance For Analysis with Yellow-Naped Amazons

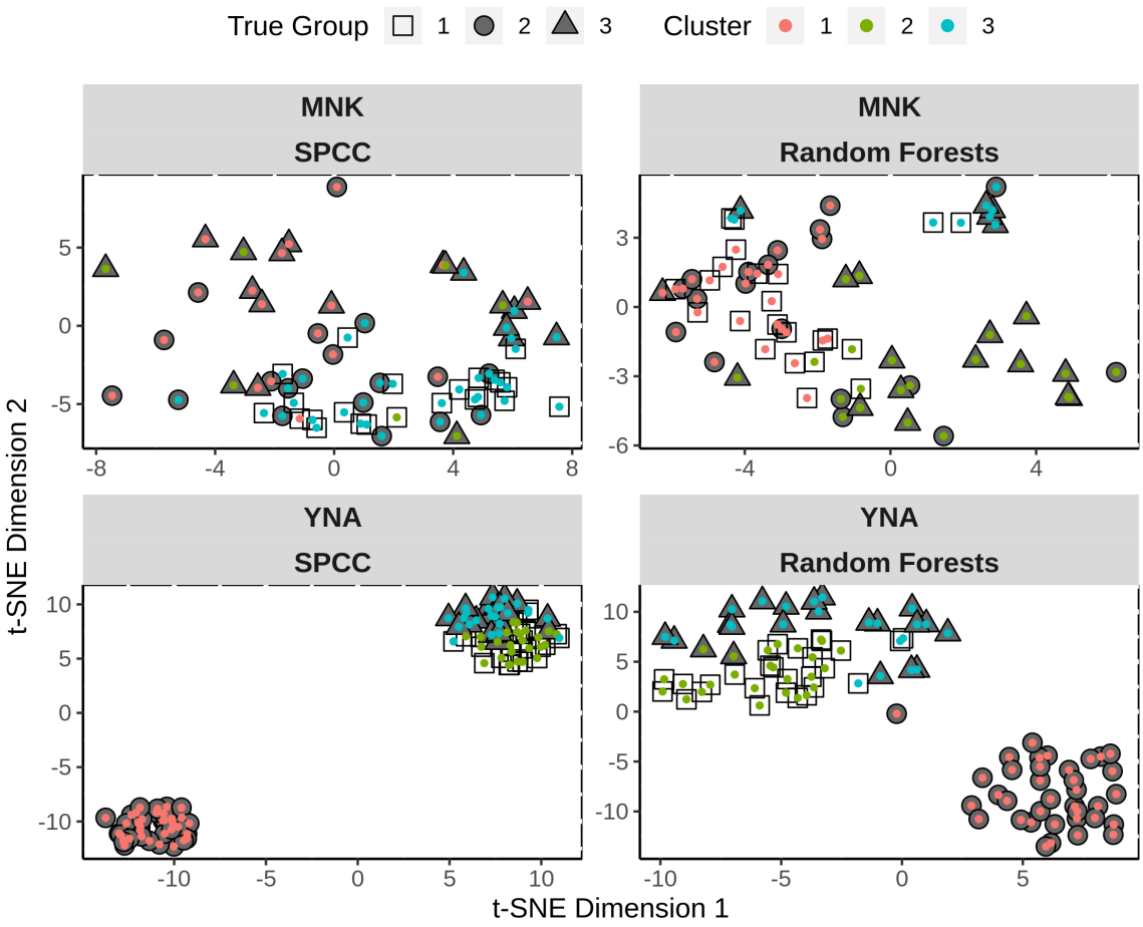
Species	Model	Implementation	Training Labels	Final Number of Trees	mtry	Training Accuracy (%)	Validation Accuracy (%) by Model-based Clustering
MNK	1	ranger	Individuals	2500	88	96.71	-
		rf	Individuals	2000	230	70.68	-
	2	**ranger**	**Individuals**	**2500**	**58**	**97.53**	**91.7**
YNA	1	ranger	Dialects	2000	65	98.47	-
		rf	Dialects	2500	382	92.92	-
	2	**ranger**	**Dialects**	**2000**	**2**	**98.98**	**100**

We validated our random forests approach by measuring acoustic similarity for yellow-naped amazons (YNA) at the site scale. Monk parakeets are abbreviated as MNK. Models 1 and 2 per species correspond to the full set of features or manually selected features, respectively. The final number of trees corresponds

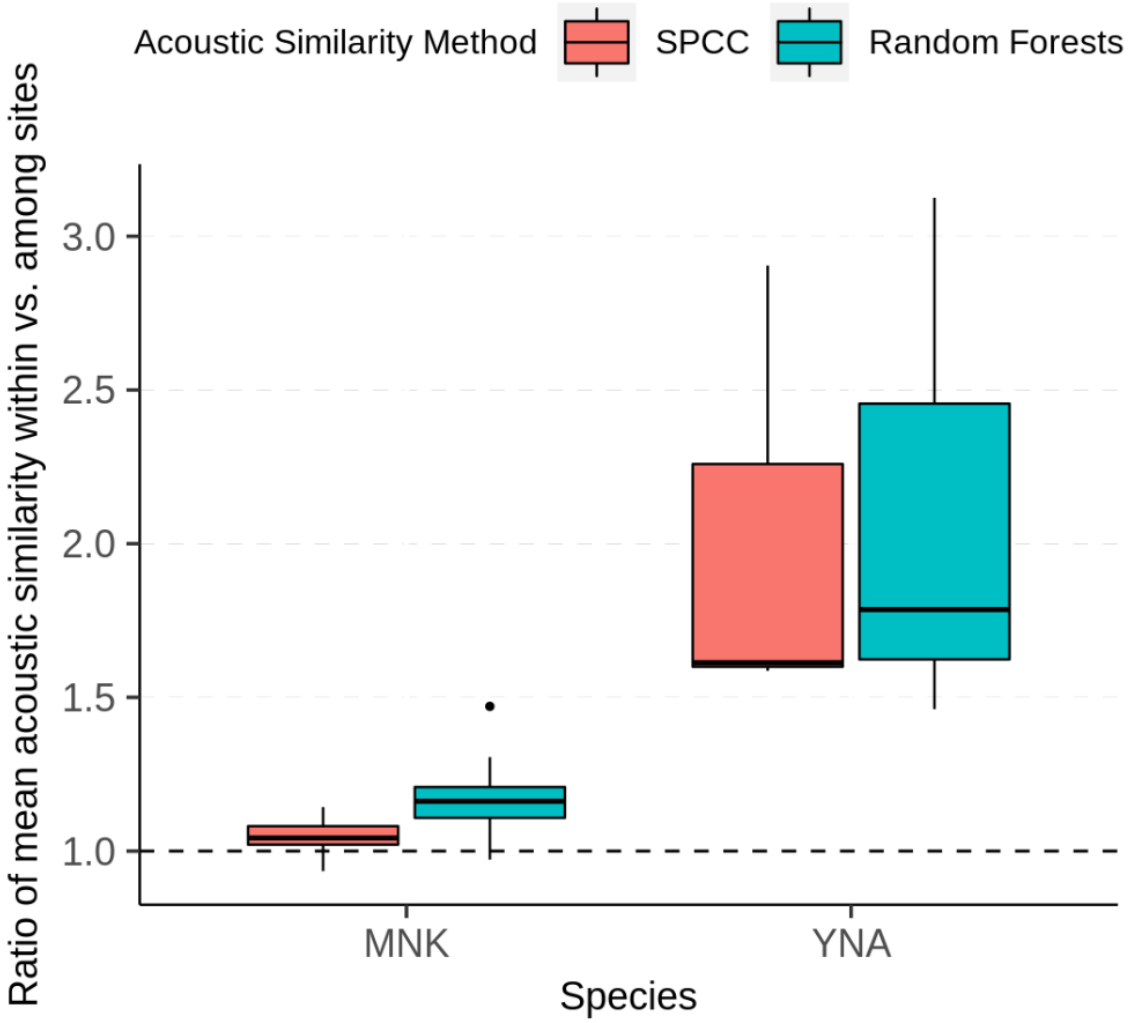
to the number of decision trees grown for each forest, and mtry is the number of variables randomly selected at each split per tree. We report the training accuracy as the percentage of correctly classified calls. For validation accuracy, we report the percentage of correctly classified calls by model-based clustering on the resulting proximity matrix. Models that we used for predicting acoustic similarity at the site social scale per species are shown in bold.

3.1.11 Predicting Random Forests Acoustic Similarity at the Site Scale with Model 2 and Comparison with SPCC Acoustic Similarity

We chose ranger Model 2 per species to predict acoustic similarity at the site scale (MNK: n_{Calls} = 598, n_{Sites}= 39, $n_{Individuals}$ = 598, and YNA: n_{Calls} = 86, $n_{Dialects}$= 2, n_{Sites}= 3, $n_{Individuals}$ = 13). As in model validation, we assessed patterns of acoustic similarity between species by performing model-based clustering on the random forests proximity matrices. Here, we restricted the clustering algorithm to the true number of clusters per dataset. We also performed model-based clustering on SPCC matrices between species in the same way. We compared clustering patterns arising from SPCC and random forests acoustic similarity (Supplementary Figure 3), as well as the ratio of within-site compared to among-site acoustic similarity between species (Supplementary Figure 4).



Supplementary Figure 3: Model-based clustering on SPCC and random forests similarity matrices at the site scale for monk parakeets (MNK) and yellow-naped amazons (YNA). We used MNK calls for higher social scales (n_{Calls} = $n_{Individuals}$ = 598, n_{Sites} = 39). We used 86 total calls for YNA ($n_{Individuals}$ = 10, n_{Sites} = 3, $n_{Dialects}$ = 2). We reduced dimensionality using t-SNE. Site identity was poorly reconstructed for MNK by both SPCC and RF, supporting the fact that acoustic similarity within sites was low. Both SPCC and RF identified previously documented patterns of high acoustic similarity within sites for YNA (Wright 1996). The YNA data set included a Nicaraguan dialect site (circles), which was more distant in acoustic space relative to the Northern dialect sites (triangles, squares) (Wright 1996).



552 Supplementary Figure 4: Acoustic similarity at the individual and site social
553 scales for monk parakeets (MNK) and yellow-naped amazons (YNA). We used
554 the same calls from random forests prediction of site scale similarity as shown in
555 Supplementary Figure 3. Acoustic similarity is represented as the ratio of within
556 versus among sites for both SPCC and random forests. The dashed line at 1
557 represents acoustic similarity within sites equal to acoustic similarity among sites.
558 Both SPCC and random forests reconstructed the previously documented pattern
559 of high acoustic similarity within sites for YNA (Wright 1996).

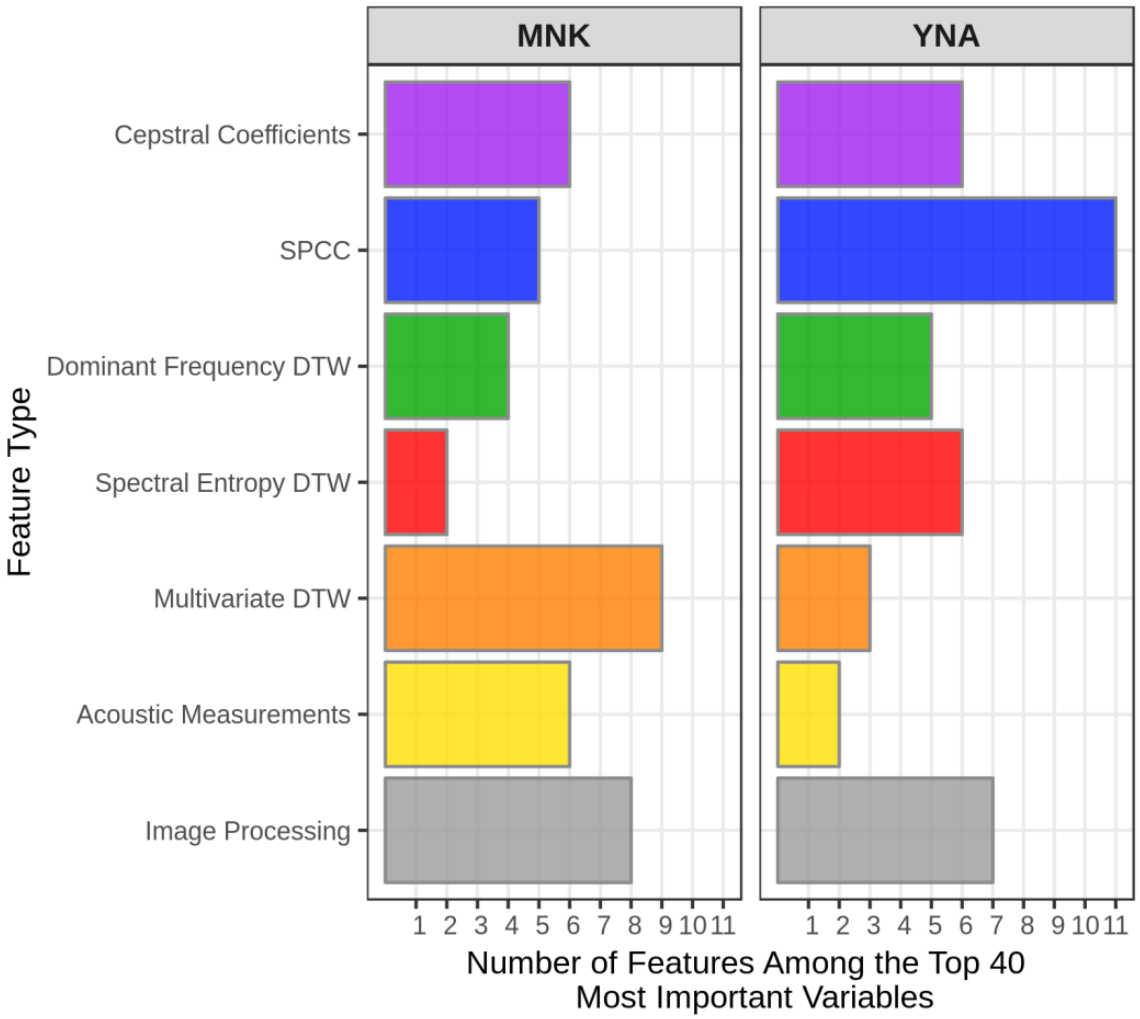
3.1.12 Random Forests is a Valuable Acoustic Similarity Method

Supervised random forests yielded highly accurate patterns of acoustic similarity for MNK and YNA. SPCC and random forests reconstructed the sharp boundaries of acoustic similarity previously found at the site social scale for YNA (Supplementary Figures 3, 4) (Wright 1996). Moreover, SPCC and random forests identified the mosaic pattern that is characteristic of dialects for yellow-naped amazons (first reported using SPCC) (Wright 1996). Both SPCC and random forests identified low acoustic similarity among Nicaraguan and Northern dialect sites, as evidenced by greater separation in acoustic space between these two dialects (Supplementary Figure 3). Neither SPCC nor random forests reconstructed discrete clustering by sites for MNK (Supplementary Figure 3). The fact that we reconstructed previously identified patterns of acoustic similarity for YNA supports the robustness of our analytical approach and findings with MNK.

3.1.13 Image Features Were Useful Measurements of Acoustic Structure

Our random forests models relied on spectrogram image features, which are not frequently used for analyzing animal acoustic signals (Shamir et al. 2014). Random forests consistently ranked image features among the most important variables across final models trained for MNK and YNA, suggesting spectrogram image features would be useful in future research (Supplementary Figure 5). Although variable importance of different feature types varied across species, image features were frequently represented among the top 40 important

582 variables per model. SPCC, multiDTW, and Mel-frequency cepstral features were
583 also highly ranked in models across species, suggesting these parameters would
584 also be of interest for future analyses (Supplementary Figure 5).



Supplementary Figure 5: Acoustic and image features represented in the top 40 most important variables during Model 2 training for monk parakeets (MNK) and yellow-naped amazons (YNA). Feature type abbreviations: Cepstral coefficients = Mel-frequency cepstral coefficients, SPCC = spectrographic cross-correlation, Dominant Frequency DTW = DTW on dominant frequency time series, Spectral Entropy DTW = DTW on spectral entropy time series, Multivariate DTW = multivariate DTW on dominant frequency and spectral entropy time series, Acoustic parameters = parameters measured across the three domains of sound using the function specan in the warbleR package, Image processing = spectrogram image processing parameters. Models relied most heavily on SPCC, Mel-frequency cepstral coefficients, multivariate DTW and spectrogram image features.

3.2 Inter-Observer Reliability of Visual Inspection

We based our quantitative approaches of measuring acoustic similarity on preliminary results from visual inspection by a single observer. We found patterns of relatively high consistency within individuals and distinctiveness among individuals, suggesting that identities of repeatedly sampled individuals could serve as reliable labels for random forests classification (Supplementary Methods 1.1). Here, we validated these preliminary findings with visual inspection by asking how reliably multiple observers classified calls at the individual scale. We used results from the Shiny app designed to collect visual classification results across multiple observers (Supplementary Methods 1.2*)*.

We performed an analysis of inter-observer reliability using calls classified at the individual scale by 12 observers (4 calls from each of 4 individual birds). At this social scale, classes generally contained a majority of calls from a single individual, such that it was possible to assign each individual to a different class and find how many calls had been misclassified across observers. The mean classification accuracy across observers was 71.82% +/- 15.94% (mean +/- SD). This relatively high classification accuracy confirmed that monk parakeet individuals produce consistent and distinctive calls, and that these patterns of acoustic similarity can be reliably identified by visual inspection.

3.3 Comparison of Visual Inspection, SPCC and Random Forests as Methods of Measuring Contact Call Similarity

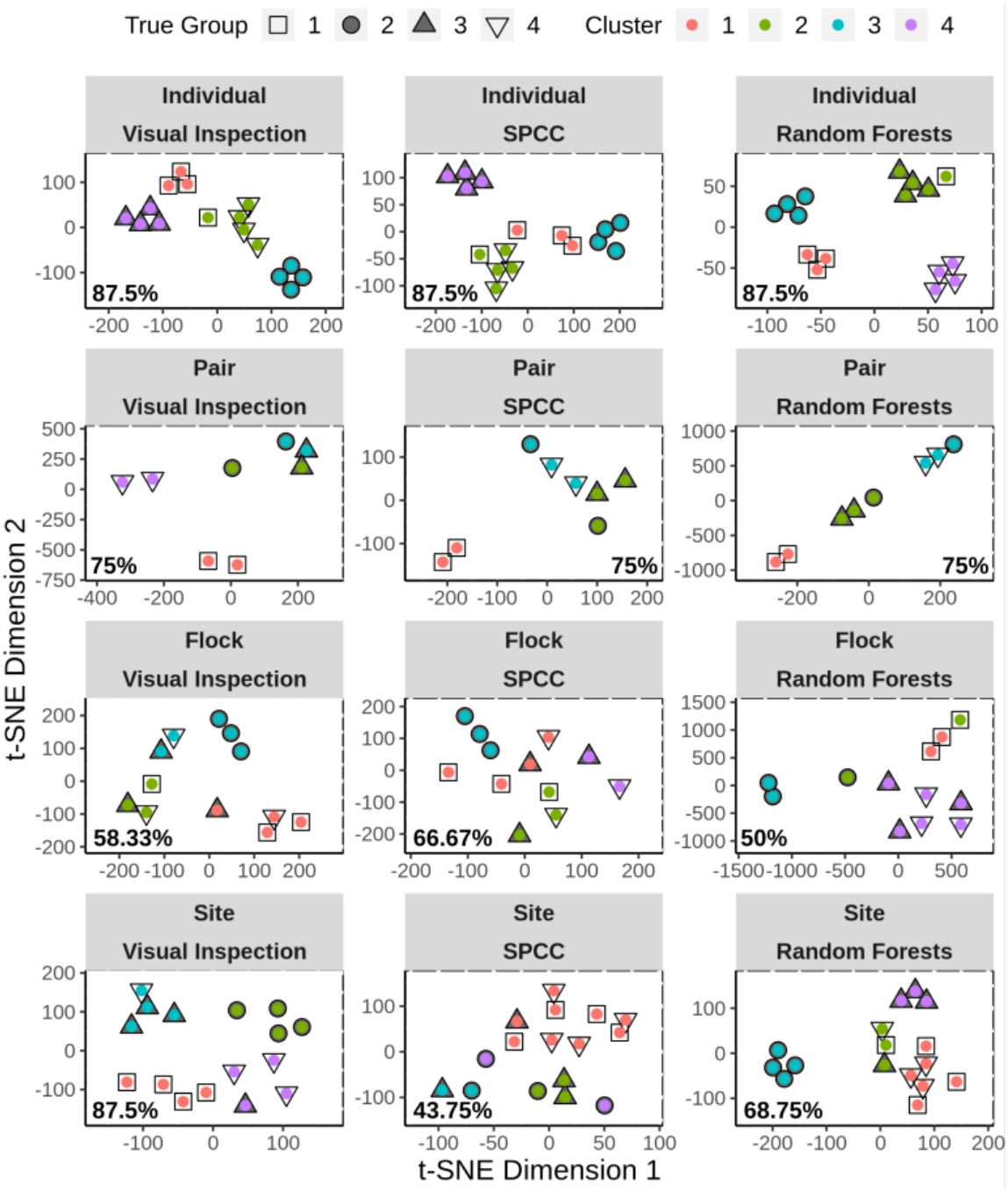
3.3.1 Obtaining Classification Accuracy Across Social Scales and Similarity Methods

We compared our three similarity methods to validate our overall analytical approach to measuring similarity of monk parakeet contact calls, which exhibit complex acoustic structure. Two of these methods, visual inspection and SPCC, have traditionally been used to assess similarity of learned acoustic signals (Nowicki and Nelson 1990; Farabaugh et al. 1992; Wright 1996; Guerra et al. 2008). Random forests has been used less frequently to assess similarity of avian acoustic signals (Keen et al. 2014; Humphries et al. 2018), and has not yet been reported as a method to assess acoustic similarity of parrot acoustic signals.

We began this analysis by converting classifications obtained by visual inspection via our Shiny app (Supplementary Methods 1.2) to quantitative measurements of visual similarity, which facilitated a direct comparison among similarity methods. Classifications varied considerably over higher social scales and observers. Classes often did not contain a majority of spectrograms belonging to a single social group. As such, it was often not possible to assign a social group to each class for higher social scales (e.g. Site X to Class A and Site Y to Class B). In turn, we could not calculate classification accuracy by evaluating how many calls had been assigned to the “wrong class” per social group. We turned to a matrix-based approach. We converted classifications per

each of the 12 observers into pairwise binary matrices, in which 1 represented two calls classified together and 0 represented two calls assigned to different classes. We added matrices across observers to obtain a single matrix that encoded the total number of times that pairs of calls had been classified together or apart. We scaled this matrix to a range of 0 – 1 to yield a matrix representative of visual similarity, and repeated this process across social scales. We subset the SPCC and random forests acoustic similarity matrices by the same calls used across social scales for visual inspection in the Shiny app. We converted the visual and acoustic similarity matrices to distance matrices by subtracting them from 1.

We used model-based clustering as a classification approach to assess how well social group identity could be reconstructed across social scales using each similarity measurement. We restricted clustering algorithms to 4 clusters, which was the true number of social groups presented at each social scale to observers. We then reduced the dimensionality of the acoustic distance matrices to 2 dimensions using t-SNE, which facilitated visualization of calls in two-dimensional acoustic space (Supplementary Figure 6). We calculated the percentage of incorrectly classified calls per social group within each social scale to evaluate how well calls were classified across social scales per similarity method (Supplementary Figure 6).



663 Supplementary Figure 6: A comparison of our three complementary similarity
664 methods. We used model-based clustering to compare how well similarity
665 methods identified patterns of acoustic similarity relative to social group
666 membership. Similarity methods are displayed in columns and social scales are
667 shown in rows. We used t-SNE for dimensionality reduction. Filled and non-filled
668 symbols correspond to the true group to which each call belongs across social

scales. Colored circles inside group symbols correspond to the clustering assignment of each call. Text in each panel corresponds to the percentage of correctly classified calls per visual or acoustic similarity method and social scale. Note that the number of incorrectly classified calls generally increases with social scales across methods, with the exception of visual inspection at the site scale (see Supplementary Methods 3.3.2 for more information).

3.3.2 Classification Accuracy Generally Decreased Over Social Scales using Visual Inspection, SPCC and Random Forests as Similarity Methods

Classification accuracy was high at the individual scale across similarity methods (87.5%, Supplementary Figure 6). These results indicated that the patterns of individual consistency and distinctiveness we used to inform our random forests approach were not an artifact of visual inspection by a single observer (Supplementary Figure 6). Classification accuracy decreased notably across higher social scales, with the exception of visual inspection (Supplementary Figure 6). Visual inspection and random forests outperformed SPCC acoustic similarity at the site social scale (87.5% and 68.75% versus 43.75% classification accuracy, respectively, Supplementary Figure 6). Interestingly, visual inspection yielded classification accuracy at the site scale that was as high as the individual scale, indicating that visual inspection could identify the weak patterns of acoustic convergence present at the site scale. However, we feel that this pattern of high classification accuracy yielded by visual similarity is in part due to the small number of social groups used for visual inspection. Given our preliminary visual inspection results (Supplementary Methods 1.1), classification accuracy would likely decrease significantly if observers were prompted to classify calls back to more social groups at this social scale.

3.4 Evaluating Differences in Social Context Between the Individual and Higher Social Scales

3.4.1 Implementation of a Permutation Test to Assess the Effect of Social Context on Acoustic Convergence

Repeatedly sampled individuals were often recorded while perched in isolation from social companions. At higher social scales, we often recorded calls from birds flying with a social group. For instance, at the site scale, we recorded calls from individuals flying in pairs (181 calls or 29.92% of the full data set for the site social scale), although we also recorded some individuals flying alone (47 calls or 7.77% of the full data set). This difference in sampling between the individual and higher social scales was primarily a difference in social context (vocalizing alone versus in a social group). It was possible that such differences in social context could have skewed our results. We did not expect other behavioral contexts (e.g. food deprivation, predator avoidance, courting) to affect sampling at one social scale more than any other scale. We reasoned that in general, in a social context, individuals might either converge more or less on calls with social group members, compared to calling alone. If so, then individuals sampled for the site scale while flying in a group would produce calls either more similar or more different to other birds sampled at the same site, compared to individuals sampled while flying alone.

We evaluated this possibility with a permutation test using SPCC acoustic distance. We obtained acoustic distances by subtracting SPCC values from 1. We identified sites at which we had sampled birds flying alone and birds flying in social groups at the site social scale. We performed permutation tests per site using inter-individual SPCC distances. We obtained inter-individual SPCC distances for each lone individual between all other individuals sampled at the same site. We also obtained inter-individual SPCC distances for each individual sampled in a social context between all other individuals sampled at the same site. We obtained the absolute value of the difference in mean SPCC distance between these two groups, which served as the observed acoustic distance between social contexts. We then combined these SPCC distances for permutation. We randomly sampled the combined SPCC distances without replacement, using the number of SPCC distances for the lone social context for the given site as the sample size. We calculated the absolute value of the mean SPCC distance between the permuted values extracted for each social context. We repeated this process over 1000 iterations. We calculated p-values as the number of times the permuted difference was greater or less than the observed difference, divided by the number of iterations. These p-values allowed us to assess whether or not the permuted difference in means was greater or less than the observed difference in means between social contexts. We also calculated the effect size for the observed difference in SPCC mean distances between social contexts and the 95% CI of this effect size (Supplementary Table 4). We

used the effsize package version 0.7.4 (Torchiano 2018) to calculate Cohen's *d* statistic, using pooled standard deviation between groups and Hedge's *g* correction to account for bias by such pooling.

We performed a separate permutation test for site 1145, where we had repeatedly sampled 5 individuals for the individual scale (including 4 unmarked birds perched in isolation from social group members), as well as birds flying in social groups for the site social scale. We repeated the permutation test as above, albeit with a few differences. We obtained inter-individual SPCC distances for individuals repeatedly sampled at site 1145 at the individual social scale. We excluded SPCC distances among an individual's own calls (e.g. intra-individual SPCC distances). We obtained inter-individual SPCC distances for individuals sampled in a social context at the site scale. We used the number of SPCC distances for individuals sampled in a social context at site 1145 to randomly sample acoustic distances in the permutation test. We combined the results of this permutation test (also run with 1000 iterations) with those from the permutation test above. We evaluated the significance of p-values after adjusting alpha of 0.05 using Bonferroni's correction for multiple testing (adjusted alpha = 0.0012, Supplementary Table 5).

Supplementary Table 4: Assessing the Effect of Social Context on Acoustic Convergence

Site	Sample Size	p_Higher	p_Lower	Effect Size	95% CI
ARAZ	28	0.183	0.817	0.33	(-0.08, 0.74)
CHAC	11	0.328	0.672	-0.37	(-1.04, 0.29)
CISN	135	0.069	0.931	0.20	(0.01, 0.39)
ELTE	66	0.594	0.406	0.09	(-0.18, 0.35)
FAGR	18	0.256	0.744	-0.29	(-0.92, 0.35)
GOLF	**105**	**0.000**	**1.000**	**0.60**	**(0.30, 0.89)**
INES-03	14	0.295	0.705	0.39	(-0.16, 0.94)
INES-04	16	0.445	0.555	0.23	(-0.34, 0.81)
INES-07	8	0.786	0.214	0.10	(-0.8, 1.00)
INES-08	26	0.342	0.658	0.25	(-0.14, 0.65)
KIYU	14	0.290	0.710	-0.36	(-1.00, 0.28)
LENA	**72**	**0.000**	**1.000**	**-0.58**	**(-0.87, -0.30)**
OJOS	44	0.155	0.845	-0.28	(-0.61, 0.05)
PAVO	**96**	**0.000**	**1.000**	**-0.39**	**(-0.65, -0.13)**
PEIX	18	0.049	0.951	0.64	(0.15, 1.13)
PFER-01	33	0.368	0.632	0.21	(-0.16, 0.57)
PIED	60	0.002	0.998	0.49	(0.20, 0.78)
QUEB	45	0.308	0.692	-0.19	(-0.54, 0.16)
ROSA	14	0.285	0.715	0.39	(-0.16, 0.94)
VALI	22	0.867	0.133	-0.05	(-0.51, 0.42)
1145	60	0.029	0.971	0.40	(0.15, 0.66)

A permutation-based test of the effect of social context on acoustic convergence. P-values represent the likelihood that the difference in mean permuted inter-individual SPCC distances was higher or lower than the observed difference in mean inter-individual SPCC distances between lone and social contexts.

Significant p-values are in bold, evaluated after adjusting alpha of 0.05 with Bonferroni's correction (adjusted alpha = 0.0012). We calculated the effect size and 95% CI for the observed difference in mean SPCC distance between the lone and social contexts per site (Cohen's d statistic with Hedges' correction).

Supplementary Table 5: Effect Sizes for Acoustic Convergence at the Individual Scale in Contact Calls

Repeatedly Sampled Individual	Effect Size of Observed Difference	95% CI of Effect Size
AAT	2.51	(2.23, 2.79)
BIRD 1	2.10	(1.91, 2.20)
BIRD 2	0.98	(0.84, 1.12)
BIRD 3	4.29	(3.59, 4.99)
BIRD 4	0.75	(0.51, 0.98)

Effect sizes and 95% CI for the difference in mean SPCC distance within compared to among repeatedly sampled individuals at site 1145 (intra-individual versus inter-individual SPCC distance). This represented the strength of acoustic convergence at the individual scale (e.g. individual signatures). We calculated effect sizes using Cohen's d statistic and Hedge's correction. These effect sizes were used as a baseline for judging the strength of the effect of social context on acoustic convergence (Supplementary Table 4). Although some sites displayed a statistically significant effect of social context on acoustic convergence

(Supplementary Table 4), the effect sizes we report here are larger than those for the effect of social context on acoustic convergence.

3.4.2 Differences in Social Context Between the Individual and Higher Social Scales Were Unlikely to Bias Acoustic Convergence Results

Of the 21 sites used for the permutation test, we found that only 3 sites (GOLF, LENA, PAVO) demonstrated a significant difference in SPCC distances between social contexts (individuals sampled while flying alone versus individuals sampled while flying in a social group). These effect sizes varied in direction: at GOLF, individuals sampled in a lone context produced slightly more different calls relative to other calls at the same site, while at LENA and PAVO, individuals sampled in a lone context produced slightly more similar calls relative to other calls at the same site. Importantly, we did not find a significant difference in mean SPCC distance among calls of repeatedly sampled individuals and individuals sampled in a social context for the site scale at site 1145.

We assessed the strength of the effect of social context on acoustic convergence. We calculated the effect size and 95% CI of the difference in mean SPCC distances within compared to among repeatedly sampled individuals at site 1145 (e.g. the effect size of individual signatures or acoustic convergence at the individual scale), to serve as a baseline for evaluating the magnitude of effect sizes reported between the lone and social contexts (Supplementary Table 5). We calculated effect sizes using the same procedure as above. Overall, the

mean effect size we found for the observed difference in mean SPCC distance within compared to among repeatedly sampled individuals (2.12 +/- 1.42) was about 4 times greater than effect sizes corresponding to the statistically significant differences in SPCC distance we identified between lone and social contexts at sites GOLF, LENA and PAVO (0.52 +/- 0.12, mean and SD calculated from absolute values, Supplementary Tables 4, 5). Our results indicate that differences in social context while sampling across social scales were unlikely to bias the acoustic convergence results we present in this study.

3.5 Accounting for Differences in Motivational Context Among Calls Recorded for the Individual Scale

Calls for repeated individuals were recorded in narrow windows of time within a single day per individual. We recorded calls from Unmarked Bird 1 over 8.50 minutes, Unmarked Bird 5 over 3.40 minutes and marked bird AAT over 5.74 minutes, respectively. All other repeatedly sampled individuals were recorded over a single day, and typically within a 2 hour window. At times, we followed marked individuals for up to 5 hours but did not successfully record calls. Individuals could have experienced differences in motivational context that could have affected call structure similarity over these narrow sampling windows.

We assessed whether acoustic similarity of calls for repeatedly sampled individuals was influenced by their position within the full temporal sequence of calls. Although we did not always have exact times per recording, so as to link

826 together call sequences across recordings, we had selected contact calls
827 sequentially within and across recordings per individual. Therefore, we assigned
828 calls per individual sequential integer values representing their position in call
829 sequences. We converted the SPCC acoustic similarity matrix for repeatedly
830 sampled individuals to a distance matrix by subtracting values from 1. We subset
831 this SPCC distance matrix to retain calls for each individual. We then generated
832 a distance matrix of temporal sequence distance per individual, and performed
833 Mantel tests to ask whether temporal sequence distance was significantly
834 correlated with acoustic distance. For individuals with more calls, we used 9999
835 permutations, although for individuals with fewer calls, permutations were limited
836 to the maximum number of permutations possible (Supplementary Table 6). We
837 adjusted alpha of 0.05 to 0.0062 using a Bonferroni correction to account for
838 multiple testing. We found no significant relationship between acoustic distance
839 and the position of calls within call sequences per individual (Supplementary
840 Table 6).

841 We repeated this analysis for 4 unmarked individuals with call sequences
842 contained in a single recording. We calculated the exact temporal distance
843 among calls per individual using start and end times within recordings, and used
844 these for a Mantel test as described above. We again found no significant
845 influence of temporal distance among calls and pairwise SPCC similarity
846 measurements. Here we used unmarked birds 2 – 5. The strongest Mantel r and
847 p-value (UM3, Mantel $r = 0.57$, $p = 0.0750$, $n_{Calls} = 5$) was not significant at an

alpha of 0.0125 (adjusted by Bonferroni correction, see supplementary code for all test statistics and p-values). Overall, these results suggest that differences in motivational context during our narrow sampling windows did not significantly influence the results we present here at the individual scale.

Supplementary Table 6: A*ssessing Differences in Motivational Context Among Calls Recorded for the Individual Scale*

Repeatedly Sampled Individual	Number of Calls	Mantel r	Mantel p	Permutations
RAW	4	-0.14	0.5833	23
ZW8	8	-0.04	0.5525	9999
AAT	12	0.05	0.3491	9999
BIRD 1	25	0.14	0.0407	9999
BIRD 2	23	0.09	0.0813	9999
BIRD 3	5	0.7	0.0500	119
BIRD 4	13	-0.24	0.9747	9999
BIRD 5	7	-0.31	0.9151	5039

Mantel test results indicate no significant correlation between SPCC acoustic distance among calls for each repeatedly sampled individual and the position of each call within temporal call sequences. Alpha was adjusted from 0.05 to 0.0062 using a Bonferroni correction to account for multiple testing. Mantel permutations were limited for individuals with fewer calls.

4. Additional R Packages Used for Data Management, Visualization and Analysis

We relied on additional R packages across our analyses: corrplot (Wei and Simko 2017), data.table (Dowle and Srinivasan 2018), dplyr (Wickham et al. 2018), dtw (Giorgino 2009), e1071 (Meyer et al. 2017), edarf (Jones and Linder 2017), facetscales (Oller Moreno 2018), forcats (Wickham 2018), ggplot2 (Wickham 2016a), gtable (Wickham 2016b), lattice (Sarkar 2008), magrittr (Bache and Wickham 2014), MLmetrics (Yan 2016), pbapply (Solymos and Zawadzki 2018), shadowtext (Yu 2017), shinyjs (Attali 2018), shinythemes (Chang 2018), shinyWidgets (Perrier et al. 2019) and tidyverse (Wickham 2017).

REFERENCES

Araya-Salas M. 2017. Rraven: Connecting R and Raven bioacoustic software (version 1.0.0).

Araya-Salas M, Smith-Vidaurre G. 2017. warbleR: An R package to streamline analysis of animal acoustic signals. Methods Ecol Evol. 8:184–191.

Attali D. 2018. shinyjs: Easily improve the user experience of your Shiny apps in seconds.

Bache SM, Wickham H. 2014. magrittr: A forward-pipe operator for R.

Bioacoustics Research Program. 2014. Raven Pro: Interactive sound analysis software (version 1.5) [Computer software]. Ithaca, NY, USA: The Cornell Lab of Ornithology. Available from http://www.birds.cornell.edu/raven.

Chang W. 2018. shinythemes: Themes for Shiny.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. 2018. shiny: Web application framework for R.

Dowle M, Srinivasan A. 2018. data.table: Extension of `data.frame` (version 1.11.4).

Farabaugh SM, Brown ED, Dooling RJ. 1992. Analysis of warble song of the budgerigar *Melopsittacus undulatus*. Bioacoustics. 4:111–130.

Giorgino T. 2009. Computing and visualizing dynamic time warping alignments in R: The dtw package (version 1.20 - 1). J Stat Softw. 31:1–24.

892 Guerra JE, Cruz-Nieto J, Ortiz-Maciel SG, Wright TF. 2008. Limited geographic
893 variation in the vocalizations of the endangered thick-billed parrot: Implications
894 for conservation strategies. Condor. 110:639–647.

895 Holm S. 1979. A simple sequentially rejective multiple test procedure. Scand J
896 Stat. 6:65–70.

897 Humphries GRW, Buxton RT, Jones IL. 2018. Machine learning techniques for
898 quantifying geographic variation in Leach's storm-petrel (*Hydrobates*
899 *leucorhous*). In: Humphries GRW, Magness DR, Huettmann F, editors. Machine
900 Learning for Ecology and Sustainable Natural Resource Management. Cham,
901 Switzerland: Springer Nature. p. 295–312.

902 Jones ZM, Linder F. 2017. edarf: Exploratory data analysis using random forests
903 (version 1.1.1).

904 Keen S, Ross JC, Griffiths ET, Lanzone M, Farnsworth A. 2014. A comparison of
905 similarity-based approaches in the classification of flight calls of four species of
906 North American wood-warblers (Parulidae). Ecol Inform. 21:25–33.

907 Krijthe JH. 2015. Rtsne: t-Distributed stochastic neighbor embedding using
908 Barnes-Hut implementation (version 0.13).

909 Liaw A, Wiener M. 2002. Classification and regression by randomForest (version
910 4.6 - 14). R News. 2:18–22.

911 Ligges U, Krey S, Mersmann O, Schnackenberg S. 2018. tuneR: Analysis of
912 music and speech (version 1.3.3).

913 van der Maaten L. 2009. Learning a parametric embedding by preserving local
914 structure. Artif Intell Stat.:384–391.

915 van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. J Mach Learn
916 Res. 9:2579–2605.

917 Mantel N. 1967. The detection of disease clustering and a generalized regression
918 approach. Cancer Res. 27:209–220.

919 Martella MB, Bucher EH. 1990. Vocalizations of the monk parakeet. Bird Behav.
920 8:101–110.

921 Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. 2017. misc functions
922 of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU
923 Wien (version 1.6 - 8).

924 Nowicki S, Nelson DA. 1990. Defining natural categories in acoustic signals:
925 Comparison of three methods applied to “chick - a - dee” calls. Ethology. 86:89–
926 101.

927 Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin
928 PR, O’Hara RB, Simpson G, Solymos P, et al. 2018. vegan: Community ecology
929 package (version 2.5 - 2).

930 Oller Moreno S. 2018. facetscales: facet_grid with different scales per facet
931 (version 0.1.0).

932 Perrier V, Meyer F, Granjon D. 2019. shinyWidgets: Custom input widgets for
933 Shiny.

934 R Core Team. 2018. R: A language and environment for statistical computing. R
935 Foundation for Statistical Computing, Vienna, Austria.

936 Sarkar D. 2008. lattice: Multivariate data visualization with R (version 0.20 - 35).
937 New York: Springer.

938 Scrucca L, Fop M, Murphy TB, Raftery AE. 2017. mclust 5: Clustering,
939 classification and density estimation using Gaussian finite mixture models
940 (version 5.4.1). R J. 8:205–233.

941 Shamir L, Orlov N, Eckley DM, Macura T, Johnston J, Goldberg IG. 2008.
942 Wndchrm - an open source utility for biological image analysis. Source Code Biol
943 Med. 3:1–13.

944 Shamir L, Yerby C, Simpson R, Tyack P, Samarra F, Miller P, Wallin J. 2014.
945 Classification of large acoustic datasets using machine learning and
946 crowdsourcing: Application to whale calls. J Acoust Soc Am. 135:953–962.

947 Smith-Vidaurre G, Araya-Salas M, Wright. 2019. Data from: Individual signatures
948 outweigh social group identity in contact calls of a communally nesting parrot.
949 *Behavioral Ecology*. http://doi.org/10.5061/dryad.w6m905qkg.

950 Soifua B. 2018. A comparison of R, SAS, and Python implementations of random
951 forests. (Master's thesis) Logan, Utah Utah State Univ.

952 Solymos P, Zawadzki Z. 2018. pbapply: Adding progress bar to "apply" functions
953 (version 1.3 - 4).

954 Sueur J, Aubin T, Simonis C. 2008. seewave: A free modular tool for sound
955 analysis and synthesis (version 2.1.0). Bioacoustics. 18:213–226.

956 Torchiano M. 2018. effsize: Efficient effect size computation.

957 Valletta JJ, Torney C, Kings M, Thornton A, Madden J. 2017. Applications of
958 machine learning in animal behaviour studies. Anim Behav. 124:203–220.

959 Wei T, Simko V. 2017. corrplot: Visualization of a correlation matrix (version
960 0.84).

961 Wickham H. 2016a. ggplot2: Elegant graphics for data analysis (version 3.1.0).
962 Springer-Verlag New York.

963 Wickham H. 2016b. gtable: Arrange grobs in tables (version 0.2.0).

964 Wickham H. 2017. tidyverse: Easily install and load the "Tidyverse."

965 Wickham H. 2018. forcats: Tools for working with categorical variables (factors)
966 (version 0.3.0).

967 Wickham H, Francois R, Henry L, Muller K. 2018. dplyr: A grammar of data
968 manipulation (version 0.7.6).

969 Wing J, Kuhn M. 2018. caret: Classification and regression training (version 6.0 -
970 80).

971 Wright TF. 1996. Regional dialects in the contact call of a parrot. Proc R Soc
972 London, B. 263:867–872.

973 Wright TF, Dahlin CR, Salinas-Melgoza A. 2008. Stability and change in vocal
974 dialects of the yellow-naped amazon. Anim Behav. 76:1017–1027.

975 Yan Y. 2016. MLmetrics: Machine learning evaluation metrics (version 1.1.1).

976 Yu G. 2017. shadowtext: Shadow text grob and layer (version 0.0.2).