

Benchmarking nearest neighbor retrieval of zebra finch vocalizations across development

Working Paper**Author(s):**

[Tomka, Tomas](#) ; Hao, Xinyu; Miao, Aoxue; Lee, Kanghui; Basha, Maris; Reimann, Stefan; Zai, Anja T.; Hahnloser, Richard H.R.

Publication date:

2023-09-04

Permanent link:

<https://doi.org/10.3929/ethz-b-000655689>

Rights / license:

[Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International](#)

Originally published in:

bioRxiv, <https://doi.org/10.1101/2023.09.04.555475>

Benchmarking nearest neighbor retrieval of zebra finch vocalizations across development

Tomas Tomka,^{1,2} Xinyu Hao,^{1,3} Aoxue Miao,¹ Kanghwi Lee,^{1,2} Maris Basha,^{1,2} Stefan Reimann,¹ Anja T. Zai,^{1,2} and Richard H. R. Hahnloser^{1,2, a}

¹*Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, 8057, Switzerland*

²*Neuroscience Center Zurich, University of Zürich and ETH Zürich, Zürich, 8057, Switzerland*

³*School of Electrical and Information Engineering, Tianjin University, Tianjin, 300222, P.R. China*

(Dated: 28 August 2023)

1 Vocalizations are highly specialized motor gestures that regulate social interactions.
2 The reliable detection of vocalizations from raw streams of microphone data remains
3 an open problem even in research on widely studied animals such as the zebra finch.
4 A promising method for finding vocal samples from potentially few labelled examples
5 (templates) is nearest neighbor retrieval, but this method has never been extensively
6 tested on vocal segmentation tasks. We retrieve zebra finch vocalizations as neighbors
7 of each other in the sound spectrogram space. Based on merely 50 templates, we
8 find excellent retrieval performance in adults (F1 score of 0.93 ± 0.07) but not in
9 juveniles (F1 score of 0.64 ± 0.18), presumably due to the larger vocal variability of
10 the latter. The performance in juveniles improves when retrieval is based on fixed-
11 size template slices (F1 score of 0.72 ± 0.10) instead of entire templates. Among
12 the several distance metrics we tested such as the cosine and the Euclidean distance,
13 we find that the Spearman distance largely outperforms all others. We release our
14 expert-curated dataset of more than 50'000 zebra finch vocal segments, which will
15 enable training of data-hungry machine-learning approaches.

^arich@ini.ethz.ch

16 I. INTRODUCTION

17 In many species including humans, vocalizations play important roles during social be-
18 haviors such as aggressions, mating, breeding, and feeding. Inferring the functions of the
19 vocalizations is a challenging task where machine learning could be promising¹. The lon-
20 gitudinal study of vocalizations involves the challenging task of segmenting vocalizations
21 from background noise. In vocal learners such as the zebra finch, the vocal segmentation
22 task is particularly difficult, because the zebra finch vocal repertoire dramatically changes
23 over the course of development^{2,3}. Songs in young zebra finches start out as unstructured
24 subsongs that lack categorical structure and that gradually differentiate into distinct classes
25 of stereotyped syllables⁴. Zebra finches also produce less stereotyped calls⁵ with acoustic
26 features that vary depending on behavioral context^{5,6}.

27 To segment vocalizations in large vocal data sets, there is a growing literature on machine-
28 learning based systems⁷⁻¹⁰. However, these systems have only recently been emerging and
29 their potential is far from being fully explored. Foremost, for segmentation systems to
30 perform well, they must be trained and tested on datasets of precisely segmented vocaliza-
31 tions. But to our knowledge, only one such dataset is publicly available^{7,11} and it contains
32 merely 473 song syllables produced by a single adult male zebra finch and fails to include
33 all vocalization types, so represents a biased sample of vocal output. Entirely lacking are
34 public datasets of precisely segmented subsongs; a recent massive-data study on this impor-
35 tant developmental phase¹² simply ignores the segmentation problem and takes as proxy of
36 vocalizations all amplitude-thresholded sound segments, semi-automatically excluding false

37 positives in such a way to introduce false negatives (see Appendix). Unfortunately, ampli-
38 tude thresholding can create severe problems if the recording quality is low¹³, which only
39 emphasizes that this severe lack of training and test data forms a bottleneck for progress
40 in large-scale research on vocal development, and it calls for the creation of gold-standard
41 data sets.

42 One method for bootstrapping large vocal data sets from few precisely labelled samples is
43 nearest neighbor (NN) retrieval¹³. NN retrieval is a highly successful information retrieval
44 method¹⁴: it is used in tasks such as tagging images¹⁵, web mining¹⁶, recommendation
45 systems^{17,18}, and for inference in language models^{19,20}. Although the computational cost of
46 NN retrieval grows linearly with the number of templates and the size of the test recordings,
47 NN search scalability has improved massively since the popularization of graphics processing
48 units (GPUs) for parallel computing²¹ and with the advent of powerful approximate nearest
49 neighbor methods²²⁻²⁵. One of the advantages of NN retrieval over neural networks is that
50 NN retrieval uses few parameters and is interpretable²⁶⁻²⁸.

51 NN retrieval has been applied previously to the problem of birdsong analysis^{29,30}. Brooker
52 and colleagues used Pearson-correlation-based NN retrieval to benchmark commercially
53 available song detection software such as Monitor^{30,31}. Anderson and colleagues even ap-
54 plied a dynamic time-warping algorithm to find data frames in the search space based on
55 their minimal path-traversing distance to template frames²⁹. However, the sample sizes and
56 scopes of these works are very restrictive: they are based on single birds and unique distance
57 measures²⁹ and they excluded certain vocalization types from the analysis³⁰.

58 We set out to scale up NN retrieval methods for annotating and proofreading vocal segments.

59 The segmentation task we consider is to determine for each time point in a sound spectro-
60 gram (i.e., 16-ms sound interval) whether it contains a vocalization or not. We benchmark
61 the performance of our approaches on two data subsets of adult (Subset 1) and juvenile
62 (Subset 2) male zebra finch vocalizations. In our WHOLE approach, we use entire tem-
63 plates for NN retrieval, whereas, in the PART approach, we use fixed windows cut from the
64 templates. The PART approach allows the detection of vocalizations from conserved parts
65 and offers the practical benefit of yielding samples of fixed dimensionality. Among the many
66 spectrogram-based distance metrics we apply during retrieval, we find that the Spearman
67 distance outperforms all other metrics. We release our gold standard (GS) data set of more
68 than 50'000 annotations, taking care of eliminating false negatives, i.e. vocalizations buried
69 in noise that are easily missed by inattentive annotators.

70 **II. METHODS**

71 **A. Sound recordings and spectrograms**

72 We used data sets from four adult and four juvenile male zebra finches (each of the latter
73 was recorded at three different ages, see Table I for details). Recording was triggered by
74 vocalizations (or other sounds); thus, recordings are unevenly spaced in time depending on
75 the activity of the bird. Each recording/file contains vocalizations with some silence before
76 and after the vocalizations.

77 All adult birds (Subset 1) were raised in the animal facility of the University of Zurich.
78 During recording, birds were housed in single cages in custom made soundproof recording

79 chambers equipped with a wall microphone (Audio-Technica Pro42), and a loudspeaker.
80 The day/night cycle was 14/10 h. Vocalizations were saved using custom song-recording
81 software (Labview, National Instruments Inc.). Sounds were recorded with a wall-attached
82 microphone and were digitized at 32 kHz. We analyzed data from birds that had already
83 spent at least three days in their cage.
84 Data from juvenile birds (Subset 2) were randomly sampled from a publication³²: We ran-
85 domly selected 4 birds and from each bird we selected 3 days. Sounds in³² were recorded at
86 a sampling rate of 44.1 kHz.
87 We computed sound spectrograms by Fourier transforming sound segments $X_t \in \mathbb{R}^b$ of $b=$
88 512 samples. Accordingly, a spectrogram column $Y_t \in \mathbb{N}^b$ at time t is given by Eq. (1), where
89 Ω is a hamming window of length $b= 512$, and $\beta = 6.54$ for Subset 1 and $\beta = 4.93$ for
90 Subset 2 is a parameter that controls the dynamic range of the int8 down conversion.

$$Y_t = \text{int8}(\ln(|\text{FFT}(X_t\Omega)|)) \cdot 128/\beta) \quad (1)$$

91 The hop size Δt between adjacent Fourier segments is 128 samples corresponding to 4 ms
92 in adults. For distance computations, we removed low frequencies (0-688 Hz in adults and
93 0-947 Hz in juveniles) due to the large background noise in these ranges.

94 **B. Generation of gold-standard annotations**

95 From each day-long recording, we annotated a subset of data by randomly selecting a
96 set of files. We annotated vocal segments (not further classified into vocalization types)
97 with high temporal accuracy. To generate these gold-standard (GS) annotations, we used

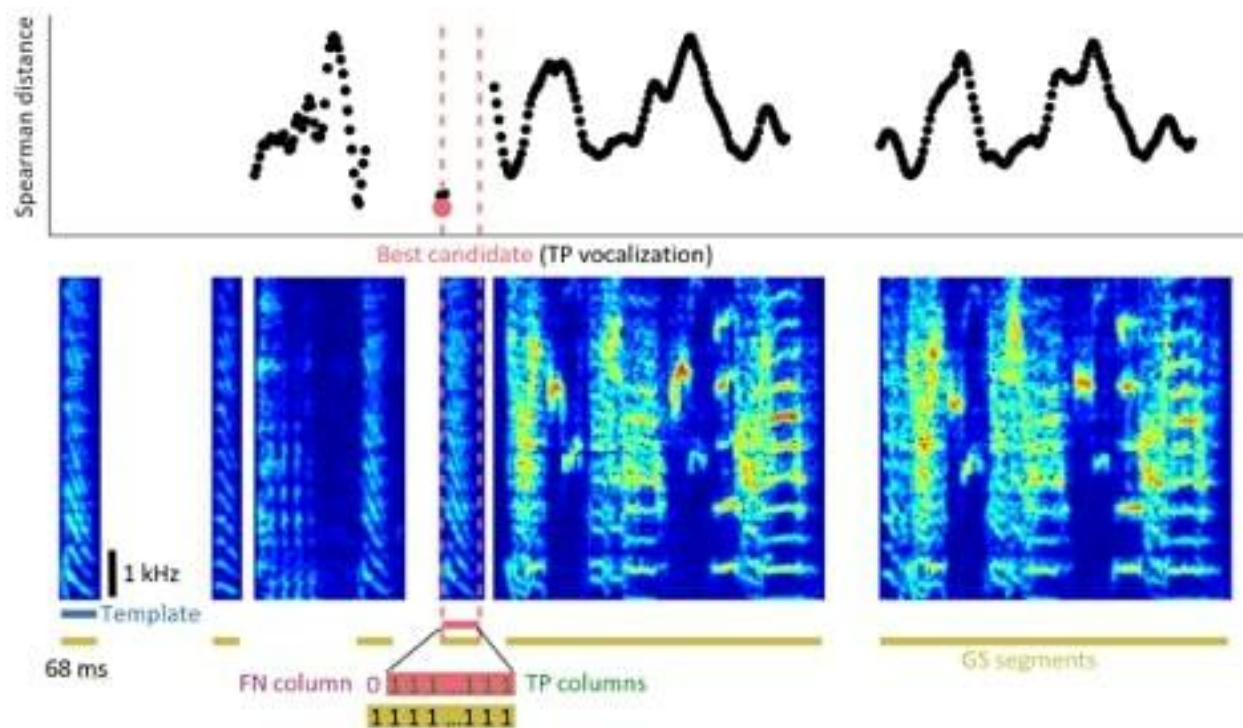
98 a semi-supervised segmentation method¹³, correcting poor segments and eliminating false
99 positives by visual inspection of spectrograms. To eliminate false negatives, the present NN
100 method was used with the cosine distance as metric. The GS dataset contains a label for
101 each spectrogram column (“1” for vocal, and “0” for non-vocal). A detailed annotation
102 protocol is provided in the “Supplementary information”.

103 **C. Nearest neighbor vocalization retrieval using gold-standard templates**

104 A simple approach to retrieving sounds segments corresponding to vocalizations is to
105 take a single template vocalization of (whole) duration τ and to compute spectrogram-
106 based distances to all candidate segments from the search space. Candidates are contained
107 in spectrogram windows of the same duration τ . The best candidate segment is the one
108 with minimal spectrogram-distance to the template and that does not temporally overlap
109 with the template, Fig. 1. To reduce computational cost, we restricted the search space to
110 non-silent periods (defined by thresholding the root-mean-squared audio signal) of duration
111 $\geq \tau$.

112 When many templates are given, we generalize this single-template procedure to many tem-
113 plates by iteratively retrieving the top segments one-by-one, as described in the following.

114



115

116 FIG. 1. **Template-based nearest-neighbor (NN) retrieval of vocal segments (WHOLE**
117 **approach)**. For an exemplary template (leftmost spectrogram) drawn from our gold-standard
118 (GS) dataset, we plot the (here Spearman) distance (top, dots aligned to candidate onsets) to
119 all candidate segments of the same duration within the search space (other spectrograms). The
120 best candidate (delimited by red dashed lines) is the one with minimal spectrogram-based distance
121 (red dot, top). With this procedure, segmentation errors can arise from mismatching segment
122 durations. Here, the best candidate starts one spectrogram column too late relative to the GS
123 segmentation, giving rise to a false negative (FN) spectrogram column (purple 0). Since this error
124 is within a reasonable tolerance (≤ 5 columns), we regard this vocal segment (red horizontal bar)
125 as containing a true positive (TP) vocalization.

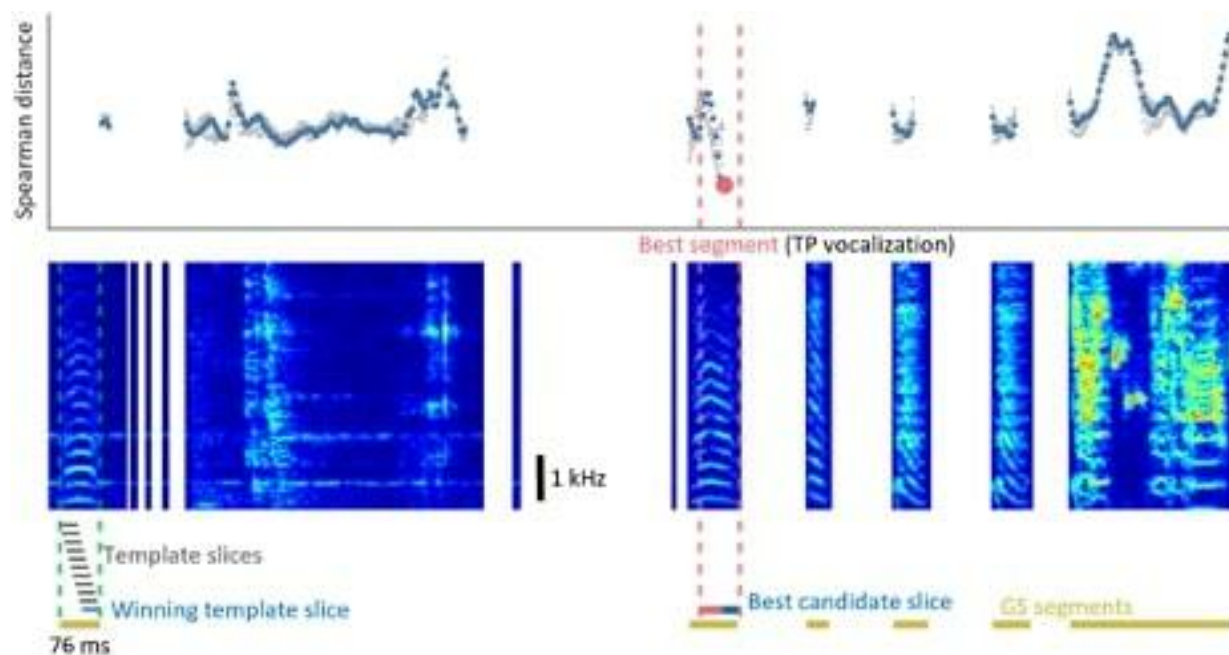
126 D. Vocalization retrieval using WHOLE approach

127 In the WHOLE approach (Fig. 2), we computed the spectrogram-based distances D_{ij} of
128 all template-candidate pairs. The distance D_{ij} represents the distance between the i -th
129 template ($i = 1, \dots, M$) and the j -th candidate in the search space. For a given template
130 i , the search space is given by the set of candidates of the same duration τ_i as the template.
131 After we computed all distance pairs, we identified the best candidate segment to any tem-
132 plate as the one with minimal distance, $\underset{i,j}{\operatorname{argmin}} D_{ij}$. After choosing the best segment, we
133 removed it from the search space, thereby also removing candidates that overlapped with
134 the best segment. Then we selected the next-best segment in an iterative procedure. By
135 iteratively selecting the segment with minimal distance to any template, we chose a very
136 greedy strategy of retrieving segments from the set of templates. In practice, we first com-
137 puted all pairwise distances and maintained an index of valid candidate-template pairs to
138 avoid re-computing any distances during the iterative procedure.
139 Because templates are of different durations τ_i , they might bias this retrieval process to short
140 templates. To address this possibility, we tested four different normalizations of distances:
141 no normalization, dividing distances D_{ij} by τ_i , by $\sqrt{\tau_i}$, or min-max normalizing them for
142 each template separately as in Eq.(2).

$$D_{ij}^{\text{norm}} = \frac{D_{ij} - \min_k D_{ik}}{\max_k D_{ik} - \min_k D_{ik}}. \quad (2)$$

143 E. Vocalization retrieval using PART approach

144 In the PART approach, we circumvent any duration-induced distance bias by slicing
145 each template into overlapping slices of w spectrogram columns (Fig. 2), where the integer
146 parameter w is shorter than a typical template. To any template i with duration $\tau_i < w$,
147 we appended a trailing zero-pad so that all templates had a duration of at least w . From
148 M templates, we obtained in total $n_w = \sum_i^M \text{floor}(\frac{\tau_i}{w})$ template slices. We then computed
149 all distance pairs D_{ij} between template slices and candidate slices. We then chose the best
150 candidate slice as the one with minimal distance to any of the n_w template slices. Based on
151 the best candidate slice, we selected the associated best segment as the sound interval with
152 the same relative timing as the template the slice was taken from (the onset and offset of the
153 best segment formed the same time lags to the slice as did the onset and offset of the sliced
154 template), Fig. 2. Thus, the best candidate segment was selected to be of equal duration
155 as the sliced template. There was one exception to this procedure: when the selected best
156 segment extended into a silent period, it was cropped.



157

158 **FIG. 2. Template-based NN retrieval of vocal segments (PART approach).** Shown is an
159 example template (delimited by green dashed lines, left) that we chopped into overlapping slices
160 (gray bars, below) of width w . For each of these slices, we computed the Spearman distances
161 (dots, top) to candidate slices. The winning template slice (thin blue bar, bottom) and the best
162 candidate slice (red dot, top; thick blue bar, bottom) are the ones with minimal distance to each
163 other. From this best candidate slice, we retrieved the best segment (delimited by dashed red
164 lines) as the sound interval that protrudes in the same way as the template relative to its winning
165 slice. Here, this candidate is a true positive, because its relative onset (+5 columns) and offset (+1
166 column) are both within the accepted tolerance (≤ 5 columns) of a GS segment.

167 **F. Spectrogram-based distance measures**

168 As metrics for distances D_{ij} , we tested the Euclidean, cosine, Jaccard, and Spearman met-
169 rics using the built-in MATLAB function `pdist2`. Additionally, for the WHOLE approach,
170 we evaluated earth mover’s distance (EMD) that measures the transport of sound-intensity
171 along a single spectrogram axis: either summing EMD distances row-wise (EMDr, transport
172 along the temporal axis) or summing column-wise (EMDc, transport along spectral axis).

173 **G. Performance evaluation**

174 We evaluated the retrieval performance of our NN approaches using scores based on time
175 bins and on sound segments:

- 176 • The time-bin based (or column-wise) score corresponds to the F1 score (the harmonic
177 mean of precision and recall) of the inferred labels of all spectrogram column relative
178 to the GS labels. Fig. 1 shows examples of true-positive and false-negative labels.
- 179 • The segment-wise or vocalization score (VocScore) is the F1 score of detected vo-
180 cal segments. A segment is considered a true-positive (TP) vocalization if both its
181 predicted onset and offset are within a temporal tolerance ϵ of the gold-standard val-
182 ues. This tolerance reflects the fact that even experts disagree on precise segment
183 boundaries. Here, we have chosen a generous tolerance of $\epsilon = 5$ spectrogram columns,
184 corresponding to a generous tolerance of 20 ms on Subset 1.

185 **III. RESULTS**

186 **A. A gold-standard (GS) dataset of juvenile and adult vocal segments**

187 From a small set of template vocalizations, we performed NN retrieval of vocal segments
188 (see Section II). We manually corrected the obtained segments to assemble a GS dataset of
189 53'326 vocalizations extracted from a total of 370 mins of data from zebra finches recorded
190 at different developmental stages (Table I). We share our guidelines for manual correction
191 that specify two decision boundaries we used to correct the segments: the decision whether
192 there is a short silent period (gap) between two vocalizations (Fig. 5), and the distinction
193 between vocal and non-vocal sounds (Fig. 6-7). In short, we advocate the definition of vocal
194 segments as tight intervals of contiguous vocal activity (no gaps) (see Appendix).

TABLE I: Dataset of zebra finch vocal segments across 4 developmental stages. The birds’ ages are specified in days-post-hatch (dph). The last four columns specify the duration of the annotated recording (including silence and noise), the number of annotated vocalizations, the fraction of time with vocal activity (“label imbalance”, vocal/total columns; perfect balance corresponds to 0.5), and the duration range of vocalizations, respectively. The Group column refers to the recording date, i.e., the number of days (20, 10, or 0) before birds learned their baseline (BL) song (Fig. 3c).

Developmental stage	Bird name	Sex	Hatch date	Age (dph)	Group	Annotated (mins)	Number of vocalizations	Label imbalance	Vocalization duration range (ms)
Adult (subset 1)	g17y2	male	14.4.2015	197		84.34	10050	0.4714	20-656
	g4p5	male	28.12.2012	115		104.18	26045	0.5155	16-300
	g19o3	male	13.11.2015	154		7.72	2045	0.4238	20-240
	g19o10	male	08.11.2015	198		7.68	1998	0.548	28-400
Juvenile (subset 2)	R3406	male	29.11.2011	35	-20BL	1.27	139	0.22	20-357
				45	-10BL	8.28	243	0.0486	9-377
				55	BL	39.42	2281	0.1077	12-372
	R3428	male	16.12.2011	39	-20BL	7.30	1316	0.2931	15-514
				49	-10BL	6.86	780	0.2496	12-418

Continued on next page

TABLE I – *Continued from previous page*

Developmental stage	Bird name	Sex	Hatch date	Age (dph)	Group	Annotated (mins)	Number of vocalizations	Label imbalance	Vocalization duration range (ms)
Juvenile (subset 2)	R3428	male	16.12.2011	59	BL	52.19	4026	0.1862	23-435
	R3549	male	17.02.2012	43	-20BL	7.33	781	0.2411	15-581
				53	-10BL	9.02	929	0.2209	15-438
				63	BL	10.52	1068	0.2372	12-343
				45	-20BL	11.67	728	0.1216	26-372
	R3625	male	13.04.2012	55	-10BL	7.23	534	0.1363	12-418
				65	BL	4.71	362	0.1575	15-293
All						370	53326	9-656	

195 To assess the annotation consistency, we asked a second expert to perform the same man-
196 ual correction of NN-retrieved segments on a subset of data (two adults and two juveniles).
197 We quantified expert disagreement by assessing the performance of Expert 2 relative to the
198 GS data (Expert 1) as a reference: While the F1 score was generally high across both subsets
199 (0.981 ± 0.014), the VocScore fluctuated more substantially (0.923 ± 0.046). A closer in-
200 spection revealed that the adult bird g19o3 produced pairs of rapidly following vocalizations
201 that Expert 2 interpreted as a single vocalization, resulting in a low VocScore (F1-Score:

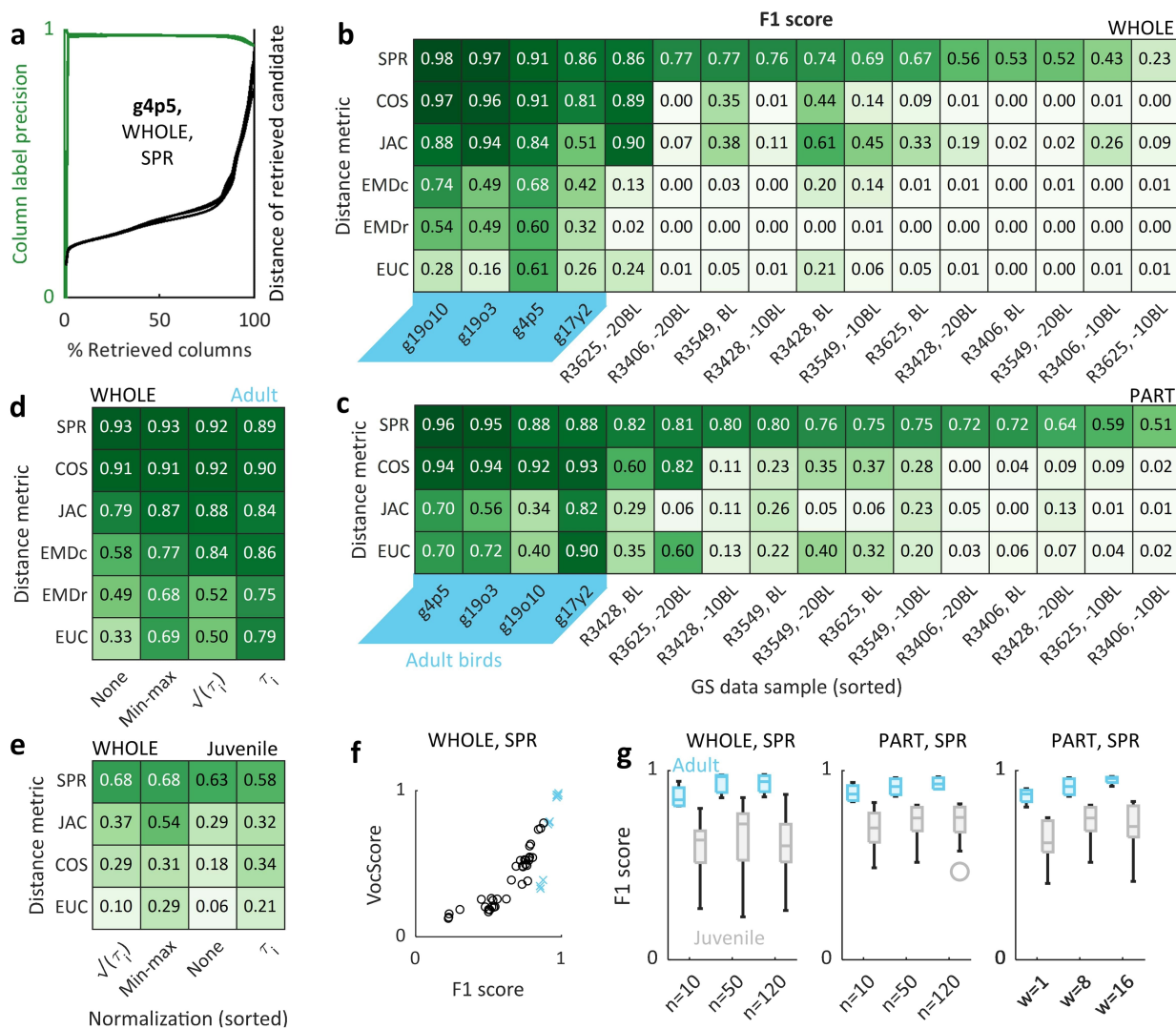
202 0.975, VocScore: 0.883), while bird g19o10 displayed no such confounding vocalization pair
203 (F1 score: 0.992, VocScore: 0.998).

204 **B. Performance of nearest neighbor retrieval**

205 We tested the two template-based vocal retrieval approaches (WHOLE and PART) on
206 our GS dataset. The NN distance of retrieved vocalizations increased monotonically with
207 increasing number of retrieved segments, as per definition (Fig. 3a, shown for three replicates
208 of 50 randomly selected templates). Less trivially, the precision of retrieved vocalizations
209 decreased with the number of retrieved vocalizations (Fig. 3a-9-10). We varied the used dis-
210 tance metric and the normalization strategy. We found that the Spearman distance metric
211 performed best, particularly in juveniles, while the Euclidean metric performed worst. In
212 juveniles also, the Jaccard metric performed better than the Cosine metric. In both adults
213 and juveniles, both EMDs performed poorly (Fig. 3b-e). In the following, we report the per-
214 formance of the Spearman metric in more detail. Using WHOLE, the Spearman distance
215 achieved an average F1 score of 0.93 ± 0.07 (range 0.86 to 0.98) for adults (Fig. 3b and
216 Fig. 3d, no normalization) and an F1 score of 0.63 ± 0.18 (range 0.23 to 0.86) for juveniles
217 (Fig. 3b and Fig. 3e, no normalization). Using PART, the performance increased for ju-
218 veniles (F1 score of 0.72 ± 0.10 , range 0.51 to 0.82) but decreased for adults (0.92 ± 0.04 ,
219 range 0.88 to 0.96), see Fig. 3c for each bird individually. This significant performance gap
220 between adults and juveniles that we observed for the Spearman metric was also true for
221 other metrics. The Cosine distance performed well on adults (F1-score range 0.97 to 0.81),
222 while on juveniles it yielded low scores. Distances such as the Euclidean distance and the two

223 Earth Mover distances performed significantly worse than the correlation-based distances
224 even in adults, while their respective F1 scores were close to zero in juveniles. In general,
225 distance metrics performed significantly better in adults than in juveniles. We normalized
226 distances in the WHOLE approach with four different strategies based on either duration
227 or sound amplitude (see Section II). For adults, not normalizing was among the best strate-
228 gies for the Spearman distance (though neither in adults nor juveniles, normalization had
229 a large impact) and it was the worst for Earth mover's, Jaccard, and Euclidean distances
230 (Fig. 3d). As expected, these latter distances benefit from division by the template dura-
231 tion to counteract the unequal dimensions of the competing candidates. The template-wise
232 min-max normalization worked well across distance metrics and GS data subsets (Fig. 3d,e).
233 Taken together, NN search performed best using the PART approach on juveniles and the
234 unnormalized WHOLE approach on adults. Across development, zebra finches can change
235 their songs to join or to separate adjacent vocalizations (Fig. 6). To quantify errors result-
236 ing from falsely joining or separating adjacent vocalizations, we used the VocScore. The
237 VocScore is very sensitive to segmentation errors occurring in between two vocalizations,
238 e.g., when a syllable gap is missed, the VocScore reports a long false-positive (FP) and
239 two short false negative (FN) vocalizations. Across both adults and juveniles, the VocScore
240 correlated with the F1 score (Fig. 3f) and the VocScore performance was quite variable
241 across datasets, which was due to some birds persistently producing hard-to-segment vocal-
242 ization pairs. The simpler F1 score of misclassified spectrogram columns was sensitive to
243 the number n of templates used, but surprisingly the F1 score barely improved from using
244 more than 50 templates (Fig. 3g). The F1 score also improved with increasing slice width

245 w (Fig. 3g), especially from the minimal width $w=1$ to $w = 8$. However, in juveniles, there
 246 was no additional improvement from increasing the slice width to $w = 16$ (Fig. 3g).

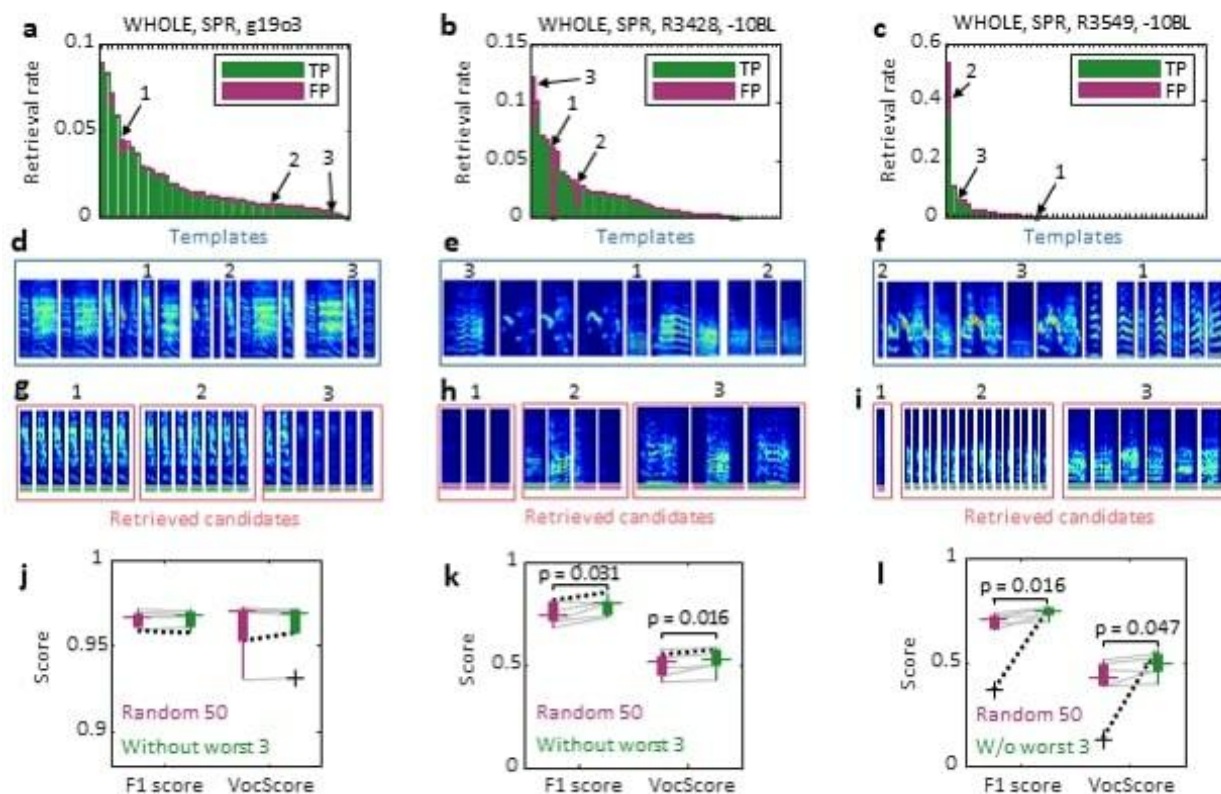


248 **FIG. 3. Performance of vocal segment retrieval for various distance metrics and normal-**
 249 **ization strategies.** (a) The column-wise precision (green) of vocal segments gradually declined
 250 (after initial fluctuation) with increasing number of retrieved segments. We retrieved a total of
 251 N - n segments ($n = 50$ templates, $N = 26045$ GS segments, bird g4p5), corresponding to theo-
 252 retical optimum of 100% of retrieved columns (x-axis). Three overlapping curves are shown for 3
 253 replicates of 50 randomly selected templates. (b,c) Mean F1 scores (from 3 replicates of 50 random

254 templates) across the dataset for different distance metrics, using the unnormalized WHOLE (b)
255 or PART (c) approach (slice $w=8$ columns). The tables are sorted along the rows and columns to
256 display the best performance on the top left. Abbreviations: SPR="Spearman", JAC="Jaccard",
257 COS="Cosine", EMDc="column-wise Earth mover's distance", EMDr="row-wise Earth mover's
258 distance", EUC="Euclidean". (d,e) Sorted tables of mean F1 scores (from b) of adults (d) and
259 of juveniles (e) for the WHOLE approach, shown for different normalization strategies. (f) The
260 relationship between F1 score and VocScore in adults (blue crosses) and juveniles (black circles),
261 computed for the Spearman distance and using the WHOLE approach (3 replicates per sample).
262 (g) Sensitivity analysis for the number of templates n and the slice width w , using the Spearman
263 distance.

264 To investigate whether the retrieval process is hampered by some detrimental templates
265 that excessively often retrieve false positives, we examined one retrieval replicate each in
266 three exemplary birds, an adult and two juveniles (Fig. 4). In both birds, we found that
267 the retrieval fractions were very non-uniform across the 50 templates (Fig. 4a-c, Figure
268 S6, S7). In the juveniles, there were a few templates that yielded excessively low retrieval
269 precision (large fraction of FPs). These detrimental templates had either background noises
270 (e.g., Fig. 4b, templates "1" and "2") or very faint harmonic extensions (e.g., Fig. 4b,
271 template "3"). To illustrate their shortcoming, we plotted the segments retrieved by the
272 three templates with the lowest retrieval precision in each bird (Fig. 4a-b, bottom row of
273 spectrograms). Removing the worst three templates (searching with 47 templates only) did
274 not increase performance in the adult (Fig. 4c), but slightly increased the performance in

275 the juvenile (Fig. 4d). This indicates that NN search can only marginally be improved by
 276 selecting representative and clean (noise-free) templates.



277

278 **FIG. 4. Retrieval performance is non-uniform across templates.** (a,b) For an example
 279 adult (a) and two juveniles (b,c), we sorted the 50 templates (from one replicate) by the fraction of
 280 segments they retrieved (summed TP and FP retrievals). (d,e,f) For each bird, example templates
 281 are shown including the worst three (numbered 1-3). (g,h,i) Example segments retrieved by the
 282 worst three templates in each bird. (j,k,l) Performance scores (6 replicates per bird) for the initial
 283 set of random 50 templates (purple box) and for the reduced set (green box) constructed by
 284 removing the worst 3 templates. A small but significant increase in both F1 score and VocScore
 285 is observed for the juveniles ($p < 0.05$, one-sided paired-sample Wilcoxon signed rank test). The

286 performance changes for the replicates in (a-i) are highlighted by black dotted lines (grey lines
287 indicate changes for the remaining 5 replicates).

288 IV. DISCUSSION

289 We have presented a simple and viable method for creating and proofreading of GS
290 datasets of animal vocalizations. Nearest neighbor retrieval is straightforward in its applica-
291 tion and is suitable both for extending manual annotations based on a few examples and for
292 proofreading existing datasets. We have used NN retrieval in a 2-step process of 1) detect-
293 ing vocalizations in raw sound recordings based on few labelled examples, and 2) systematic
294 screening the remaining data for false negative samples. We evaluated NN retrieval on vo-
295 calizations from individual birds including the notoriously challenging subsongs produced
296 during an early developmental phase. We benchmarked two NN variants and found that
297 adult vocalizations were better retrieved using whole templates (WHOLE approach, Fig. 1)
298 whereas juvenile vocalizations were better retrieved using template slices (PART approach,
299 Fig. 2). We found that as few as 50 templates were sufficient for reaching plateau perfor-
300 mance, which imposes a minimal requirement on the human effort for adopting this method.
301 In theory, NN retrieval can be performed with as little as one single positive example. In
302 practice, we recommend selecting clean templates and disregarding templates that contain
303 background noises or outlier features (Fig. 4), because otherwise the noise itself becomes a
304 target of NN retrieval. A good strategy might be to perform a two-stage search: first with
305 stereotyped templates, then with apparent outliers. The Spearman distance outperformed

306 the other tested metrics (Fig. 3) – especially on juvenile data. Surprisingly, the Euclidean
307 metric, often the first choice when comparing songbird vocalizations^{3,29,33,34}, exhibited the
308 overall worst performance. That the Spearman distance outperformed the Euclidean dis-
309 tance on both juveniles and adults suggests that commonly used analysis methods based on
310 the Euclidean distance^{3,33} could be improved simply by the use of Spearman distance. The
311 finding that correlation-based metrics (including Spearman and cosine distances) outper-
312 form the Euclidean and EMD distances emphasizes the importance of discounting for vocal
313 variability: Under the Euclidean and EMD metrics, a loud candidate vocalization will have
314 a large distance to its softer template. Variability of sound intensity can arise from varying
315 distances and directions of a bird to the microphone and so they should not affect retrieval.
316 In contrast, correlation-based metrics are invariant to global changes in signal intensity (or
317 loudness). Furthermore, correlation-based metrics work well with templates of different du-
318 rations since the correlation between two vectors does not scale with the vector dimension.
319 These results are in line with a general trend away from the Euclidean distance towards
320 correlation-based metrics: The advantage of Spearman distance over the cosine distance
321 is that the former captures non-linear monotonic relations^{35,36}. This property is generally
322 believed to contribute to the good performance of the Spearman distance in applications
323 as diverse as spam email detection³⁷ and indoor localization based on received Wi-Fi signal
324 strength³⁸. We see the strength of NN retrieval in proofreading the predictions generated by
325 other systems, in particular when labelled data are scarce. By contrast, when labelled data
326 are abundant, NN retrieval is unlikely going to be competitive with state-of-art approaches
327 for birdsong segmentation such as deep neural networks^{7,8}. The main disadvantage of NN

328 retrieval (e.g. compared to neural networks), is that the computational cost scales with the
329 number of labelled examples, although workarounds could be to sub-sample or summarize
330 the templates using for example k-means clustering. Very large datasets are amenable to
331 NN retrieval by virtue of powerful methods for approximative NN retrieval^{22–25}. Therefore,
332 there is no fundamental barrier for scaling up this method. We benchmarked NN retrieval
333 on vocal segmentation, which is a task that is feasible in both adults and juveniles and allows
334 for comparison of performance across age. In adults with their stereotyped repertoire, it is
335 possible to target retrieval to renditions of specific syllable types rather than any vocalization
336 from the repertoire. Coincidentally, we used such type-specific retrieval to generate the GS
337 annotations for adults. In practice, we found that best performance is achieved when first
338 searching for renditions of long vocalization types and then successively for shorter types.
339 Such a hierarchical retrieval strategy avoids confounds from repeated notes among syllables
340 in adult zebra finch song³⁹, which may also be the reason for the lower performance of PART
341 in adults compared to WHOLE. By contrast, the reason why for juveniles, PART seems to
342 work better than WHOLE could be that on a larger time scale juveniles have no repeating
343 vocal units — thus, if we model their vocalizations as random vectors then these are all far
344 from each other since in large spaces, random pairs of vectors tend to be orthogonal to each
345 other. Our retrieval approach (in particular the WHOLE approach) suffers from inflexibility
346 of segment durations, namely that the retrieved segments must exhibit the same durations
347 as the templates. Therefore, WHOLE will struggle to find the overall shortest vocalization
348 performed by an animal. One possible approach to overcome this limitation is to use dy-
349 namic time warping²⁹ as a means to create artificially short templates, thereby increasing

350 the number and diversity of templates. NN retrieval is attractive because it controls for
351 out-of-distribution detection with a well-defined and interpretable distance measure. NN
352 retrieval shifts the challenge of modeling vocalizations to the challenges of identifying a
353 good metric. We tested only a set of well-known metrics here, but in follow-up work it
354 may be worthwhile train custom metrics on the same retrieval task to learn to optimally
355 account for natural variability. Metrics can be learned from embeddings and the approach of
356 computing embeddings in a self-supervised manner⁴⁰ is getting more popular also in sound
357 processing⁴¹, in particular speech^{42,43}. The role of NN search we foresee in future work is to
358 assist in creation of vocal annotations and in proofreading automated annotations produced
359 by trained systems. One promising idea is to develop human-in-the loop iterative proce-
360 dures of labelling, training, searching, and fine-tuning of machine-learning systems. Our
361 expert-curated dataset of annotated individual vocal repertoires counts more than 50'000
362 vocalizations from 8 zebra finches. We release this dataset so that data-hungry deep learn-
363 ing systems for large scale vocal analysis can be trained and evaluated. To make our work
364 reproducible, we also share our segmentation guidelines as illustrations of the manual an-
365 notation challenges and of our chosen decision boundaries (see Appendix). We hope that
366 our annotation guidelines will help to standardize vocal annotation tasks and so promote
367 comparative work across species.

368 **DATA AVAILABILITY**

369 We will release our dataset (Table I) upon publication of our work in a peer-reviewed
370 journal.

371 **CONFLICT OF INTEREST**

372 The authors declare that the research was conducted in the absence of any commercial
373 or financial relationships that could be construed as a potential conflict of interest.

374 **AUTHOR CONTRIBUTION**

375 RHRH, TT, SR, and XH contributed to the conceptualization of the study. ATZ con-
376 ducted experiments of Subset 1. TT, RHRH, XH, and AM contributed to data annotation.
377 TT and ATZ curated the dataset for release. TT and RHRH implemented the retrieval
378 algorithms. TT, RHRH, XH, SR, and KL were involved in data analysis. TT, ATZ and
379 RHRH wrote the manuscript. SR, AM, KL and MB provided feedback on the manuscript.

380 **ACKNOWLEDGEMENT**

381 We thank Dina Lipkind for making her previously published data (Subset 1) available
382 for post-hoc analyses. Experimental procedures involved in Subset 1 of the GS dataset were
383 approved by the Veterinary Office of the Canton of Zurich.

384 **FUNDING**

385 This study was partly funded by the Swiss National Science Foundation (Grant 31003A_182638;
386 and the NCCR Evolving Language, Agreement No. 51NF40_180888) and by the China
387 Scholarship Council (Grant No. 202006250099).

388 **REFERENCES**

389 ¹C. Rutz, M. Bronstein, A. Raskin, S. C. Vernes, K. Zacarian, and D. E. Blasi, “Us-
390 ing machine learning to decode animal communication,” *Science* **381**(6654), 152–155
391 (2023) <https://www.science.org/doi/abs/10.1126/science.adg7314> doi: [10.1126/
392 science.adg7314](https://doi.org/10.1126/science.adg7314).

393 ²O. Tchernichovski, P. Mitra, T. Lints, and F. Nottebohm, “Dynamics of the vocal imitation
394 process: how a zebra finch learns its song.’,” *Science* **291**(5513), 2564–2569, doi: [10.1126/
395 science.1058522](https://doi.org/10.1126/science.1058522).

396 ³S. Kollmorgen, R. Hahnloser, and V. Mante, “Nearest neighbours reveal fast and
397 slow components of motor learning.’,” *Nature* **577**(7791), 526–530, doi: [10.1038/
398 s41586-019-1892-x](https://doi.org/10.1038/s41586-019-1892-x).

399 ⁴D. Lipkind, A. Geambasu, and C. Levelt, “The development of structured vocalizations
400 in songbirds and humans: A comparative analysis.’,” *Top Cogn Sci* **12**(3), 894–909, doi:
401 [10.1111/tops.12414](https://doi.org/10.1111/tops.12414).

402 ⁵J. Elie and F. Theunissen, “The vocal repertoire of the domesticated zebra finch: a data-
403 driven approach to decipher the information-bearing acoustic features of communication
404 signals.’,” *Anim Cogn* **19**(2), 285–315, doi: [10.1007/s10071-015-0933-6](https://doi.org/10.1007/s10071-015-0933-6).

405 ⁶E. Perez, J. Elie, C. Soulage, H. Soula, N. Mathevon, and C. Vignal, “The acoustic
406 expression of stress in a songbird: does corticosterone drive isolation-induced modifications
407 of zebra finch calls?’,” *Horm Behav* **61**(4), 573–581, doi: [10.1016/j.yhbeh.2012.02.004](https://doi.org/10.1016/j.yhbeh.2012.02.004).

408 ⁷E. Steinfath, A. Palacios-Muñoz, J. Rottschäfer, D. Yuezak, and J. Clemens, “Fast and
409 accurate annotation of acoustic signals with deep neural networks.’,” eLife **10** doi: [10.7554/eLife.68837](https://doi.org/10.7554/eLife.68837).

411 ⁸Y. Cohen, D. Nicholson, A. Sanchioni, E. Mallaber, V. Skidanova, and T. Gardner, “Auto-
412 mated annotation of birdsong with a neural network that segments spectrograms.’,” eLife
413 **11** doi: [10.7554/eLife.63853](https://doi.org/10.7554/eLife.63853).

414 ⁹D. Baggi, M. Premoli, A. Gnutti, S. Bonini, R. Leonardi, M. Memo, and P. Migliorati,
415 “Extended performance analysis of deep-learning algorithms for mice vocalization segmen-
416 tation,” Scientific Reports **13** (2023) doi: [10.1038/s41598-023-38186-7](https://doi.org/10.1038/s41598-023-38186-7).

417 ¹⁰D. Pessoa, L. Petrella, P. Martins, M. Castelo-Branco, and C. Teixeira, “Automatic seg-
418 mentation and classification of mice ultrasonic vocalizations,” The Journal of the Acous-
419 tical Society of America **152**(1), 266–280 (2022) <https://doi.org/10.1121/10.0012350>
420 doi: [10.1121/10.0012350](https://doi.org/10.1121/10.0012350).

421 ¹¹J. Clemens, “Zebra finch - train and test data’,” GRO.data **V1** <https://doi.org/10.25625/ZXJJJY>
422 <https://doi.org/10.25625/ZXJJJY> doi: [10.25625/ZXJJJY](https://doi.org/10.25625/ZXJJJY) accessed:.

423 ¹²S. Brudner, J. Pearson, and R. Mooney, “Juvenile zebra finch syllables for data-driven
424 analysis of development” (2022), <https://doi.org/10.7924/r4j38x43h>, doi: [10.7924/](https://doi.org/10.7924/r4j38x43h)
425 [r4j38x43h](https://doi.org/10.7924/r4j38x43h).

426 ¹³C. Lorenz, X. Hao, T. Tomka, L. Rüttimann, and R. Hahnloser, “Interactive extraction
427 of diverse vocal units from a planar embedding without the need for prior sound seg-
428 mentation’,” Frontiers in Bioinformatics **2** <https://www.frontiersin.org/articles/>

429 [10.3389/fbinf.2022.966066](https://doi.org/10.3389/fbinf.2022.966066) accessed:.

430 ¹⁴T. Cover and P. Hart, “Nearest neighbor pattern classification’,” *IEEE Trans. Inform.*
431 *Theory* **13**(1), 21–27, doi: [10.1109/TIT.1967.1053964](https://doi.org/10.1109/TIT.1967.1053964).

432 ¹⁵M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, “Tagprop: Discriminative metric
433 learning in nearest neighbor models for image auto-annotation’,” in *2009 IEEE 12th Inter-*
434 *national Conference on Computer Vision, IEEE*, p. 309–316, doi: [10.1109/ICCV.2009.](https://doi.org/10.1109/ICCV.2009.5459266)
435 [5459266](https://doi.org/10.1109/ICCV.2009.5459266).

436 ¹⁶V. Bijalwan, P. Kumari, J. Espada, V. Semwal, and V. Kumar, “Knn based machine learn-
437 ing approach for text and document mining,” *International Journal of Database Theory*
438 *and Application* **7** (2014) doi: [10.14257/ijdta.2014.7.1.06](https://doi.org/10.14257/ijdta.2014.7.1.06).

439 ¹⁷D. Adeniyi, Z. Wei, and Y. Yongquan, “Automated web usage data mining and recommen-
440 dation system using k-nearest neighbor (knn) classification method’,” *Applied Computing*
441 *and Informatics* **12**(1), 90–108, doi: [10.1016/j.aci.2014.10.001](https://doi.org/10.1016/j.aci.2014.10.001).

442 ¹⁸T. Anwar, U. Vijayasundaram, M. Hussain, and M. Pantula, “Collaborative filtering
443 and knn based recommendation to overcome cold start and sparsity issues: A com-
444 parative analysis,” *Multimedia Tools and Applications* **81**, 1–19 (2022) doi: [10.1007/](https://doi.org/10.1007/s11042-021-11883-z)
445 [s11042-021-11883-z](https://doi.org/10.1007/s11042-021-11883-z).

446 ¹⁹U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, “Generalization
447 through memorization: Nearest neighbor language models”’, doi: [10.48550/arxiv.1911.](https://doi.org/10.48550/arxiv.1911.00172)
448 [00172](https://doi.org/10.48550/arxiv.1911.00172). arXiv,.

449 ²⁰J. He, G. Neubig, and T. Berg-Kirkpatrick, “Efficient nearest neighbor language models,”
450 CoRR **abs/2109.04212** (2021) <https://arxiv.org/abs/2109.04212>.

451 ²¹V. Garcia, E. Debreuve, and M. Barlaud, “Fast k nearest neighbor search using gpu’,” in
452 *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*
453 *Workshops, IEEE*, p. 1–6, doi: [10.1109/CVPRW.2008.4563100](https://doi.org/10.1109/CVPRW.2008.4563100).

454 ²²A. Becker, L. Ducas, N. Gama, and T. Laarhoven, “New directions in nearest neighbor
455 searching with applications to lattice sieving’,” in *Proceedings of the Twenty-Seventh An-*
456 *nual ACM-SIAM Symposium on Discrete Algorithms*, edited by R. Krauthgamer, Society
457 for Industrial and Applied Mathematics, Philadelphia, PA, p. 10–24, doi: [10.1137/1.](https://doi.org/10.1137/1.9781611974331.ch2)
458 [9781611974331.ch2](https://doi.org/10.1137/1.9781611974331.ch2).

459 ²³M. Muja and D. Lowe, “Fast approximate nearest neighbors with automatic algorithm
460 configuration’,” in *Proceedings of the Fourth International Conference on Computer Vi-*
461 *sion Theory and Applications, SciTePress*, Science and and Technology Publications, p.
462 331–340, doi: [10.5220/0001787803310340](https://doi.org/10.5220/0001787803310340).

463 ²⁴A. Andoni and P. Indyk, “Near-optimal hashing algorithms for approximate nearest
464 neighbor in high dimensions’,” *Commun ACM* **51**(1), 117–122, doi: [10.1145/1327452.](https://doi.org/10.1145/1327452.1327494)
465 [1327494](https://doi.org/10.1145/1327452.1327494).

466 ²⁵P. Indyk and R. Motwani, “Approximate nearest neighbors: Towards removing the curse
467 of dimensionality’,” in *Proceedings of the thirtieth annual ACM symposium on Theory*
468 *of computing - STOC '98*, ACM Press, New York, New York, USA, p. 604–613, doi:
469 [10.1145/276698.276876](https://doi.org/10.1145/276698.276876).

470 ²⁶B. Mitra and N. Craswell, “Neural models for information retrieval” , doi: [10.48550/](https://doi.org/10.48550/arxiv.1705.01509)
471 [arxiv.1705.01509](https://doi.org/10.48550/arxiv.1705.01509). arXiv,.

472 ²⁷J. Guo, “A deep look into neural ranking models for information retrieval’,” *Information*
473 *Processing Management* 102067, doi: [10.1016/j.ipm.2019.102067](https://doi.org/10.1016/j.ipm.2019.102067).

474 ²⁸X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang, “Representation learning using
475 multi-task deep neural networks for semantic classification and information retrieval’,” in
476 *Proceedings of the 2015 Conference of the North American Chapter of the Association for*
477 *Computational Linguistics: Human Language Technologies*, Association for Computational
478 Linguistics, Stroudsburg, PA, USA, p. 912–921, doi: [10.3115/v1/N15-1092](https://doi.org/10.3115/v1/N15-1092).

479 ²⁹S. Anderson, A. Dave, and D. Margoliash, “Template-based automatic recognition of bird-
480 song syllables from continuous recordings.’,” *J Acoust Soc Am* **100**(2 Pt 1), 1209–1219,
481 doi: [10.1121/1.415968](https://doi.org/10.1121/1.415968).

482 ³⁰S. Brooker, P. Stephens, M. Whittingham, and S. Willis, “Automated detection and clas-
483 sification of birdsong: An ensemble approach’,” *Ecological Indicators* **117**, 106609, doi:
484 [10.1016/j.ecolind.2020.106609](https://doi.org/10.1016/j.ecolind.2020.106609).

485 ³¹J. Katz, S. Hafner, and T. Donovan, “Tools for automated acoustic monitoring within the r
486 package monitor’,” *Bioacoustics* **25**(2), 197–210, doi: [10.1080/09524622.2016.1138415](https://doi.org/10.1080/09524622.2016.1138415).

487 ³²D. Lipkind, A. Zai, A. Hanuschkin, G. Marcus, O. Tchernichovski, and R. Hahnloser,
488 “Songbirds work around computational complexity by learning song vocabulary indepen-
489 dently of sequence.’,” *Nat Commun* **8**(1), 1247, doi: [10.1038/s41467-017-01436-0](https://doi.org/10.1038/s41467-017-01436-0).

- 490 ³³O. Tchernichovski, F. Nottebohm, C. Ho, B. Pesaran, and P. Mitra, “A procedure for
491 an automated measurement of song similarity.’,” *Anim Behav* **59**(6), 1167–1176, doi: [10.](https://doi.org/10.1006/anbe.1999.1416)
492 [1006/anbe.1999.1416](https://doi.org/10.1006/anbe.1999.1416).
- 493 ³⁴T. Sainburg, M. Thielk, and T. Gentner, “Finding, visualizing, and quantifying latent
494 structure across diverse animal vocal repertoires.’,” *PLoS Comput Biol* **16**(10), 1008228,
495 doi: [10.1371/journal.pcbi.1008228](https://doi.org/10.1371/journal.pcbi.1008228).
- 496 ³⁵C. Spearman, ““footrule” for measuring correlation’,” *British Journal of Psychology* **2**(1),
497 89–108, doi: [10.1111/j.2044-8295.1906.tb00174.x](https://doi.org/10.1111/j.2044-8295.1906.tb00174.x).
- 498 ³⁶L. Kaufman and P. Rousseeuw, eds., *Finding Groups in Data: An Introduction to Cluster*
499 *Analysis* (John Wiley Sons, Inc, Hoboken, NJ, USA).
- 500 ³⁷A. Sharma and A. Suryawanshi, “A novel method for detecting spam email using knn
501 classification with spearman correlation as distance measure’,” *IJCA* **136**(6), 28–35, doi:
502 [10.5120/ijca2016908471](https://doi.org/10.5120/ijca2016908471).
- 503 ³⁸Y. Xie, Y. Wang, A. Nallanathan, and L. Wang, “An improved k-nearest-neighbor in-
504 door localization method based on spearman distance’,” *IEEE Signal Process Lett* **23**(3),
505 351–355, doi: [10.1109/LSP.2016.2519607](https://doi.org/10.1109/LSP.2016.2519607).
- 506 ³⁹A. Vyssotski, V. Anisimov, A. Latanov, and R. Hahnloser, “Formation of hierarchical
507 network of vocalizations in songbird groups’,” *Society for Neuroscience* **07** [http://www.](http://www.abstractsonline.com/pp8/index.html#!/4071/presentation/2570)
508 [abstractsonline.com/pp8/index.html#!/4071/presentation/2570](http://www.abstractsonline.com/pp8/index.html#!/4071/presentation/2570) accessed: Jun.
- 509 ⁴⁰J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidi-
510 rectional transformers for language understanding’,” in *Proceedings of the 2019 Confer-*

511 *ence of the North*, Association for Computational Linguistics, Minneapolis, Minnesota, p.
512 4171–4186, doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).

513 ⁴¹A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-
514 supervised learning of speech representations”’, doi: [10.48550/ARXIV.2006.11477](https://doi.org/10.48550/ARXIV.2006.11477).

515 ⁴²A. Radford, J. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust speech
516 recognition via large-scale weak supervision,” in *Proceedings of the 40th International
517 Conference on Machine Learning, PMLR*, p. 28492–28518, [https://proceedings.mlr.
518 press/v202/radford23a.html](https://proceedings.mlr.press/v202/radford23a.html), accessed:.

519 ⁴³V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni,
520 A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and
521 M. Auli, “Scaling speech technology to 1,000+ languages” (2023).

522 ⁴⁴A. Kershenbaum, D. T. Blumstein, M. A. Roch, Akçay, G. Backus, M. A. Bee, K. Bohn,
523 Y. Cao, G. Carter, C. Cäsar, M. Coen, S. L. DeRuiter, L. Doyle, S. Edelman, R. Ferrer-i
524 Cancho, T. M. Freeberg, E. C. Garland, M. Gustison, H. E. Harley, C. Huetz, M. Hughes,
525 J. Hyland Bruno, A. Ilany, D. Z. Jin, M. Johnson, C. Ju, J. Karnowski, B. Lohr, M. B.
526 Manser, B. McCowan, E. Mercado III, P. M. Narins, A. Piel, M. Rice, R. Salmi, K. Sasa-
527 hara, L. Sayigh, Y. Shiu, C. Taylor, E. E. Vallejo, S. Waller, and V. Zamora-Gutierrez,
528 “Acoustic sequences in non-human animals: a tutorial review and prospectus,” *Bio-
529 logical Reviews* **91**(1), 13–52 (2016) [https://onlinelibrary.wiley.com/doi/abs/10.
530 1111/brv.12160](https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12160) doi: <https://doi.org/10.1111/brv.12160>.

531 ⁴⁵M. D. Hauser, N. Chomsky, and W. T. Fitch, “The faculty of language: What is it, who
532 has it, and how did it evolve?,” *Science* **298**(5598), 1569–1579 (2002) [https://doi.org/](https://doi.org/10.1126/science.298.5598.1569)
533 [10.1126/science.298.5598.1569](https://doi.org/10.1126/science.298.5598.1569) doi: [10.1126/science.298.5598.1569](https://doi.org/10.1126/science.298.5598.1569).

534 ⁴⁶C. Slobodchikoff, J. Kiriazis, C. Fischer, and E. Creef, “Semantic information distin-
535 guishing individual predators in the alarm calls of gunnison's prairie dogs,” *Animal Be-*
536 *haviour* **42**(5), 713–719 (1991) [https://doi.org/10.1016/s0003-3472\(05\)80117-4](https://doi.org/10.1016/s0003-3472(05)80117-4) doi:
537 [10.1016/s0003-3472\(05\)80117-4](https://doi.org/10.1016/s0003-3472(05)80117-4).

538 ⁴⁷W. P. Dittus, “Toque macaque food calls: Semantic communication concerning food dis-
539 tribution in the environment,” *Animal Behaviour* **32**(2), 470–477 (1984) [https://doi.](https://doi.org/10.1016/s0003-3472(84)80283-3)
540 [org/10.1016/s0003-3472\(84\)80283-3](https://doi.org/10.1016/s0003-3472(84)80283-3) doi: [10.1016/s0003-3472\(84\)80283-3](https://doi.org/10.1016/s0003-3472(84)80283-3).

541 ⁴⁸M. D. HAUSER, “Functional referents and acoustic similarity: field playback experiments
542 with rhesus monkeys,” *Animal Behaviour* **55**(6), 1647–1658 (1998) [https://doi.org/10.](https://doi.org/10.1006/anbe.1997.0712)
543 [1006/anbe.1997.0712](https://doi.org/10.1006/anbe.1997.0712) doi: [10.1006/anbe.1997.0712](https://doi.org/10.1006/anbe.1997.0712).

544 ⁴⁹R. M. Seyfarth, D. L. Cheney, and P. Marler, “Monkey responses to three different
545 alarm calls: Evidence of predator classification and semantic communication,” *Science*
546 **210**(4471), 801–803 (1980) <https://doi.org/10.1126/science.7433999> doi: [10.1126/](https://doi.org/10.1126/science.7433999)
547 [science.7433999](https://doi.org/10.1126/science.7433999).

548 ⁵⁰J. FISCHER, “Barbary macaques categorize shrill barks into two call types,” *Animal*
549 *Behaviour* **55**(4), 799–807 (1998) <https://doi.org/10.1006/anbe.1997.0663> doi: [10.](https://doi.org/10.1006/anbe.1997.0663)
550 [1006/anbe.1997.0663](https://doi.org/10.1006/anbe.1997.0663).

- 551 ⁵¹S. Gouzoules, H. Gouzoules, and P. Marler, “Rhesus monkey (*macaca mulatta*) screams:
552 Representational signalling in the recruitment of agonistic aid,” *Animal Behaviour*
553 **32**(1), 182–193 (1984) [https://doi.org/10.1016/s0003-3472\(84\)80336-x](https://doi.org/10.1016/s0003-3472(84)80336-x) doi: 10.
554 [1016/s0003-3472\(84\)80336-x](https://doi.org/10.1016/s0003-3472(84)80336-x).
- 555 ⁵²K. Zuberbühler, D. L. Cheney, and R. M. Seyfarth, “Conceptual semantics in a nonhuman
556 primate.,” *Journal of Comparative Psychology* **113**(1), 33–42 (1999) [https://doi.org/](https://doi.org/10.1037/0735-7036.113.1.33)
557 [10.1037/0735-7036.113.1.33](https://doi.org/10.1037/0735-7036.113.1.33) doi: [10.1037/0735-7036.113.1.33](https://doi.org/10.1037/0735-7036.113.1.33).
- 558 ⁵³P. Marler, A. Dufty, and R. Pickert, “Vocal communication in the domestic chicken:
559 II. is a sender sensitive to the presence and nature of a receiver?,” *Animal Behaviour*
560 **34**, 194–198 (1986) [https://doi.org/10.1016/](https://doi.org/10.1016/0003-3472(86)90023-0)
561 [0003-3472\(86\)90023-0](https://doi.org/10.1016/0003-3472(86)90023-0) doi: [10.1016/](https://doi.org/10.1016/0003-3472(86)90023-0)
- 562 ⁵⁴T. N. Suzuki, “Semantic communication in birds: evidence from field research over the past
563 two decades,” *Ecological Research* **31**(3), 307–319 (2016) [https://doi.org/10.1007/](https://doi.org/10.1007/s11284-016-1339-x)
564 [s11284-016-1339-x](https://doi.org/10.1007/s11284-016-1339-x) doi: [10.1007/s11284-016-1339-x](https://doi.org/10.1007/s11284-016-1339-x).
- 565 ⁵⁵S. A. Gill and A. M.-K. Bierema, “On the meaning of alarm calls: A review of functional
566 reference in avian alarm calling,” *Ethology* **119**(6), 449–461 (2013) [https://doi.org/10.](https://doi.org/10.1111/eth.12097)
567 [1111/eth.12097](https://doi.org/10.1111/eth.12097) doi: [10.1111/eth.12097](https://doi.org/10.1111/eth.12097).
- 568 ⁵⁶T. Sainburg and T. Q. Gentner, “Toward a computational neuroethology of vocal com-
569 munication: From bioacoustics to neurophysiology, emerging tools and future directions,”
570 *Frontiers in Behavioral Neuroscience* **15** (2021) [https://doi.org/10.3389/fnbeh.2021.](https://doi.org/10.3389/fnbeh.2021.811737)
571 [811737](https://doi.org/10.3389/fnbeh.2021.811737) doi: [10.3389/fnbeh.2021.811737](https://doi.org/10.3389/fnbeh.2021.811737).

572 ⁵⁷F. Goller and M. A. Daley, “Novel motor gestures for phonation during inspiration enhance
573 the acoustic complexity of birdsong,” *Proceedings of the Royal Society of London. Series B:
574 Biological Sciences* **268**(1483), 2301–2305 (2001) [https://doi.org/10.1098/rspb.2001.
575 1805](https://doi.org/10.1098/rspb.2001.1805) doi: [10.1098/rspb.2001.1805](https://doi.org/10.1098/rspb.2001.1805).

576 ⁵⁸R. Suzuki, J. R. Buck, and P. L. Tyack, “Information entropy of humpback whale songs,”
577 *The Journal of the Acoustical Society of America* **119**(3), 1849–1866 (2006) [https://doi.
578 org/10.1121/1.2161827](https://doi.org/10.1121/1.2161827) doi: [10.1121/1.2161827](https://doi.org/10.1121/1.2161827).

579 ⁵⁹T. Mizuhara and K. Okanoya, “Do songbirds hear songs syllable by syllable?,” *Behavioural
580 Processes* **174**, 104089 (2020) <https://doi.org/10.1016/j.beproc.2020.104089> doi:
581 [10.1016/j.beproc.2020.104089](https://doi.org/10.1016/j.beproc.2020.104089).

582 **APPENDIX**

583 **1. Vocal segmentation conventions for microphone recordings of single birds**

584 Vocal signals tend to arise from discrete acoustic units, which is a characteristic shared
585 across the polymorphic landscape of vocalizing species^{44,45}. Animal studies in monkeys, dogs,
586 chicken, and songbirds have shown that animal calls can be used to communicate semantic
587 meaningful information such as detection of predators, discovery of food, or attraction of
588 mates^{46–55}. Nevertheless, the functions of animal vocalizations are generally unknown for
589 most calls and species^{44,56}. To advance our understanding of vocal communication in ani-
590 mals, we need to study large and well-annotated data sets. Here we address the problem of
591 how to segment audio recordings of a given species. The segmentation problem is to distin-

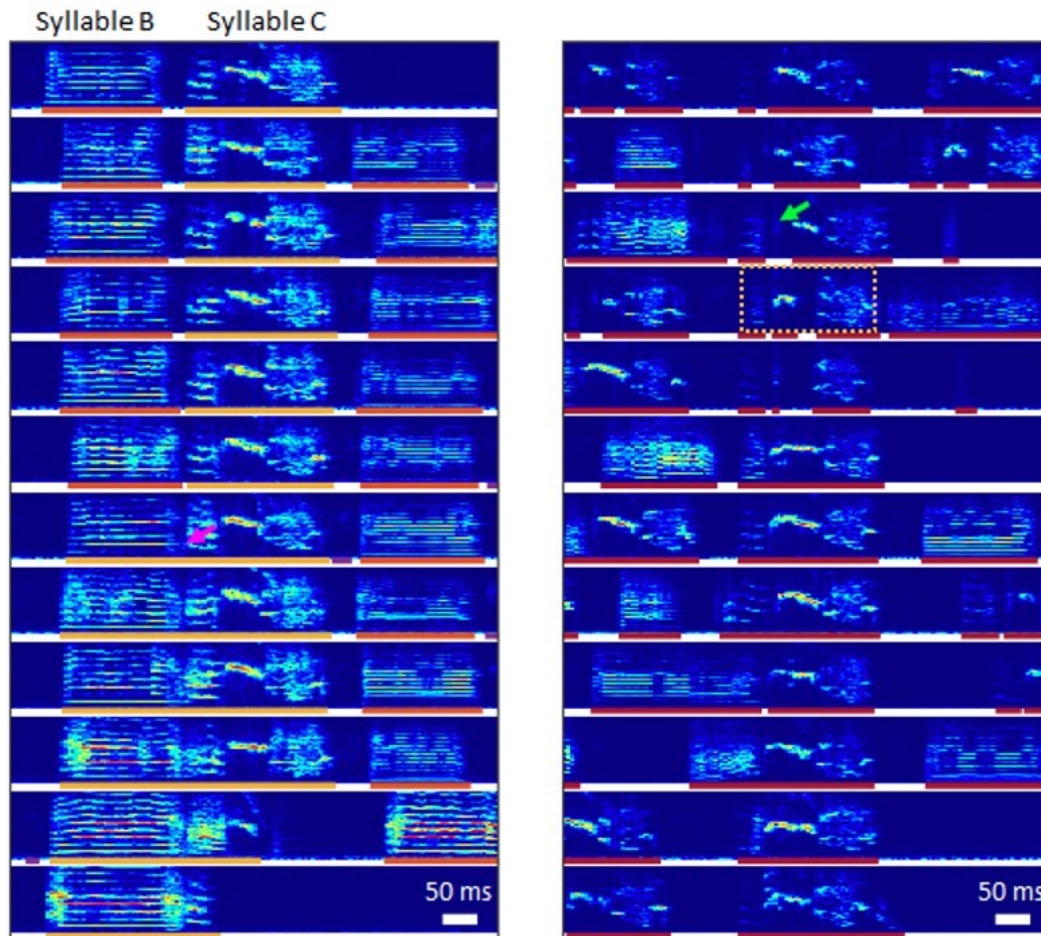
592 guish the times at which an animal vocalizes from the times at which it does not. One of
593 the simplest methods of segmenting vocalizations from continuous recordings is to consider
594 sound amplitude and to define as vocalizations all sounds that are above a given threshold.
595 However, this procedure will misclassify certain noises as vocalizations, which is why more
596 refined approaches are needed that potentially make use of the statistics of the individual³³.
597 In the extreme case, we need to inspect every single potential vocalization and decide based
598 on expert knowledge where to cut the dividing line between vocalization and noise.

599 To standardize the segmentation task, we have created this set of guidelines based on two
600 decisions boundaries for a vocalization:

- 601 • The decision whether there is a silent period between two sounds, which we take by
602 inspecting spectrograms (Fig. 5, left).
- 603 • The decision whether a sound is vocal or non-vocal (Fig. 5, right; Fig. 6-7).

604 Birds, especially when young, tend to vary the gaps between vocalizations. An example is
605 shown in Fig. 5 (yellow dotted box): This sequence of three vocal elements looks like a pre-
606 cursor of syllable C that the juvenile tries to imitate, but they appear with sufficiently large
607 gaps, which is why we sometimes classify them as 3 distinct syllables. Thus for (a) we infer
608 a gap where we can visually detect one, irrespective of other singing attempts in the animal.
609 The second decision boundary (b) is harder to define universally from single-microphone
610 recordings, ideally we would like to have simultaneous recordings from the trachea to mea-
611 sure sounds and air flow there. In practice, it is a human expert, who judges whether
612 a sound is vocal or non-vocal by listening to examples and inspecting the corresponding
613 spectrograms. Again, this task is relatively simple for highly stereotyped vocalizations, but

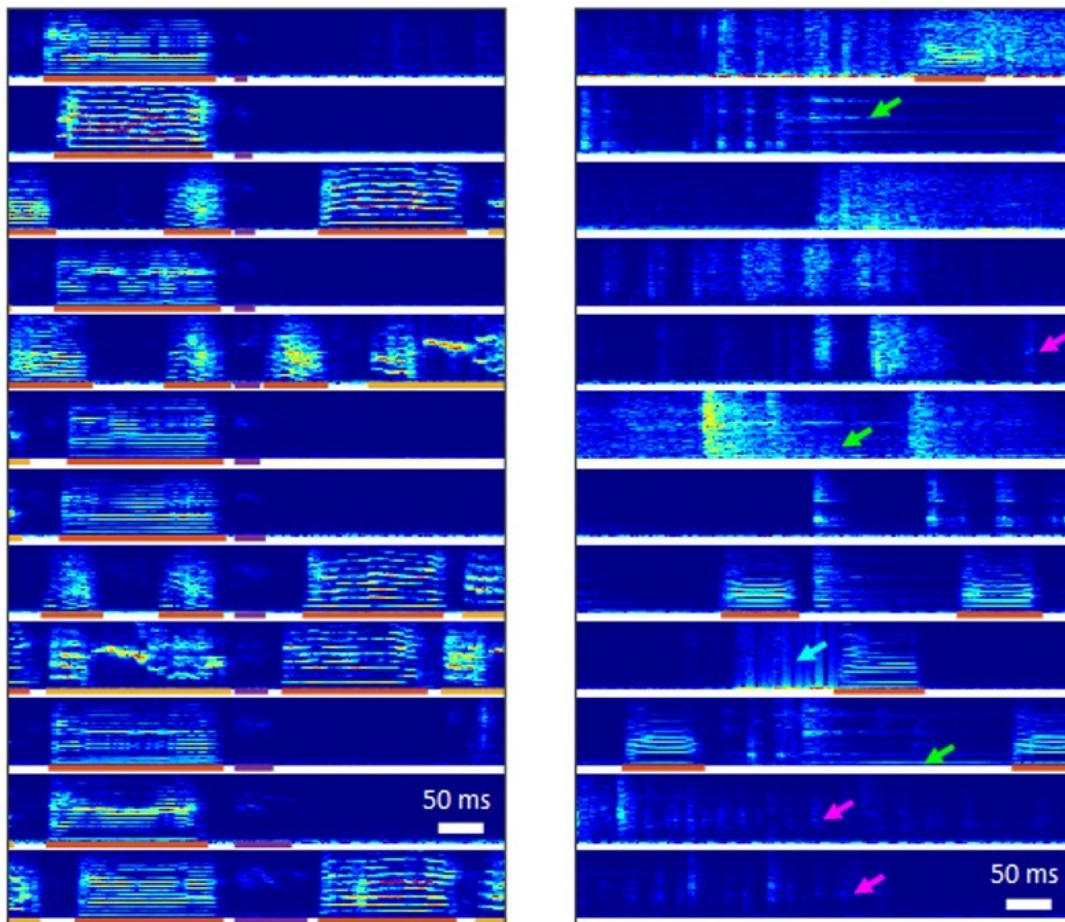
614 more difficult for faint, short and variable vocalizations in juveniles (Fig. 5, right; Fig. 6,
615 left, Fig. 7). A special case consists of faint sounds (usually at around 6kHz) that frequently
616 occur after (or, less frequently, before) vocalizations (Fig. 2, left). We consider them to be
617 inhalation sounds^{33,57} and exclude them from the vocal dataset (default setting).



618

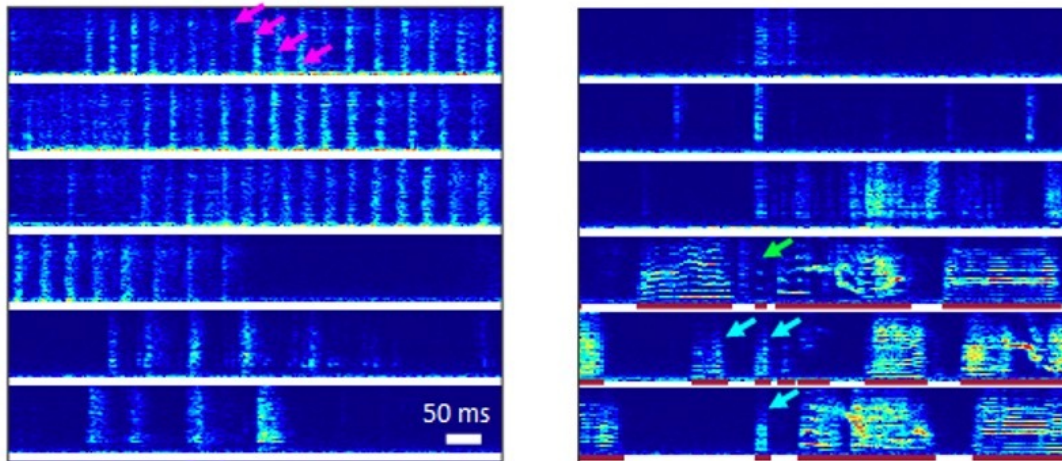
619 **FIG. 5. Definition of vocal segments as continuous intervals of vocal activity.** (left)
620 Zebra finch song examples at 59 day-post-hatch, aligned to notes that resemble the beginning of
621 syllable C. At this stage, syllable C is surrounded by clear gaps most of the time (top 6 examples).
622 However, in a minority of cases, no silent gap is visible between the preceding syllable B and
623 the first note of syllable C (bottom 6 examples, boundary case indicated with magenta arrow).

624 Gold-standard segmentation labels of syllable-C-notes (yellow) and of other vocalizations (orange,
625 purple) are indicated by bars below the spectrograms. (right) Vocalizations recorded at 49 day-
626 post-hatch (red bars), aligned to examples that resemble syllable C. Short noisy sounds within
627 syllable precursors (green arrow) have not been classified as vocal activity based on isolated visual
628 inspection, but likely would be, if the context would be taken into account. The yellow dotted box
629 marks three vocal elements that could potentially be interpreted as a unitary precursor of syllable
630 C, if the developmental endpoint were to be taken into account. Bars as on the left.



631

632 FIG. 6. **Decision-boundary between vocal and non-vocal sounds.** (left) Spectrogram ex-
633 amples of putative inhalation sounds (indicated with purple bars) observed in a zebra finch at 59
634 day-post-hatch (excluded in the gold standard by default). (right) Examples of non-vocal noises
635 which may include prominent tones (green arrows), wide-band noise (blue arrows), or very faint
636 signals (magenta arrows).

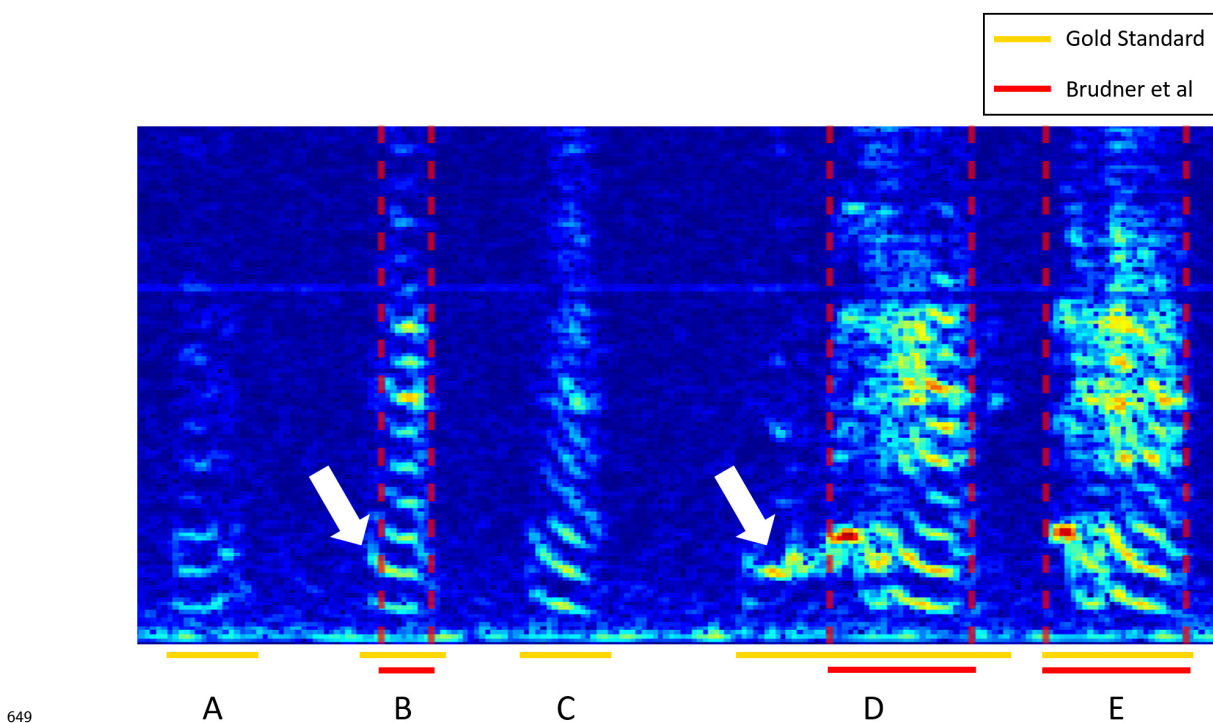


637

638 FIG. 7. **Detailed decision-boundary between vocal sounds and wing flaps.** Spectrogram
639 examples short noises. Wing flaps are easy to detect on spectrograms when occurring in serial
640 repetition (i.e., when the bird is flying; magenta arrows). For short sounds, indicators of vocal
641 activity can be harmonics (green arrow) or a strong skew in the spectral density towards certain
642 frequencies (low frequency sounds indicated with blue arrows).

643 **2. Analysis of an open dataset**

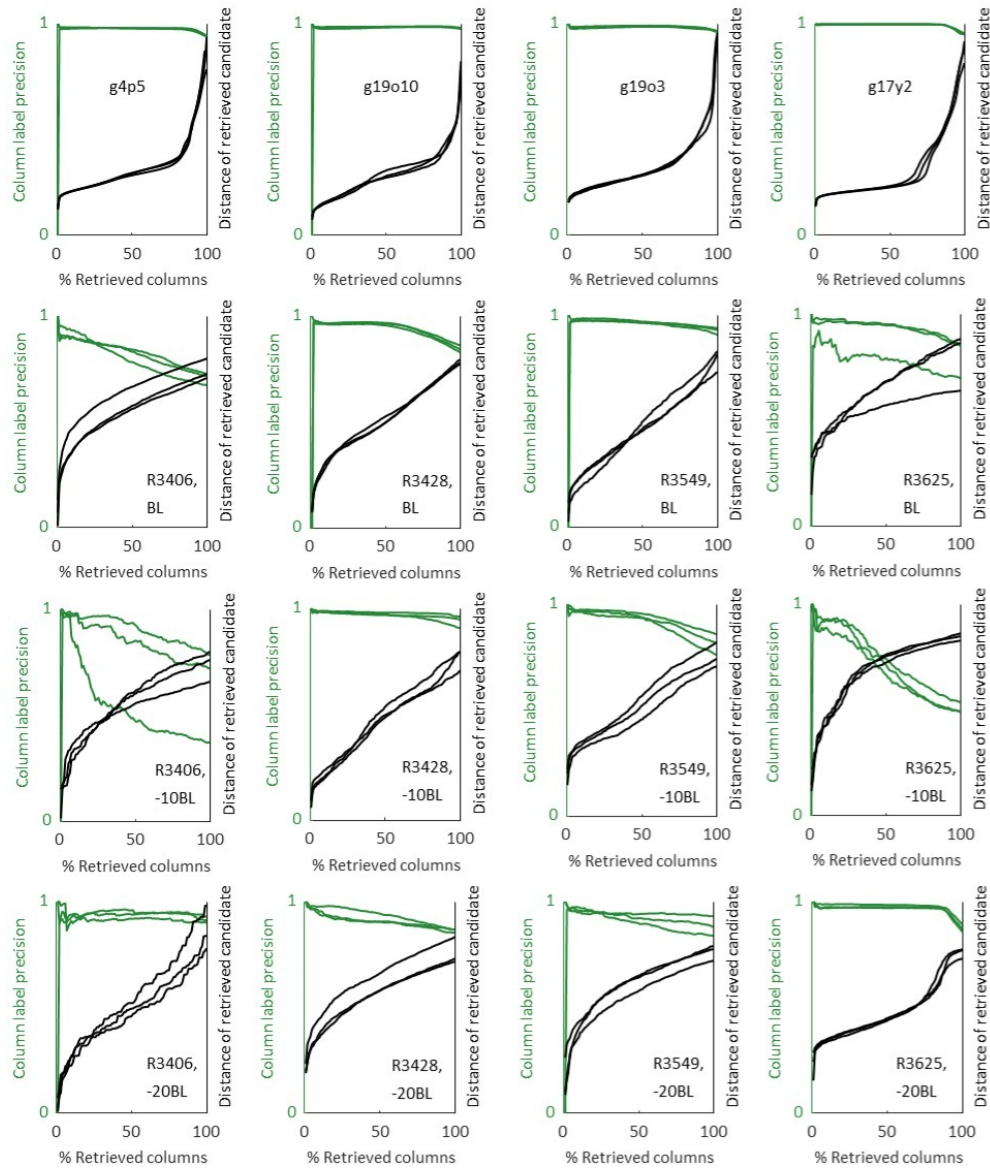
644 A recent publication¹² includes a large dataset of vocal segments from 5 zebra finches.
645 According to the data documentation, the segmentation was performed using a sound-
646 amplitude based method that included some hand tuning. Although we found the published
647 segmentation results to be valuable, they were insufficient to qualify as gold standard, due
648 to the existence of false negatives and inaccurate segment boundaries Fig. 8.



649 **FIG. 8. Example segmentation inaccuracies of the¹² dataset.** The published segments (red
650 horizontal bars) deviate from the (gold-standard) manual annotations (gold horizontal bars) in
651 terms of a false negative sample (Syllables A and C) and in terms of inaccurate segment boundaries
652 (white arrows).
653

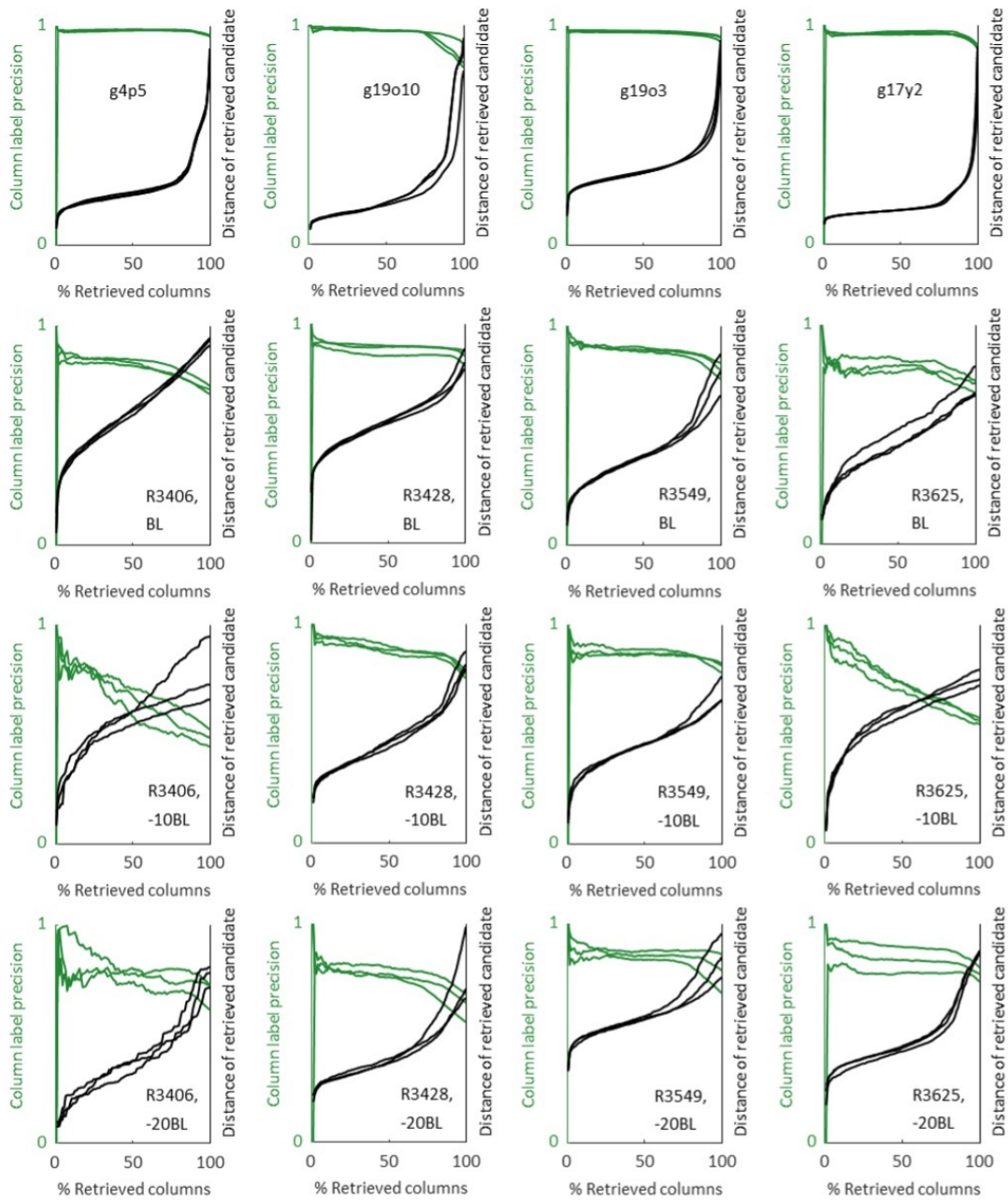
654 **3. Discussion**

655 The examples we provided illustrate our decision boundaries and the difficulties with
656 segmentation approaches. In summary, we advocate the definition of vocal segments as
657 tightly restricted intervals of continuous vocal activity. These segments should be defined
658 independently from functional considerations. How to extract functional units from vocal
659 segments is an open question, the answer may depend on whether the vocal units are assessed
660 in the domain of perception (receiver) or production (sender). Still, it is regarded as ideal
661 to validate chosen segmentations based on the functional roles of the vocal signals^{44,56,58}.
662 However, recent work in songbirds suggests that “syllables may not be perceptual units for
663 songbirds as opposed to common assumption”⁵⁹.



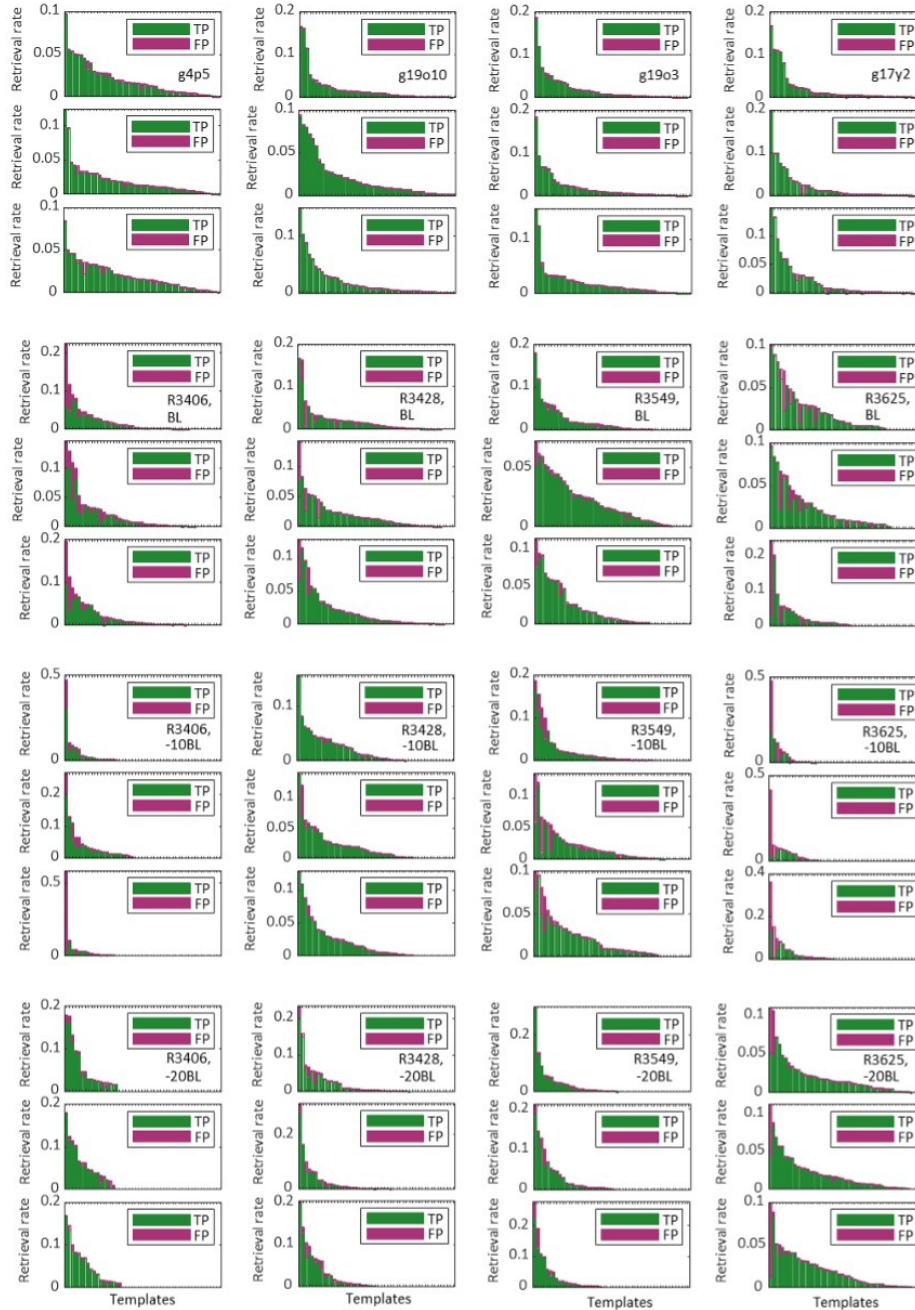
664

665 FIG. 9. Extended set of precision and distance curves as a function of retrieval pro-
666 gression, using the WHOLE approach (replicated for all birds). The top row shows adult
667 birds, while the subsequent rows show juveniles at different ages relative to baseline. See Figure
668 3a for a detailed description.



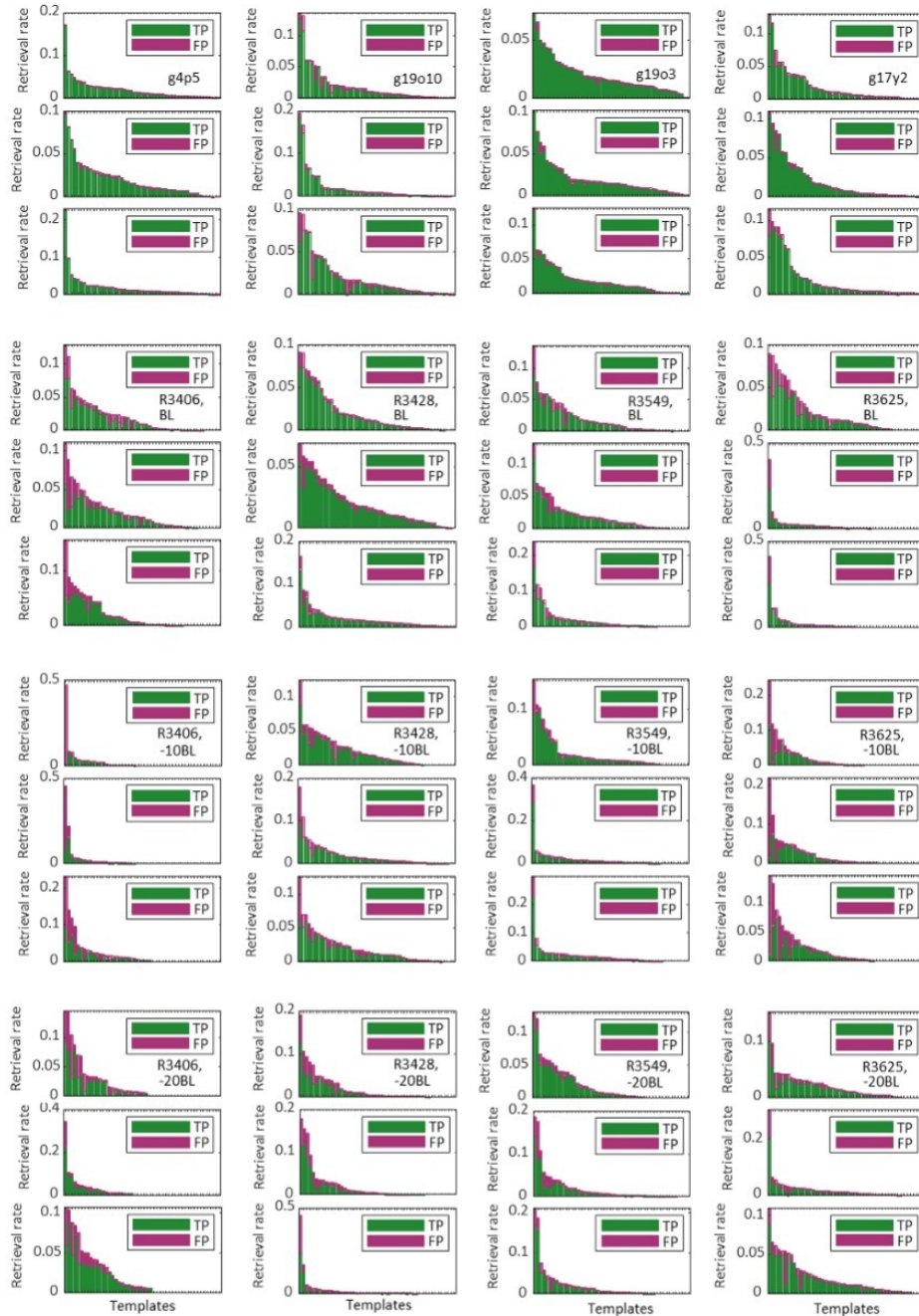
669

670 FIG. 10. Extended set of precision and distance curves as a function of retrieval pro-
671 gression, using the PART approach (replicated for all birds). The top row shows adult
672 birds, while the subsequent rows show juveniles at different ages relative to baseline. See Figure
673 3a for a detailed description.



674

675 FIG. 11. Extended set of histograms of retrieval rates across templates, using the
676 **WHOLE** approach (3 retrieval replicates for each bird). The top row (consisting of 3
677 panels for each retrieval replicate) shows adult birds, while the subsequent rows show juveniles at
678 different ages relative to baseline. See Fig. 4a-c for a detailed description.



679

680 FIG. 12. Extended set of histograms of retrieval rates across templates, using the PART
681 approach (3 retrieval replicates for each bird). The top row (consisting of 3 panels for each
682 retrieval replicate) shows adult birds, while the subsequent rows show juveniles at different ages
683 relative to baseline. See Fig. 4a-c for a detailed description