Analysing vocal complexity in relation to sociality in orcas of British Columbia: An application of long-term computational passive acoustics

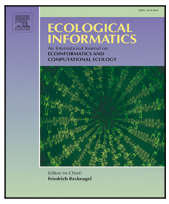
Paul Best, Marion Poupard, Ricard Marxer, Paul Spong, Helena Symonds, Hervé Glotin

Ecological Informatics 90 (2025) 103211

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Analysing vocal complexity in relation to sociality in orcas of British Columbia: An application of long-term computational passive acoustics

Paul Best [a,b,*], Marion Poupard [a,b], Ricard Marxer [a,b], Paul Spong [c], Helena Symonds [c], Hervé Glotin [a,b,*]

[a] Université de Toulon, Aix Marseille Univ, CNRS, DYNI, LIS, Toulon, France
[b] Centre international d'Intelligence Artificielle en Acoustique Naturelle (CIAN), France [1]
[c] OrcaLab, Hanson Island, P.O. Box 510, Alert Bay, BC, V0N 1A0, Canada [2]

ARTICLE INFO

Dataset link: https://zenodo.org/records/15462307

Keywords:
Vocal complexity
Orcinus orca
Passive acoustic monitoring
Machine learning

ABSTRACT

Orcas are both highly social and highly vocal animals. In coastal waters of the North-Eastern Pacific Ocean, the Northern Resident orca population is well monitored, providing a great opportunity to learn about their social and communicative behaviour. Here, we report a series of acoustic analyses that lead to the empirical assessment of factors that might impact vocal complexity.

Automatically processing long-term passive acoustic data, we detected and classified calls to transcribe vocal activity. Detailed post-hoc analyses show that the detection model is imperfect, especially in detecting calls of low energy. Also, diarisation is not possible with this data and transcriptions might gather a mixture of several emitters. Taking these limitations into account, we measured communicative complexity considering the groups' vocal production as a whole. Acoustic and visual cues also enabled the identification of specific groups with estimated numbers of individuals.

Results highlight a positive correlation between vocal and social complexity, which could be due to the mere effect of having more potential emitters. Nonetheless, this brings a first demonstration of the non-trivial link between the number of emitters and complexity in the composition of sequences. We also demonstrate significant impacts of other proximate factors such as behaviour on vocal complexity measurements, and advocate for multi-factor considerations when evaluating communicative complexity.

This work demonstrates the pertinence of joint efforts between passive acoustics, visual observations and machine learning to enhance the scale of behavioural studies and assess the validity of evolutionary hypotheses of communication systems.

1. Introduction

Toothed whales have both complex vocal behaviour and social organisation. This makes them an interesting case of communication in species distantly related to humans. Their communication systems are studied for their acoustic structure, and how this may be linked to their behaviour or sociality. Their vocalisations are learned, and presumably serve for social signalling and social bonding (Janik, 2014). However, the usage varies depending on the species. For instance, bottlenose dolphins (*Tursiops truncatus*) use individual-specific signature whistles (Tyack, 2012), whereas orcas (*Orcinus orca*) use community-specific pulsed-calls (Ford, 1987).

Orca calls, for instance, form a repertoire of discrete categories called call types (acoustic similarity is based on characteristics such as frequency contours). For several populations, these repertoires were shown to correlate with social structures (Yurk et al., 2002; Ford, 1987; Filatova et al., 2007). In a given orca population, social structure is commonly characterised by three hierarchical levels: the matriline (group of individuals that share a living female ancestor), the pod (several matrilines related by a common deceased female ancestor), and the clan (set of pods that share elements of their acoustic repertoires) (Towers, 2020). The pod level is the most stable social level, where the members spend the majority of their time together. This is the level where dialects are observed (Bigg et al., 1990; Ford, 1987).

In terms of function, orca calls have been suggested to facilitate social cohesion, and no straightforward relationship between call types and behavioural states were found (Ford, 1989; Filatova et al., 2013).

* Corresponding authors at: Université de Toulon, Aix Marseille Univ, CNRS, DYNI, LIS, Toulon, France.
E-mail addresses: paul.best@univ-amu.fr (P. Best), glotin@univ-tln.fr (H. Glotin).
[1] https://cian.lis-lab.fr
[2] https://orcalab.org/

https://doi.org/10.1016/j.ecoinf.2025.103211
Received 11 September 2024; Received in revised form 19 May 2025; Accepted 19 May 2025
Available online 6 June 2025

Ford (1989) demonstrated that the distribution of call types varies with activity, though no particular call usage was linked to a specific behaviour. Another study has shown that orcas have a context dependent modulation of call amplitude (Saulitis et al., 2005) (between lone individuals and hunting ones). Alternatively, it has also been reported that activity does not affect the proportions of call categories but multi-pod interactions does (Filatova et al., 2013). Stereotyped calls thus appear to have a role akin to contact calls observed in some terrestrial species, but the size of the repertoire cannot be explained solely by this function (Ford, 1989). In addition to keeping contact with relatives in an environment where visibility is limited, the repertoire might serve to maintain group integrity (Filatova et al., 2015), particularly since pods often mix together.

Besides their functions, structure of orca vocalisation sequences have also been studied. Whether of pulsed calls or of whistles, sequences were found to be non-random — call type probability distribution alone cannot explain call transition frequencies (Ford, 1989; Riesch et al., 2008), but no strict call combination pattern emerges (Selbmann et al., 2023). As an alternative to look at sequential structures or call type distributions in relation to behaviour, measuring the complexity of vocal sequences could also shed light on their usage and functions. Complexity in orca communication has previously been reported in comparative studies (Kershenbaum et al., 2021, 2014), however it has not yet been subject to extensive analyses in relation to socio-behavioural factors.

In non-human animal communication research, a large number of studies focus on measuring the complexity of communication systems (McCowan et al., 2002; Kershenbaum, 2014; Suzuki et al., 2006). It is indeed appealing to use complexity metrics to compare vocal behaviour across taxa. This, in turn, potentially helps in phylogenetic studies of language. Commonly used complexity metrics are the number of call types in a repertoire (indicating the potential diversity of the communication system), and the Shannon entropy of the call occurrence distribution (which reflects the potential number of bits carried by call sequences or the level of uncertainty) (Peckre et al., 2019).

Complexity of communication systems, according to the social complexity hypothesis, has been stimulated by social complexity. Therefore, the two have co-evolved across taxa (Freeberg et al., 2012). Several empirical findings have supported this hypothesis by comparing communication systems across closely related species, whether in birds such as in the wren family (*Troglodytidae*) (Kroodsma, 1977), or in mammals such as in ground-dwelling sciurids (Blumstein and Armitage, 1997) and primates (Bouchet et al., 2013) (see Freeberg et al., 2012 for a review). A phylogenetic analysis of toothed cetaceans has already shown a correlation between vocal complexity and social complexity across species (May-Collado et al., 2007). However, in this study, vocal complexity measurements were based on the number of inflection points in tonal sounds, but did not look at their arrangement in sequences nor at repertoire sizes.

In our study, we focus on the Northern Resident Killer Whale (NRKW) population, recurrently observed in the Johnstone Strait/Blackfish Sound area of Northern Vancouver Island (Western Canada). Thanks to the significant efforts of the OrcaLab land-based laboratory and boat-based surveys, the social structure of the NRKWs is well known. Combining photo-identification (Towers, 2020) with acoustic monitoring also allowed to define the acoustic repertoires of each pod (Ford, 1987).

The extensive knowledge we have on this population provides a good opportunity to better understand the complexity of their communication system. Similarly to repertoire composition, we can test how complexity correlates with behaviour, or social structures. Thus, given long-term passive acoustic recordings of NRKWs, we use machine learning methods to examine their vocal behaviour. We achieved this by transcribing acoustic signals into sequences of units, following previously defined call types (Ford, 1987). Focusing on one clan of this population, the A clan (pods A1, A4 and A5) (Towers, 2020), we report here various complexity metrics and how factors such as approximate group size or behaviour might correlate with them.

This brings a novel perspective on the correlation of social factors with communicative complexity. Firstly, it is yet impossible to identify individual emitters from this type of passive acoustic recordings. Hence, instead of measuring the complexity of vocal sequences produced by a single individual, we measured complexity from 'mixture' sequences produced by whole groups. This is different to the traditional way of measuring complexity in animal communication research, but is nonetheless of interest to learn how communication systems are collectively used. Secondly, in contrast to phylogenetic analyses, measuring how complexity varies with proximate factors within a single population does not directly inform us on evolutionary drivers. However, analysing proximate variations in the usage of communication systems can provide insights on ultimate influences that drive the evolution of communication systems across species. Finally, we conducted these measurements with automated systems applied to long-term data (over 5 years), collected without interference with the animals.

2. Material and methods

2.1. Data acquisition

For 20 years, the OrcaLab NGO[3] developed and maintained a unique multi-hydrophone acoustic monitoring station around Hanson Island (Northern Vancouver Island, Canada) to study orcas. It is now composed of seven hydrophones extending over 50 km^2 of open water (Fig. 1). In 2015, we set up a continuous recording from six of these hydrophones (Fig. 1). The aim was to allow the observation and modelling of bioacoustic activities for various species, including details of their ecoacoustic niche and under various geophysical and anthropophonic conditions (Poupard et al., 2019b). Besides, as in this present study, it helps build new knowledge about orca vocal behaviour (Poupard et al., 2019a). The material used here gathers full time recordings from summers 2015 to 2020 (July 1st to September 30th), when NRKW presence is most notable. Accounting for occasional system malfunctions, this database is equivalent to 403 days of recordings over 6 channels.

2.2. CNN based detection and classification

To analyse this large quantity of data, a convolutional neural network (CNN) was trained to detect NRKW pulsed calls on spectrogram representations. The deep learning approach has shown promising developments for bioacoustic tasks in recent years (Xie et al., 2023), most often for binary detection tasks (Bergler et al., 2022; Zhong et al., 2020; Best et al., 2020). Here we chose the object detection paradigm, inspired by computer vision, in which the model provides time–frequency bounding-boxes for vocalisations in spectrograms (Stowell, 2022).

Mel-spectrograms were computed on acoustic signals sampled at 22,050 Hz. The Short-Term-Fourier-Transform applied Hann windows of 0.05 s with a 87.5% overlap, and was followed by a Mel filter-bank of 128 bins spread logarithmically from 0 to 11,025 Hz. The model architecture and training procedure was borrowed from YOLO (Redmon et al., 2016), a framework widely used for object detection in images, but also in bioacoustic applications (Coffey et al., 2019; Parcerisas et al., 2024; Escobar-Amado et al., 2023). Hyper-parameters were left to their default values, and the model was trained for 500 epochs.

As training data, we used a set of 1750 calls with manually annotated call types and time–frequency boundaries. The call types were among 21 categories following the catalogue described by Ford (1987) (the distribution of labels is shown in Fig. 2). A random sample of 100 calls was selected from this set to form a validation set which was used to compute the confusion matrix presented in Fig. 2, and the F-score of

[3] https://orcalab.org

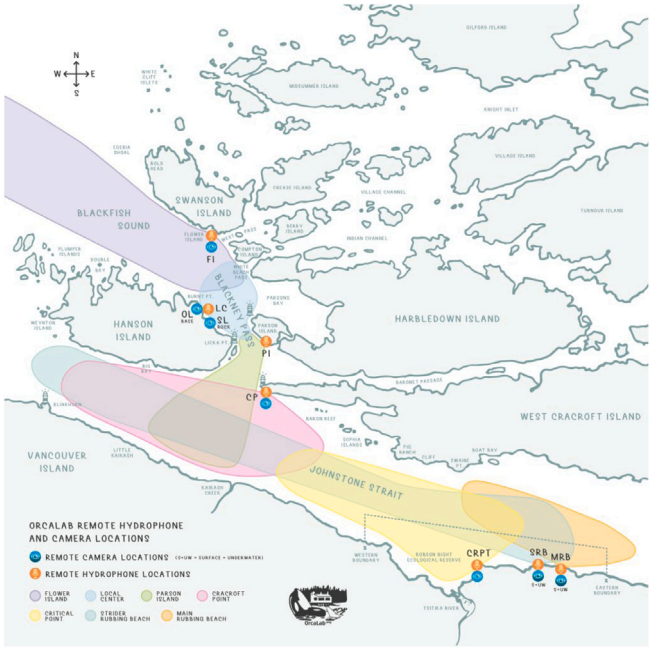


Fig. 1. Map of the area and listening range for the 7 hydrophones of the OrcaLab station (taken from visited in Sept. 2023). Hydrophones used in this study are FI, LC (named LL in this paper), PI, CP, CRPT and MRB (named RB in this paper).

0.76 (at a confidence of 0.5). We would like to emphasise that the YOLO system used here conducts class specific detection rather than strict sense classification, which makes the commonly reported accuracy metric impossible to measure. Nonetheless, the confusion matrix shows that there were no confusions between call types, except sporadic ones with the 'background' class, denoting false alarms or missed calls.

2.3. Extraction of sequences

For the following analysis, sequences of calls were extracted from CNN predictions, considering that vocalisations belong to the same sequence if they are separated by less than 10 s. This threshold was chosen on visual inspection of the inter-vocalisation interval distribution which showed a dip around this value (Sup. Fig. 3), and is comparable to the threshold used in a previous vocal sequence study on orca whistle sequences (Riesch et al., 2008). It is important to note that unlike many vocal sequence analysis, the recordings used in this study gather the emissions of potentially multiple individuals. For instance, the calls extracted in the exemplary sequence shown in Fig. 3 have varying energy levels, suggesting a combination of several emitters, with the ones further away being missed by the detection system. Nonetheless, analysing multi-emitter vocal sequences is relatively common in cetacean communication studies (Ford, 1989; Riesch et al., 2008), especially in passive acoustic monitoring setups.

Only sequences with at least 3 calls, and with no call categorised as 'unknown', 'bark', 'squawk', or 'whistle' were kept. These vocalisations, which are not in the pulsed-call catalogue used here (Ford, 1987), were discarded in our analysis because we did not annotate them by type as we did for pulsed calls. Thus similarly to other studies (Ford, 1989; Riesch et al., 2008), we focus our analysis on transcriptions of a single vocalisation category. This yielded more than 20,000 sequences with a total of 137,760 calls. The distribution of sequence sizes is shown in Fig. 4 in a log-survivor plot. For instance, there are 1000 sequences of at least 20 calls, and this amount exponentially grows as we decrease in minimal size. When apart by less than 5 min, sequences were again grouped by passage (or bout). Passages describe the whole duration in which a group of orcas evolves within the hydrophones' detection range, allowing the assumption that the acoustic events within this temporal window are emitted by the same group of individuals.

The count of call types detected in sequences already enables a preliminary analysis of the communication system by testing if it is Zipfian and measuring its Power Law Coefficient (PLC). In several studies, Zipf's Law (Zipf, 1949) has been used to quantitatively evaluate animal communication system repertoires (for humans Yu et al., 2018 and non-humans McCowan et al., 2005; Kershenbaum et al., 2021). This analysis relies on the estimation of the PLC which reflects the relationship between the rank of a repertoire's token r (for the most frequent token $r = 1$) and its frequency of occurrence f, following Eq. (1) (α is set at highest token frequency).

$$f = \alpha \times r^{\mathrm{PLC}}. \tag{1}$$

The PLC is a comparative measure for repertoire complexity (McCowan et al., 2005), PLC = 0 indicating a uniform distribution, and PLC $\ll -1$ implying a highly skewed one. Zipf (1949) states that for a system that follows constraints of efficiency ('least effort'), the PLC would converge to −1. This is supported by the fact that most human languages have a PLC close to −1 (Yu et al., 2018). A PLC close to −1 would thus be a necessary condition for a communication system to be 'language-like' (Kershenbaum et al., 2021; McCowan et al., 2005).

2.4. Extraction of pod information

We used two different approaches to attribute the detected sequences to a specific pod: manual and automated. The first was to use data from the land based manual audio–visual monitoring conducted at OrcaLab in the summer of 2016. This consists in 98 passages of pod A1, 90 of pod A4, and 27 of pod A5 (see Table 2). These data were limited in temporal coverage and in volume, which motivated a parallel automated approach.

To automatically infer the pod or group of pods responsible for detected sequences, we used the N09 call, which has 3 variations that each can be attributed to a single pod of the A clan (N09i, N09ii, and N09iii to A1, A4 and A5 respectively). Passages with one instance of these sub-types detected, and none of the other two, were attributed to the corresponding pod, passages with two different N09 sub-types detected were attributed to an aggregation of the two corresponding pods and so on.

The number of passages and sequences inferred from the automated approach are summarised in Table 1. It is worth noting that no detection of a pod (for instance pod A4 in 2016) does not mean that they were not present on site, but rather that either they did not emit their own N09 call, that it was not detected by the CNN, or that it co-occurred with N09 sub-types of another pod of the A clan.

2.5. Complexity metrics

2.5.1. Repertoire size

A first approach to measure the complexity of a system is to count its number of possible outcomes. For a communication system, this relies on the discretisation of acoustic signals into units and their following categorisation. As such, the number of categories, commonly referred to as repertoire size, gives a first indication on potential communicative complexity. It is formalised by N in Eq. (2), with C being the set of possible call types.

$$\mathrm{N} = \#C. \tag{2}$$

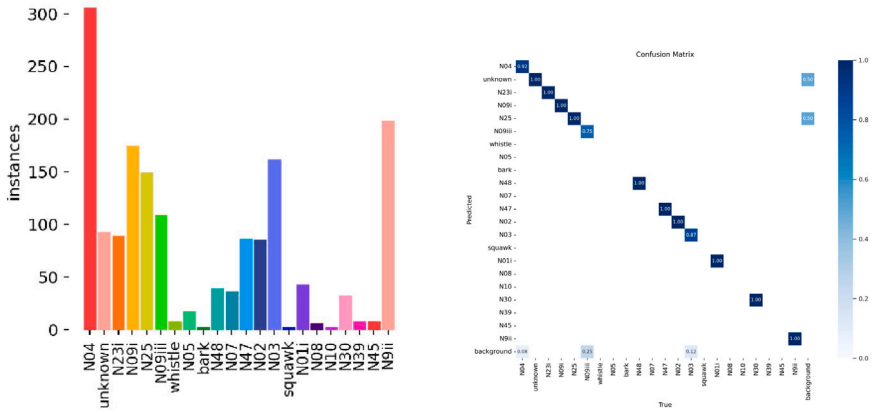


Fig. 2. (left) Distribution of annotations by call type in the training set. The various colours used do not carry any information. (right) Confusion matrix of the model on the validation set.

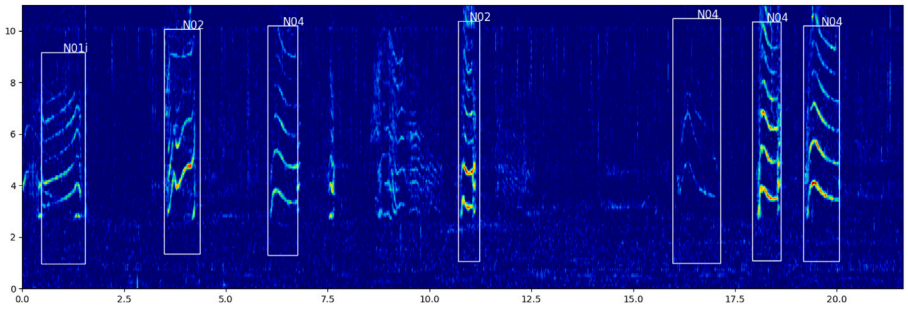


Fig. 3. Example spectrogram of an orca call sequence with YOLO detections. The time axis is given in seconds and frequencies in kHz.

Table 1
Data volumes for each year and pod or pod aggregation. Here, the attribution of passages was inferred through the automatic classification of the N09 call.

	Sequences							Passages						
	A1	A4	A5	A1–A4	A1–A5	A4–A5	A1–A4–A5	A1	A4	A5	A1–A4	A1–A5	A4–A5	A1–A4–A5
2015	540	1	7	34	319	2	189	138	1	4	4	36	1	8
2016	946	0	67	34	541	7	86	253	0	22	5	71	1	6
2017	2,416	4	87	74	1,138	0	152	641	1	28	11	153	0	9
2018	1,502	171	116	86	803	5	533	22	59	12	167	2	16	
2019	1,590	0	94	13	738	0	31	426	0	53	3	132	0	4
2020	615	37	54	20	522	8	86	209	4	25	4	104	3	12
Total	7,609	213	425	261	4,061	22	655	2,198	28	191	39	663	7	55

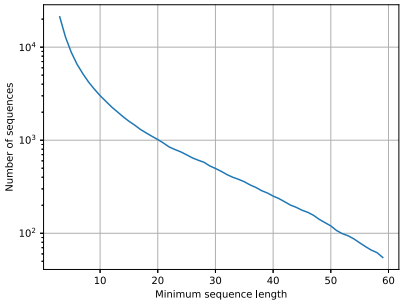


Fig. 4. Log-survivor plot (Fagen and Young, 1978) **of the extracted sequences' lengths.** This reads as "there are 1000 sequences of at least 20 calls" for instance.

Table 2
Data volumes for acoustic detections that were attributed a pod from the visual observations of 2016.

	Sequences	Passages
A1	320	75
A4	15	4
A5	98	23
A5-A4	16	4

2.5.2. Shannon entropy index

Repertoire size gives an indicator of potential emissions of a signaller, but does not make sense of the actual usage of units. For instance, we might consider as less complex a system that uses a single unit most of the time, and others only on rare occasions. The Shannon entropy index captures this nuance by quantifying the uncertainty of a system as given by the probability distribution of units. It is formulated in Eq. (3), using the probability of a call x to belong to a call type c, and normalised by the maximum entropy given by $\log_2(N)$. This normalisation is especially relevant in the case of varying repertoire sizes (Rebout et al., 2021).

$$\mathrm{H} = \frac{-1}{\log_2(\mathrm{N})} \sum_{c \in C} P(x = c) \log_2(P(x = c)). \tag{3}$$

2.5.3. Entropy rate

The Shannon entropy index quantifies uncertainty of call type occurrence but ignores potential syntactic structures. For instance, in NRKW call sequences, the same type is often repeated several

times (Ford, 1989). In a way, this reduces the uncertainty of the communication system, but only if we take into account the preceding call to refine our guess of the next. The Entropy Rate (ER), proposed by Kershenbaum (Kershenbaum, 2014) to measure animal vocal complexity, solves this issue by computing the Shannon entropy of call transitions, as given by Eq. (4), with x_1 and x_2 being two subsequent vocalisations.

$$\mathrm{ER} = \frac{-1}{\log_2(\mathrm{N})} \sum_{c_1 \in C} P(x = c_1) \sum_{c_2 \in C} P(x_1 = c_1, x_2 = c_2) \log_2(P(x_1 = c_1, x_2 = c_2)). \tag{4}$$

2.5.4. Entropy of association rate

In passive acoustic datasets that record groups of individuals simultaneously with no information on specific call emitters, transition probabilities might be misleading as they will include intertwined call sequences of multiple individuals (this is also known as the cocktail party problem). However, the occurrence of a call type might still suggest the close appearance of another call and reduce the system's uncertainty, say if a given behaviour changes the call type probability distribution for instance. We thus propose a slight modification of the ER formula, by using association rates (AR) instead of transition probabilities. Here, by association rate, we refer to the probability that given a call type c_1, another call type c_2 is found in the same sequence (independently of either of their positions). We formulate this in Eq. (5), with S being the set of all recorded sequences. This association rate can be seen as analogous to the association index widely used in animal behaviour studies to measure the frequencies of social interactions between individuals (Cairns and Schwager, 1987).

$$\mathrm{AR}(c_1, c_2) = \frac{\sum_{S_i \in S} \mathbb{I}(c_1, c_2 \in S_i)}{\sum_{S_i \in S} \mathbb{I}(c_1 \in S_i)}. \tag{5}$$

Given this AR value for each call type pair, and following the incentive to measure combinatorial complexity using the Shannon entropy (Kershenbaum, 2014; Rebout et al., 2021), we can compute the entropy of association rates (EAR) in a similar fashion as the ER (Eq. (6)).

$$\mathrm{EAR} = \frac{-1}{\log_2(\mathrm{N})} \sum_{c_1 \in C} P(x = c_1) \sum_{c_2 \in C} \mathrm{AR}(c_1, c_2) \log_2(\mathrm{AR}(c_1, c_2)). \tag{6}$$

3. Results

3.1. Classifier performance assessment

Besides the performance measurements from the validation set, to verify the reliability of the neural network for the analysis of call sequences, we manually inspected its output in two manners. The first was to count the number of missed calls (false negative) or false alarm (false positive) for 100 randomly sampled sequences. Out of them, 16 sequences consisted exclusively of false alarms due to sound card malfunctions, and the distribution of evaluations for the remaining is shown in Fig. 5. Noticing that often, missed calls were of significantly lower energy than others in the sequence, we ran the test twice with different selection criteria. First, we considered all perceivable calls (for the sequence in Fig. 3 there would be 3 missed calls), and second we considered only foreground calls with an energy level similar to the rest of the sequence (for the sequence in Fig. 3 there would be only 1 missed calls).

The other approach was to measure the confusion of call types. To achieve this, for each type, we randomly sampled 50 calls and manually assessed the level of accuracy for them. The resulting confusion matrix is given in Fig. 6.

As for detection performance, out of the 16 sequences containing only noise, 13 were recorded at the Rubbing Beach (RB) hydrophone (81%). Also, it is worth noting that 12 out these 13 sequences contain

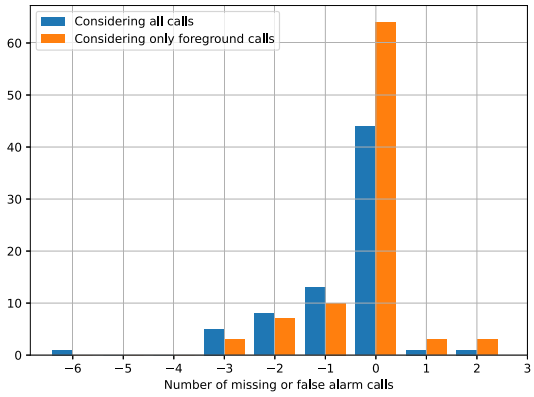


Fig. 5. Post-hoc analysis of detection performance. Bar heights denote the number of sequences found with varying numbers of missed calls (negative numbers) or false alarms (positive numbers) with 0 being a perfect detection. We ran this test twice with the same data, first considering that all perceivable calls should be detected, and second considering that only foreground calls should be detected.

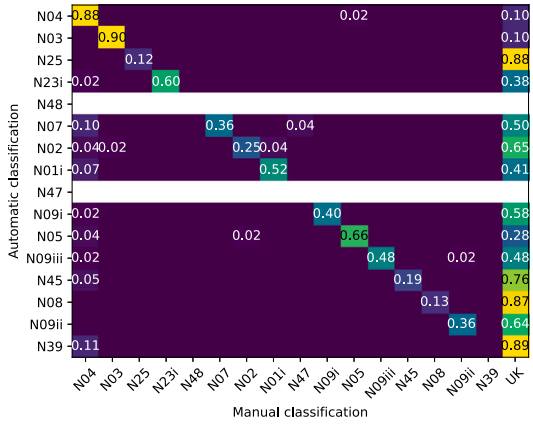


Fig. 6. Post-hoc confusion matrix for 50 randomly sampled predictions (uniformly over years and hydrophones) of each type. Counts were normalised by line, with bright colours denoting higher frequencies, and the unknown column (UK) referring to calls for which the annotator was unsure of the type due to a too low signal-to-noise ratio.

only one call type repeated several times. This has strong implications for the complexity measurements conducted in the following sections. For the remaining sequences, the majority had no false nor missed detections, but a significant number had between 1 and 3 missed detections. The fact that the model has a stronger tendency to miss calls than to detect some that were not present implies that complexity metrics might be underestimated rather than overestimated.

As for type classification, in addition to the commonly used confusion matrix on the test set (Fig. 2), we propose to report the proportion of confusions for new model predictions (Fig. 6). An experienced annotator was shown calls without being informed of the model predictions and asked to attribute types to them. The calls tested here consisted of 50 per type used in the subsequent analysis (having included the 'unknown' and non-pulsed call classes), sampled randomly from the model predictions.

For a significant number of instances, the annotator was unable to unambiguously attribute a type due to a too low signal-to-noise ratio (SNR). This highlights the common gap between annotated data used in training (often from high SNR conditions) and real-world scenarios. These low SNR conditions were especially present for the less frequent calls such as N39, N45 or N08. Besides these ambiguous cases, for each category, the majority of samples had their types correctly identified by the model, the most common confusion being the N04

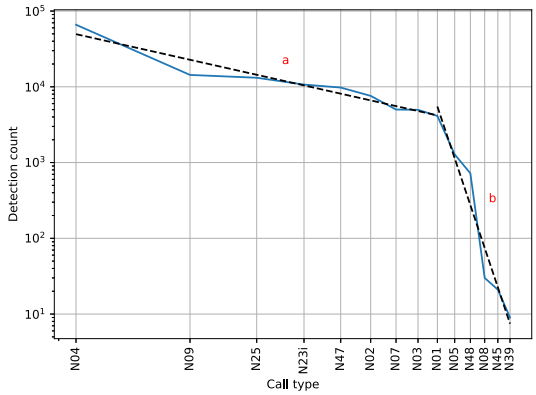


Fig. 7. Distribution of call types detected in sequences (solid line). We display this distribution as commonly done in Zipfian analyses, showing the log-count as a function of the log-rank of the tokens, descending from the most frequent to the least frequent. The dashed lines denote fitted Zipfian slopes (Kershenbaum et al., 2021) (for part a: PLC = −1.12 and $R^2 = 0.93$; for part b: PLC = −14.9 and $R^2 = 0.94$).

call (between 2% and 11% depending on the type), possibly because of its over-representation in the training data (Sup. Fig. 2).

Overall, this post-hoc model validation, instead of being computed on isolated calls of a test set, is directly sampled from the data used in our sequence analysis. As such, it informs us more precisely on the expected level of error and its potential impact on subsequent results.

3.2. Zipf analysis of detected type distribution

Throughout the summers of 2015 to 2020, close to 140,000 calls were detected in sequences of at least three vocalisations. The distribution of detected call types is shown in Fig. 7, and appears divided in two parts (noted 'a' and 'b'). The need to fit separate curves to estimate the PLC of a frequency distribution has previously been proposed and is explained by the broken power law effect (Kershenbaum et al., 2021) (element generation operates differently at extreme high and low frequencies). Linear regression analysis on log ranks of call types versus log detection counts gave PLCs of −1.1 for part 'a', and −14.9 for part 'b'.

3.3. Non-randomness of call sequences

From automatic detections and the extraction of call sequences, we were able to compute the transition probability matrix between call types. This approach can reveal certain patterns in a communication system, such as the high association rate between two units. Moreover, transition probabilities are used to compute the ER complexity metric, which measures how predictable vocal units are, given their precedent one. The transition matrix is displayed in Fig. 8, along with that of a previous study conducted on the same population 30 years before (Ford, 1989) (the difference between the two matrices is shown in Sup. Fig. 1).

This visualisation confirms the previous finding of non-randomness of call usages (Ford, 1989), in other words, that call type probability distributions with an a priori call (each line in Fig. 8) are different from the global probability distribution (column total in Fig. 8). Especially, as in Ford (1989), there is a tendency for repetition (strong diagonal of the matrix), and a propensity for N04 calls — the tendency for repetition has also been demonstrated for another orca group (Kershenbaum et al., 2014). This concordance shows the relative stability of the orca's vocal behaviour through time, and validates the potential use of automated methods to study them.

With the large scale of this study, we were able to extend this matrix to the probability of call emission depending on the two preceding calls.

Table 3
Pair-wise comparison (Kruskal–Wallis H-test) of complexity across annotated pods. Compared distributions are shown in Fig. 9. Symbols indicate the direction of the distribution difference when the *P* value<0.001 ('+' being a congruent increase of complexity with approximate pod size).

		A1
A5	N	+
	H	+
	ER	+
	EAR	ns

This is shown in Sup. Fig. 2, and suggests that the tendency of calls to be repeated extends even with another call in-between (the probability of N01 is higher when the second to last call was of the same type).

3.4. Comparing complexity across pods

The three pods of the NRKW A clan each have specific communication systems (despite overlaps in call repertoires) (Ford, 1987), and this could be an opportunity to test the reliability of vocal complexity metrics and how they might link to social complexity (if we consider group size a relevant proxy to measure it Peckre et al., 2019). In 2014, the A1, A4 and A5 pods each consisted of 22, 16, and 12 living individuals (Towers et al., 2015). In 2019, the same pods grew to 24, 19 and 15 living individuals respectively (Towers, 2020). We do not have the exact birth dates of the new individuals, but pods grew similarly (either 2 or 3 additional individuals). We thus assume that the pods conserved a relatively stable size difference between 2014 and 2019, and for simplicity, we chose to retain the pod sizes of 2019 to approximate social complexity for the following analysis. Hence, we can plot vocal complexity as a function of approximate group size, summing pod sizes in cases of pod aggregation (Fig. 9).

To enable a fair comparison and statistical significance tests, we randomly sampled the same number of sequences for each pod(s) (either identified manually in 2016, or automatically with the N09 call variants). The policy employed was to remove groups that had a too small amount of samples (A4 and A4–A5 for manual attributions and A4–A5 for automated ones), and to use half of the remaining smallest group's size as the sample size to compute complexity metrics (49 and 106 for manual and automated identifications respectively). For each group (pod or pod aggregation), we thus randomly sampled sequences to compute complexity metrics, and repeated the operation 100 times to get a distribution for these measurements. Results are presented in Fig. 9 in violin plots that were generated using the Matplotlib Python package (Hunter, 2007).

We used the Kruskal–Wallis H-test to assess if, across pods, complexity distributions were significantly different or not (parametric tests such as ANOVA could not be used because some data were not normally distributed). We tested each group pairs (using the SciPy Python package Virtanen et al., 2020), considering that distributions are different if the *P* value is less than 0.001, and report results in Tables 3 and 4. We used a *P* value threshold of 0.001 instead of 0.05, because the non-parametric statistical test used is more permissive than others like ANOVA for instance.

A large majority of the statistical tests (49 out of 64, or 77%) showed a significant difference of complexity distributions congruent with the approximate pod size increase (a higher complexity distributions for the larger group of individuals). For 11% of the tests, distributions were non-significantly different, and the remaining were significantly different but with opposite directions of differences (a lower complexity distribution for bigger pods).

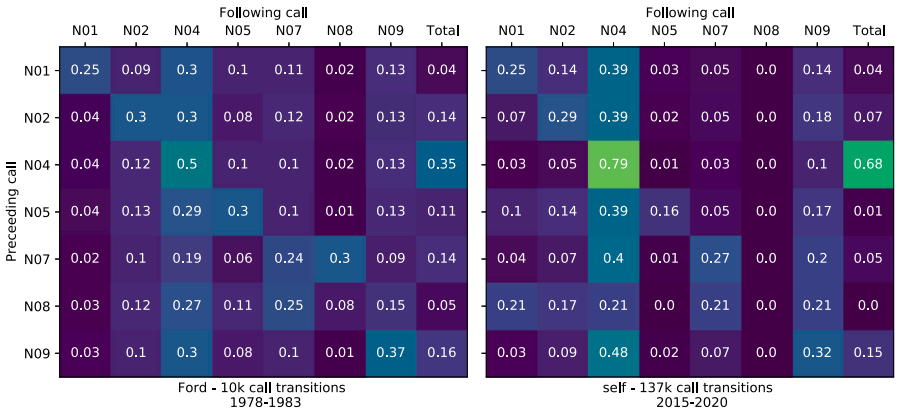


Fig. 8. Comparison between the transition matrix from Ford (1989) (left) and the present study (right). The 'Total' columns denote the proportion of each call in the dataset. Only calls present in the two studies are reported, with N09i, N09ii, and N09iii subtypes grouped into a single N09 type.

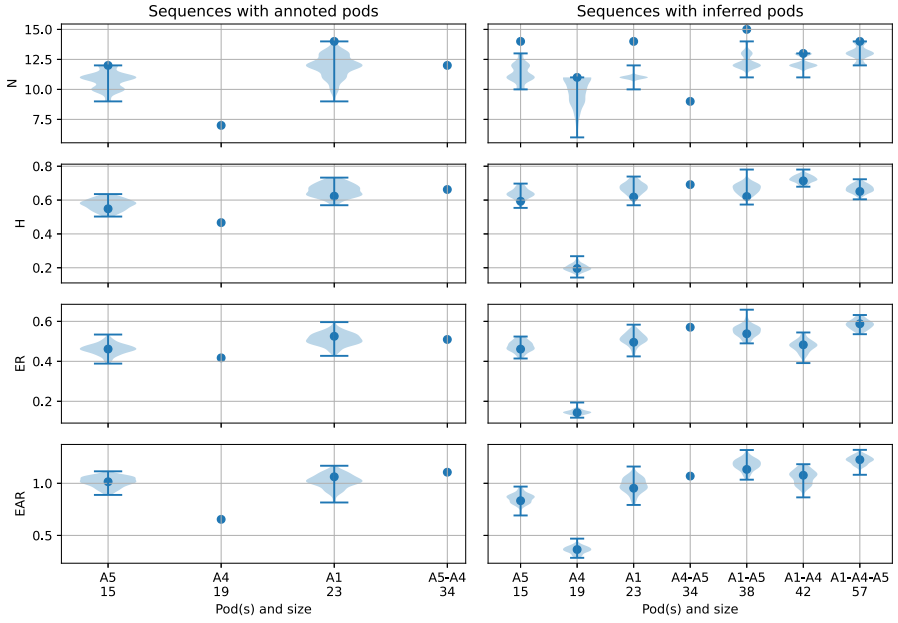


Fig. 9. Complexity metric distributions across pods manually identified (left) and automatically identified (right). For groups that had a too few observations for sampling (A4 and A5–A4 in annotated pods and A4–A5 in inferred pods), only the global complexity score is shown with a point. The results of statistical tests for distribution differences are reported in Tables 3 and 4. Points denote complexity measurements when gathering all available sequences.

3.5. Comparing complexity across behaviour and locations

One of OrcaLab's 6 hydrophones is located at Main Rubbing Beach (Fig. 1), a 500 m relatively shallow section of shoreline where NRKWs have been observed rubbing their bodies against pebbles regularly. Ford (1989) did not observe a complete change in call type probability distribution during this activity as compared to travelling, socialising or foraging, but complexity measures could still differ. In Fig. 10, we show complexity measurements across recording location, along complexity across manually annotated behaviours (only available for the summer of 2016). We used the same sampling approach as for the comparison across pods: half of the smallest group's size is sampled out of each group 100 times to compute complexity metrics, which represents 11 and 628 sequences for behaviour and location respectively.

Again, we used the Kruskal–Wallis H-test to assess if, across behaviour or locations, complexity distributions were significantly different or not, and report results in Tables 5 and 6. Out of all statistical tests for significant difference of complexity distribution across behaviours, 25% (6 out of 24) were non-significant. As for comparisons across locations, 18% (11 out of 60) of tests were non-significant.

The majority of pair-wise comparisons of vocal complexity showed significant differences, both when testing different behaviours or for locations. This highlights the fact that the vocal behaviour of this species is highly variable, and that its complexity does not only depend on the composition of the group. If a different usage of the communication system could be expected during different behaviours, the significant differences at different locations is more surprising. It is however possible that the orcas do not behave uniformly in the recording area, for instance due to variations in prey availability, water depth, current strengths, or even specific land features such as at Rubbin Beach.

3.6. Correlation tests

As a further assessment of a potential link between group size and vocal complexity, we computed the Pearson correlation coefficient between group sizes and the different complexity metrics (using a data point per passage, $N = 3,181$ with automated pod attribution). For N, E, ER and EAR, we found moderate correlations (respectively $R =$ 0.51, 0.36, 0.40 and 0.48), all with P values below 0.001 for the null-hypothesis that the two variables are uncorrelated. As a control, we

Table 4
Pair-wise comparison (Kruskal–Wallis H-test) of complexity across automatically inferred pods. Distributions compared are shown in Fig. 9. Symbols indicate the direction of the distribution difference when the *P* value<0.001 ('+' being a congruent increase of complexity with approximate pod size).

		A4	A1	A1–A5	A1–A4	A1–A4–A5
A5	N	–	ns	+	+	+
	H	–	+	+	+	+
	ER	–	+	+	ns	+
	EAR	–	+	+	+	+
A4	N		+	+	+	+
	H		+	+	+	+
	ER		+	+	+	+
	EAR		+	+	+	+
A1	N			+	+	+
	H			ns	+	ns
	ER			+	–	+
	EAR			+	+	+
A1–A5	N				ns	+
	H				+	ns
	ER				–	+
	EAR				–	+
A1–A4	N					+
	H					–
	ER					+
	EAR					+

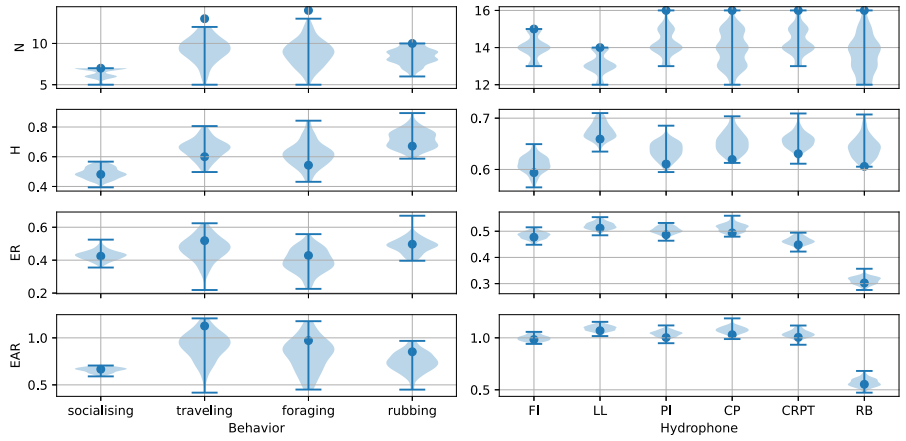


Fig. 10. Complexity metric distributions across annotated behaviours (left) and location (right). The results of statistical tests for distribution differences are reported in Tables 5 and 6. Points denote complexity measurements when gathering all available sequences.

Table 5
Pair-wise comparison (Kruskal–Wallis H-test) of complexity across annotated behaviours. Distributions compared are shown in Fig. 10. Symbols indicate the direction of the distribution difference when the *P* value<0.001 ('+' meaning an increase from row to column).

		Travelling	Foraging	Rubbing
Socialising	N	+	+	+
	H	+	+	+
	ER	+	ns	+
	EAR	+	+	+
Travelling	N		ns	–
	H		ns	+
	ER		–	ns
	EAR		–	–
Foraging	N			ns
	H			+
	ER			+
	EAR			ns

Table 6
Pair-wise comparison (Kruskal–Wallis H-test) of complexity across location. Distributions compared are shown in Fig. 10. Symbols indicate the direction of the distribution difference when the *P* value<0.001 ('+' meaning an increase from row to column).

		LL	PI	CP	CRPT	RB
FI	N	–	ns	ns	ns	–
	H	+	+	+	+	+
	ER	+	+	+	–	–
	EAR	+	+	+	+	–
LL	N		+	+	+	+
	H		–	–	–	–
	ER		–	–	–	–
	EAR		–	–	–	–
PI	N			ns	ns	–
	H			+	+	ns
	ER			ns	–	–
	EAR			+	ns	–
CP	N				ns	ns
	H				ns	–
	ER				–	–
	EAR				–	–
CRPT	N					–
	H					–
	ER					–
	EAR					–

Table 7
Pearson correlation coefficients (*R*) between each complexity metric and two dependent variables (group size and year).

	N	E	ER	EAR
group size	0.51	0.36	0.40	0.48
year	−0.10	−0.06	−0.07	−0.10

also tested for a correlation between recording dates and complexity, which for all metrics gave scores between −0.1 and −0.06, also with *P* values below 0.001 (Table 7).

4. Discussion

4.1. Deep learning and passive acoustic monitoring

This study demonstrates how deep learning and passive acoustics can contribute to the understanding of communication systems for wild ranging animals without disturbance. Such an automated approach enables us to search for potential structures in vocalisation sequences at large spatio-temporal scales. Moreover, the large-scale multi-emitter vocal sequence transcription allows to test hypotheses on proximate drivers of communication complexity, which in turn can give insights on ultimate evolutionary drivers such as with the social complexity hypothesis. Nonetheless, this study would not have been possible without the knowledge gained from the long-term audio/visual human monitoring of this population, which brought information on pod compositions and repertoires.

The automated approach presented here is not fully reliable yet. False detections and misclassification occur (Figs. 5 and 6), and could bias the subsequent complexity measurements. Missed calls especially, which are relatively frequent (Fig. 5), might alter type transition probabilities, but all complexity metrics except the ER do not rely on them. As Figs. 3 and 5 suggest also, the detection system mostly misses 'background' calls (of significantly lower energy than the detected sequence). Thus possibly, either the transcribed sequences only take a subpart of the group's emitters into account, or another more distant group is responsible for background calls not included in the transcriptions.

Taking these detection flaws into account, there is no a priori reason for the automated system to bias sequence transcription and complexity measurements for some groups but not others. Thus, we argue that these limitations on the vocal transcription procedure do not hinder the claims based on the generated data.

4.2. Assessment of structure in call sequences

First, we show that call type transition probabilities resulting from automatic classification are relatively similar to that of manual classification observed by Ford (1989) three decades beforehand (Fig. 8). The transition from N08 to N01 stands out both from the difference of bi-gram transition matrices and from the tri-gram transition matrix (Sup. Fig. 1 and 2), but the confusion matrix (Fig. 2) does not indicate frequent misclassification in this regard (Fig. 6), suggesting that the anomaly is likely due to the low amount of samples for the N08 call (30 occurrences, see Fig. 7). Besides this singularity, the similarity between the two transition matrices measured three decades apart highlights some stability of the communication system over time, but also, along with manual validations of the classification model (Figs. 5 and 6), demonstrates the relative reliability of the proposed automated method. Note that here, the temporal stability of the communication system is shown only for type transition probabilities, which overlooks within-type call structure evolution which has been shown for the same population (Deecke et al., 2000).

A second result of this study is that the orcas' propensity for call type repetition or vocal matching, found in previous studies (Ford, 1989; Miller et al., 2004; Kershenbaum et al., 2014), can be observed even with intermediate calls. This is seen in the transition probabilities for tri-grams starting with N01 and N09 calls (Supplementary Fig. 2). Particularly, the emission of a call type increases its probability of occurrence even if another type is emitted in between (a 'N09 N04 N09' sequence for instance). This observation could result from the current impossibility of diarisation (we do not know the individual emitter of each call), for instance if an individual only repeats a single call but another individual's calls are emitted in-between. Another possibility is that an underlying behavioural state of the group is responsible for an increased probability for a specific call to be emitted, or that the emission of a specific type "spreads contagiously among group members" (Ford, 1989), *i.e.*, vocal matching (Miller et al., 2004).

Despite the large amount of data analysed, no specific call-association pattern significantly stood out such as those found in Icelandic killer whales (Selbmann et al., 2023), but this could be due to the nature of this studies material which mixes calls potentially from different emitters.

4.3. Vocal complexity measurements

Besides transition probabilities, the call type sequences generated automatically allowed us to measure several metrics of communicative complexity. Unlike many studies of communication complexity which operate on single-emitter utterances, multiple individuals are potentially responsible for the vocal sequences studied here. We argue nonetheless, that studying group-level communicative complexity is relevant to better understand vocal behaviours.

First, the Zipf analysis conducted on the call type counts in detected sequences resulted in a broken power law distribution (Fig. 7). The first part has a slope (PLC) of −1.1, very close to that of a precedent similar analysis also conducted on orcas (Kershenbaum et al., 2021). However, the tail of the distribution with the least detected calls has a much steeper slope (−14.9), which is probably due to the sample size being not sufficient for rare calls (a common difficulty even in human corpora Kershenbaum et al., 2021). Accounting for these potential limits of Zipfian analysis, we turned to a more 'Shannonian' paradigm to compare communicative complexity across social, spatial, and behavioural parameters.

In this study, we report both common complexity metrics (repertoire size and Shannon entropy) along with more advanced ones such as the entropy rate which measures the predictability of type transitions, and the entropy of association rates which measures the predictability of type combinations within sequences regardless of their position. This proposal for a new metric of vocal sequence complexity is motivated by the common difficulty or impossibility to diarise (discriminate between individual emitters) vocal sequences recorded passively, which can strongly alter call transition probabilities used in former complexity metrics (Kershenbaum, 2014). However, thanks to advanced localisation techniques (Rhinehart et al., 2020), in some cases, individual diarisation is possible (Poupard et al., 2021). If vocal sequences were known to be emitted from a single individual exclusively, call transition probabilities become more reliable, but association rates might still give relevant information on a longer-term combinatorial complexity for a given communication system (as significant associations might occur as long range dependencies Ferrer-i Cancho and McCowan, 2012).

The large scale of this study also enabled a comparison of these measurements across different parameters, namely the pod which emitted the sequences (thanks to both a manual monitoring and classifying N09 variants), but also the surface behaviour (via visual monitoring), and the location of the recording. We compared complexity distributions of each condition using statistical pair-wise tests. They showed that in 77% of the group-wise comparison, the complexity distribution of multi-emitter vocal sequences increased with the number of individuals (Tables 3 and 4). Moreover, the pair-wise tests were further supported by Pearson tests, correlation coefficients between group size and vocal complexity being all positive with moderate values ($0.13 < R^2 < 0.27$ depending on the metric used, Table 7). As a comparison, testing for a similar relationship between social and communicative complexity, but comparing species and not intraspecific groups as in this study, Blumstein and Armitage (Blumstein and Armitage, 1997) found a R^2 of 0.40 in scuirids, and May-Collado et al. (2007) found a R^2 of 0.08 in cetaceans.

The number of individuals in each pod used here only gives a rough estimation of social complexity, also accounting for the fact that we only approximated it with the 2019 population count. Social complexity could be better measured as the diversity of dyadic interactions for instance, if this type of data was available. Specifically, similar measures to those used here to measure vocal complexity can be used to measure the complexity of social organisations (Rebout et al., 2021). The link between social and communicative complexity across pods, as tested here, looks at a very different scale to the cross-species comparisons usually done. Thus, our findings do not reinforce any hypothesis on evolutionary drivers that might have stimulated a co-evolution of communicative and social behaviours.

Nonetheless, such an observation opens a new perspective on the social complexity hypothesis. Indeed, within one specific population, when considering multi-emitter sequences, vocal complexity measurements vary significantly. This suggests that evolution is not the only driver at play, and that other shorter-term parameters are also to be considered. Varying factors could explain this increase in vocal complexity linked to group size.

It might result from the fact that here, vocal sequences could not be diarised, and that in such a context, complexity metrics could be influenced by the number of vocalisation emitters. A larger vocalising group does not necessarily imply a greater complexity of the multi-emitter vocal sequences. Several emitters could be using the same call type, for instance in the case of vocal matching (King and McGregor, 2016) which has been observed in NRKWs (Miller et al., 2004). It cannot be excluded also that not all individuals vocalise in a group, resulting in similar numbers of emitters even for pods of different sizes. Nonetheless, the increases of vocal complexity with pod size observed in this study is most probably to be attributed to a larger number of emitters. Since the global repertoire size of each pod is almost equal across the three studied pods: 14 for A1 and A4, and 13 for A5 (Ford, 1987), the hypothesis of a more complex individual-wise communication system seems unlikely.

Furthermore, orca pods show a dialectic behaviour, each having specific communication systems that evolve culturally (Deecke et al., 2000). This phenomenon could also impact vocal complexity measurements, alone or in combination with the number of vocalisation emitters. We thus suggest to account for these factors when testing the social complexity hypothesis in vocal learning species, as cultural phenomena can occur at a shorter-term than species' evolution.

Other factors than group size such as behaviour and location also appeared to have an impact on vocal complexity measurements (Fig. 10 and Tables 5 and 6). The impact of behaviour on vocalisation emission had previously been observed for the same species (Ford, 1989; Filatova et al., 2013), but not in terms of vocal complexity. A large majority of the pair-wise tests comparing vocal complexity across behaviour and location showed significant differences between distributions (75% for behaviours and 82% for locations). This shows that group composition is not the only factor influencing vocal complexity measurements. It might seem paradoxical that during socialising behaviour, vocal complexity is lower than during other behaviours, but the low amount of samples for this behaviour (only 22 available sequences) might explain the distribution. Possibly also, the socialisation categorised as such by human observers is more based on visual than on acoustic cues. Different locations could be a proxy of behaviour if they occur more often in certain places than others, and as behaviour can impact call type probability distribution (Ford, 1989), it is plausible that complexity measurements are affected as well. This supports the incentive proposed by Peckre et al. (2019) to study other mechanisms than social complexity as potential underlying contributors of vocal complexity.

Across all pair-wise comparisons, two sets appeared to particularly stand out from the rest: the A4 pod and the Rubbing Beach location (Fig. 9 and Fig. 10 respectively). As for the Rubbing Beach location, 13 of the 16 observed false sequences being from this location, and 12 of these 13 sequences consisting of a single call (see Section 3.1), it is most probable that a malfunction of the call detection system is responsible for the drop in vocal complexity measures (note that this drop is only noticeable in ER and EAR measurements, which would be the metrics most affected by a highly repetitive sequences).

As for the A4 pod, a clear explanation on this significant drop is lacking. The low sample size for this set could not be the only responsible factor since the A1–A4 pod aggregation has a similarly small number of sequences available (213 and 261 respectively). However, passages with only the A4 pod occurred only among 18 separate days

(as compared to 36 for the A1–A4 pod aggregation). Thus a possible explanation for this drop is that the observed data are not representative of the group's regular behaviour. It should also not be excluded that this group's communication system, as for this measurement methodology, is less complex than the others'.

Focusing on the A clan of the NRKWs, a population with a specific vocal culture and social structure, this study showed that multi-emitter vocal sequence complexity varies across groups. As other studies have shown that social organisation and vocal behaviour varies greatly across orca populations, especially depending on diet preference (Ford and Ellis, 2014; Beck et al., 2012), links between social factors and communicative complexity are also to be expected. Again, this observation calls for consideration of all ecological factors to explain communicative complexity and to test the social complexity hypothesis.

5. Conclusion

This study is the first to report an empirical analysis of vocal complexity in relation to sociality in orcas. Despite not looking at the same scale as what is commonly done in social complexity hypothesis studies, results are, for the most part, in agreement: vocal complexity correlates with social complexity. With the current data however, it is not possible to disentangle contributions from communication systems themselves, and the potentially varying number of emitters in the recorded sequences.

This study's results thus highlight limitations and precautions to be had in conducting these tests, as many different factors might impact vocal complexity measurements in passive acoustic monitoring settings. As other contributions, this paper shows how different metrics can be complementary in measuring communicative complexity (their consistency strengthens results, and inconsistencies can indicate false assumptions). We also propose the Entropy of Association Rates (EAR) as a means to cope with the cocktail party problem, which is common when studying vocal behaviours with a passive acoustic setting.

As for future works, passive acoustic recordings hold potential to develop our understanding of non-human communication systems, but research efforts are needed to contextualise the information they provide. The contextual information that appears most needed for studies similar to this one is the identity of the emitter, and the current behaviour.

CRediT authorship contribution statement

Paul Best: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marion Poupard:** Writing – review & editing, Investigation, Data curation. **Ricard Marxer:** Writing – review & editing, Supervision, Formal analysis, Investigation, Methodology. **Paul Spong:** Data collection, Writing – review & editing, Validation. **Helena Symonds:** Data collection, Writing – review & editing, Validation. **Hervé Glotin:** Writing – review & editing, Supervision, Project administration, Formal analysis, Investigation.

Financial disclosure

Hervé Glotin received the grants ANR-21-CE04-0019, ANR-21-CE04-0020 and ANR-20-CHIA-0014, and Ricard Marxer received the grant ANR-20-CE23-0012-01 from Agence Nationale de la Recherche.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Herve Glotin reports financial support was provided by French National Research Agency. Ricard Marxer reports financial support was provided by French National Research Agency. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to sincerely thank the OrcaLab team, without whom we would know so little about the NRKW population.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103211.

Data availability

A sample of call detections and metadata such as date and location used in the experiments are made publicly accessible, along with acoustic recordings https://zenodo.org/records/15462307.

References

Beck, Suzanne, Kuningas, Sanna, Esteban, Ruth, Foote, Andrew D, 2012. The influence of ecology on sociality in the killer whale (Orcinus orca). Behav. Ecol. 23 (2), 246–253.

Bergler, Christian, Smeele, Simeon Q, Tyndel, Stephen A, Barnhill, Alexander, Ortiz, Sara T, Kalan, Ammie K, Cheng, Rachael Xi, Brinkløv, Signe, Osiecka, Anna N, Tougaard, Jakob, et al., 2022. ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning. Sci. Rep. 12 (1), 21966.

Best, Paul, Ferrari, Maxence, Poupard, Marion, Paris, Sébastien, Marxer, Ricard, Symonds, Helena, Spong, Paul, Glotin, Hervé, 2020. Deep learning and domain transfer for orca vocalization detection. In: 2020 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–7.

Bigg, MA, Olesiuk, PF, Ellis, Graeme M, Ford, JKB, Balcomb, Kenneth C, 1990. Social organization and genealogy of resident killer whales (Orcinus orca) in the coastal waters of British Columbia and Washington state. 12, pp. 383–405, Report of the International Whaling Commission.

Blumstein, Daniel T., Armitage, Kenneth B., 1997. Does sociality drive the evolution of communicative complexity? A comparative test with ground-dwelling sciurid alarm calls. Amer. Nat. 150 (2), 179–200.

Bouchet, Hélène, Blois-Heulin, Catherine, Lemasson, Alban, 2013. Social complexity parallels vocal complexity: a comparison of three non-human primate species. Front. Psychol. 4, 390.

Cairns, Sara J., Schwager, Steven J., 1987. A comparison of association indices. Anim. Behav. 35 (5), 1454–1469.

Ferrer-i Cancho, Ramon, McCowan, Brenda, 2012. The span of correlations in dolphin whistle sequences. J. Stat. Mech. Theory Exp. 2012 (06), P06002.

Coffey, Kevin R., Marx, Ruby E., Neumaier, John F., 2019. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology 44 (5), 859–868.

Deecke, Volker B., Ford, John K.B., Spong, Paul, 2000. Dialect change in resident killer whales: implications for vocal learning and cultural transmission. Anim. Behav. 60 (5), 629–638.

Escobar-Amado, Christian, Badiey, Mohsen, Wan, Lin, 2023. Computer vision for bioacoustics: Detection of bearded seal vocalizations in the Chukchi shelf using YOLOV5. IEEE J. Ocean. Eng..

Fagen, R.M., Young, D.Y., 1978. Temporal patterns of behaviors: durations, intervals, latencies, and sequences. Quant. Ethol. 79–114.

Filatova, Olga A, Fedutin, Ivan D, Burdin, Alexandr M, Hoyt, Erich, 2007. The structure of the discrete call repertoire of killer whales orcinus orca from Southeast Kamchatka. Bioacoustics 16 (3), 261–280.

Filatova, OA, Guzeev, MA, Fedutin, ID, Burdin, AM, Hoyt, Erich, 2013. Dependence of killer whale (*Orcinus orca*) acoustic signals on the type of activity and social context. Biol. Bull. 40 (9), 790–796.

Filatova, Olga A, Samarra, Filipa IP, Deecke, Volker B, Ford, John KB, Miller, Patrick JO, Yurk, Harald, 2015. Cultural evolution of killer whale calls: background, mechanisms and consequences. Behaviour 152 (15), 2001–2038.

Ford, JKB, 1987. A catalogue of underwater calls produced by killer whales (*Orcinus orca*) in British Columbia. Canadian Data Report of Fisheries and Aquatic Sciences no. 633, Fisheries Research Branch, Pacific Biological Station.

Ford, John K.B., 1989. Acoustic behaviour of resident killer whales (*Orcinus orca*) off Vancouver Island, British Columbia. Can. J. Zool. 67 (3), 727–745.

Ford, John K.B., Ellis, Graeme M., 2014. You are what you eat: foraging specializations and their influence on the social organization and behavior of killer whales. In: Primates and Cetaceans: Field Research and Conservation of Complex Mammalian Societies. Springer, pp. 75–98.

Freeberg, Todd M., Dunbar, Robin I.M., Ord, Terry J., 2012. Social complexity as a proximate and ultimate factor in communicative complexity. Phil. Trans. R. Soc. B 367 (1597), 1785–1801.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9 (3), 90–95.

Janik, Vincent M., 2014. Cetacean vocal learning and communication. Curr. Opin. Neurobiol. 28, 60–65.

Kershenbaum, Arik, 2014. Entropy rate as a measure of animal vocal complexity. Bioacoustics 23 (3), 195–208.

Kershenbaum, Arik, Bowles, Ann E, Freeberg, Todd M, Jin, Dezhe Z, Lameira, Adriano R, Bohn, Kirsten, 2014. Animal vocal sequences: not the Markov chains we thought they were. Proc. R. Soc. B Biol. Sci. 281 (1792), 20141370.

Kershenbaum, Arik, Demartsev, Vlad, Gammon, David E, et al., 2021. Shannon entropy as a robust estimator of Zipf's law in animal vocal communication repertoires. Methods Ecol. Evol. 12 (3), 553–564.

King, Stephanie L., McGregor, Peter K., 2016. Vocal matching: the what, the why and the how. Biol. Lett. 12 (10), 20160666.

Kroodsma, Donald E., 1977. Correlates of song organization among North American wrens. Amer. Nat. 111 (981), 995–1008.

May-Collado, Laura J., Agnarsson, Ingi, Wartzok, Douglas, 2007. Phylogenetic review of tonal sound production in whales in relation to sociality. BMC Evol. Biol. 7, 1–20.

McCowan, Brenda, Doyle, Laurance R., Hanser, Sean F., 2002. Using information theory to assess the diversity, complexity, and development of communicative repertoires. J. Comp. Psychol. 116 (2), 166.

McCowan, Brenda, Doyle, Laurance R, Jenkins, Jon M, Hanser, Sean F, 2005. The appropriate use of Zipf's law in animal communication studies. Anim. Behav. 69 (1), F1–F7.

Miller, Patric JO, Shapiro, AD, Tyack, Peter Lloyd, Solow, AR, 2004. Call-type matching in vocal exchanges of free-ranging resident killer whales, orcinus orca. Anim. Behav. 67 (6), 1099–1107.

Parcerisas, Clea, Schall, Elena, Te Velde, Kees, Botteldooren, Dick, Devos, Paul, Debusschere, Elisabeth, 2024. Machine learning for efficient segregation and labeling of potential biological sounds in long-term underwater recordings. Front. Remote. Sens. 5, 1390687.

Peckre, Louise, Kappeler, Peter M., Fichtel, Claudia, 2019. Clarifying and expanding the social complexity hypothesis for communicative complexity. Behav. Ecol. Sociobiol. 73, 1–19.

Poupard, Marion, Best, Paul, Schlüter, Jan, Symonds, Helena, Spong, Paul, Glotin, Hervé, 2019a. Large-scale unsupervised clustering of Orca vocalizations: a model for describing orca communication systems. PeerJ Preprints 7, e27979v1. http://dx.doi.org/10.7287/peerj.preprints.27979v1.

Poupard, Marion, Best, Paul, Schlüter, Jan, et al., 2019b. Deep learning for ethoacoustics of oreas on three years pentaphonie continuous recording at orealab revealing tide, moon and diel effects. In: OCEANS 2019-Marseille. IEEE, pp. 1–7.

Poupard, Marion, Symonds, Helena, Spong, Paul, Glotin, Hervé, 2021. Intra-group orca call rate modulation estimation using compact four hydrophones array. Front. Mar. Sci. 8, 681036. http://dx.doi.org/10.3389/fmars.2021.681036.

Rebout, Nancy, Lone, Jean-Christophe, De Marco, Arianna, Cozzolino, Roberto, Lemasson, Alban, Thierry, Bernard, 2021. Measuring complexity in organisms and organizations. R. Soc. Open Sci. 8 (3), 200895.

Redmon, Joseph, Divvala, Santosh, Girshick, Ross, Farhadi, Ali, 2016. 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, In: You Only Look Once: Unified, Real-Time Object Detection, vol. 2, pp. 779–788. http://dx.doi.org/10.1109/CVPR.2016.91.

Rhinehart, Tessa A, Chronister, Lauren M, Devlin, Trieste, Kitzes, Justin, 2020. Acoustic localization of terrestrial wildlife: Current practices and future opportunities. Ecol. Evol. 10 (13), 6794–6818.

Riesch, Rüdiger, Ford, John K.B., Thomsen, Frank, 2008. Whistle sequences in wild killer whales (*Orcinus orca*). J. Acoust. Soc. Am. 124 (3), 1822–1829.

Saulitis, Eva L., Matkin, Craig O., Fay, Francis H., 2005. Vocal repertoire and acoustic behavior of the isolated AT1 killer whale subpopulation in southern Alaska. Can. J. Zool. 83 (8), 1015–1029.

Selbmann, Anna, Miller, Patrick JO, Wensveen, Paul J, Svavarsson, Jörundur, Samarra, Filipa IP, 2023. Call combination patterns in Icelandic killer whales (orcinus orca). Sci. Rep. 13 (1), 21771.

Stowell, Dan, 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152.

Suzuki, Ryuji, Buck, John R., Tyack, Peter L., 2006. Information entropy of humpback whale songs. J. Acoust. Soc. Am. 119 (3), 1849–1866.

Towers, Jared R., 2020. Photo-Identification Catalogue and Status of Northern Resident Killer Whale Population in 2019. Fisheries and Oceans Canada, Pacific Region.

Towers, Jared R., Ellis, Graeme, Ford, John Kenneth Baker, 2015. Photo-Identification Catalogue and Status of the Northern Resident Killer Whale Population in 2014/by JR Towers, GM Ellis and JKB Ford. <bound method Organization. get_name_with_acronym of<Organization ....

Tyack, Peter L., 2012. Review of the signature-whistle hypothesis for the Atlantic bottlenose dolphin 10. Bottlenose Dolphin 199.

Virtanen, Pauli, Gommers, Ralf, Oliphant, Travis E., et al., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Xie, Jiangjian, Zhong, Yujie, Zhang, Junguo, Liu, Shuo, Ding, Changqing, Triantafyllopoulos, Andreas, 2023. A review of automatic recognition technology for bird vocalizations in the deep learning era. Ecol. Inform. 73, 101927.

Yu, Shuiyuan, Xu, Chunshan, Liu, Haitao, 2018. Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. arXiv preprint arXiv:1807.01855.

Yurk, Harald, Barrett-Lennard, Lance, Ford, J.K.B., Matkin, C.O., 2002. Cultural transmission within maternal lineages: vocal clans in resident killer whales in southern Alaska. Anim. Behav. 63 (6), 1103–1119.

Zhong, Ming, Castellote, Manuel, Dodhia, Rahul, Lavista Ferres, Juan, Keogh, Mandy, Brewer, Arial, 2020. Beluga whale acoustic signal classification using deep learning neural network models. J. Acoust. Soc. Am. 147 (3), 1834–1841.

Zipf, George Kingsley, 1949. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison- Wesley Press.