Quantifying group specificity of animal vocalizations without specific sender information

Heike Vester,[1,*] Kurt Hammerschmidt,[2,†] Marc Timme,[3,‡] and Sarah Hallerberg[3,§]
[1]*Ocean Sounds, Sauoya 01, 8312 Henningsvaer, Norway*
[2]*Cognitive Ethology Lab, German Primate Center, Kellnerweg 4, 37077 Göttingen, Germany*
[3]*Network Dynamics, Max Planck Institute for Dynamics and Self-Organization, 37077 Göttingen, Germany*
(Received 8 October 2015; published 25 February 2016)

Recordings of animal vocalization can lack information about sender and context. This is often the case in studies on marine mammals or in the increasing number of automated bioacoustics monitorings. Here, we develop a framework to estimate group specificity without specific sender information. We introduce and apply a bag-of-calls-and-coefficients approach (BOCCA) to study ensembles of cepstral coefficients calculated from vocalization signals recorded from a given animal group. Comparing distributions of such ensembles of coefficients by computing relative entropies reveals group specific differences. Applying the BOCCA to ensembles of calls recorded from group of long-finned pilot whales in northern Norway, we find that differences of vocalizations within social groups of pilot whales (*Globicephala melas*) are significantly lower than intergroup differences.

DOI: 10.1103/PhysRevE.93.022138

I. INTRODUCTION

Large-scale acoustic monitoring of wildlife through acoustic recording stations becomes a more and more common approach in many places of the world [1–4]. Although the amounts of data generated through automated recordings are huge, data available for a specific species of interests might be scarce. In this setting there is no additional information about the sender, i.e., the (individual) animal that produced the sound. These restriction also applies to recordings of marine mammals, vocalizing out of sight of the observer (i.e., under water). In the presence of these challenges there is an increasing need for fast but robust methods to infer biologically and ecologically relevant information on the basis of acoustic signals. In this contribution, we propose an ensemble-based approach to quantify group-specific differences of animal vocalizations.

As a test data set for studying the validity of the approach we use vocalizations of long-finned pilot whales (*Globicephala melas*), recorded in northern Norway [5]. In a multidimensional aquatic environment it is important to recognize group members for offspring care, protection against predators, and cooperative social and feeding behavior. The existence of vocal cultures and dialects has been suggested by observer-based analysis of variations in call type usage of killer whales [6,7] and sperm whale codas [8]. Less is known about the vocalizations of long-finned pilot whales, especially about the population living in the northeast Atlantic. Long-finned pilot whales in the *northwest* Atlantic produce typical dolphin sounds, such as clicks, buzzes, grunts, and a variety of pulsed calls including whistles [9,10]. In a previous study [5] it has been found that calls are complex with different structural components, such as elements and segments, and one-fifth of the calls we observe are biphonal with a lower (LFC) and an upper frequency component (UFC) [5,10,11]. In total about 140 different call types have been found through observer-based audiovisual classification [5].

There are several approaches to the automatic processing of cetacean vocalizations (*calls*) [12–17] or sound in general [18,19]. Most of them consist in capturing the temporal changes of selected sound features and training different classifiers to categorize and classify single call types. We, however, test whether one can study the communication of whales without categorizing and classifying *single calls* and without considering the temporal changes of the signals. Instead we propose and apply an automated analysis method, the bags-of-calls-coefficients approach (BOCCA) to ensembles (also called *bags* in a machine learning context) of calls, recordings, and sound features. Additionally, we refuse to focus only on a low-dimensional subset of selected sound features that seem relevant to the human observer, since they might be irrelevant for the sensory processing system of the animals. In more detail, we work with ensembles of all available features as computed through a cepstral decomposition of the sound signal [20]. Cepstral coefficients have been proposed as features for speech recognition [20] and in this context term *cepstral decomposition* is used in analogy to the common *spectral decomposition*, with the dependent variable called *quefrency* in analogy to *frequency*. Computing distributions of cepstral coefficients for each ensemble allow us to quantify group specificity in a statistical significant way. This approach is conceptually related to the bag-of-words approach [21].

This article is organized as follows: In Sec. II we present BOCCA, a new method to detect group specificity of animal vocalizations using ensembles of sounds and relative entropies. In the following two sections we demonstrate the relevance of this approach by applying it to quantify the group specificity of vocalizations of long-finned pilot whales. In more detail, Sec. III provides all details concerning the data set of recordings and a conventional observer-based analysis approach [overlap of call usage (OCU)]. We then present the results of the BOCCA and compare it to results of the OCU approach by computing similarity rankings in Sec. IV. We summarize the results and discuss further applications of the BOCCA in Sec. V.

*heike_vester@ocean-sounds.org
†khammerschmidt@dpz.eu
‡timme@nld.ds.mpg.de
§shallerberg@nld.ds.mpg.de

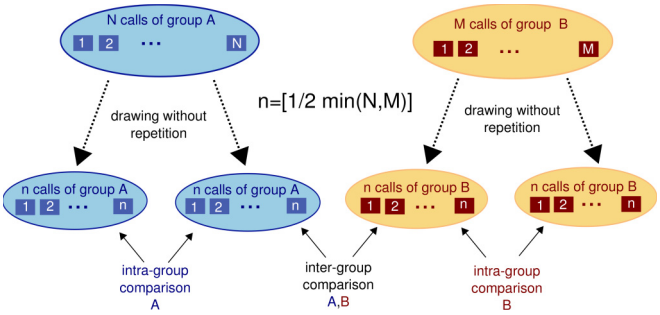

FIG. 1. How to construct ensembles (bags-of-calls) for inter- and intragroup comparisons.

II. ENSEMBLE-BASED IDENTIFICATION OF GROUP SPECIFICITY

A. Constructing ensembles for intra- and intergroup comparisons

To compare group-specific differences in vocalizations, we construct ensembles of calls (*bags-of-calls*), each representing a sample of recordings from one group of whales. The detailed procedure of constructing bags-of-calls is described as follows (see Fig. 1 for an illustration): Suppose we have N_i recordings of calls of whale group i and N_j recordings of group j. An ensemble $E_i(n)$ of n calls is drawn through random number generators of all calls that have been recorded for group i. We then randomly draw a second ensemble $E_j(n)$ consisting of n of N_j recordings of the second group j, with $j \neq i$. For each pairwise comparison of groups, the size of the ensembles n is adjusted according to the smaller number of available recordings, i.e., to the largest integer smaller than half of the smaller number of available recordings, $n = \lfloor \frac{1}{2}\min(N_i, N_j) \rfloor$. We then compare groups i and j by comparing properties of the ensembles $E_i(n)$ and $E_j(n)$. In order to check whether the resulting differences can be attributed to the difference in group, we also compare pairs of two random ensembles drawn from the same group, using the $N_i - n$ or $N_j - n$ remaining recordings (i.e., recordings not used for the previous comparison of two different groups). In other words, we generate additional ensembles $\tilde{E}_i(n)$ of size n using now only the recordings that are *not* part of ensemble $E_i(n)$. The differences in distribution within group i are then estimated by comparing properties of $E_i(n)$ and $\tilde{E}_i(n)$.

B. Computing cepstral coefficients

Several features have been proposed and studied in order to characterize bioacoustic signals [22]. Bias can be introduced by selectively choosing features that the observer considers to be relevant. We try to reduce this bias by completing the previous findings with a study based on (relatively) unselected features, i.e., cepstral coefficients [20], often used in speech processing. Unlike so-called *hand-crafted* features, they are computed for any arbitrary input signal, without requiring knowledge about the sounds under study. The only bias inherent to cepstral coefficients is the choice of the representation (FFT based) and the choice of the window lengths. Mel cepstral coefficients [23], often used to describe human vocalizations, project cepstral coefficients onto the Mel scale [24] to emphasize the frequencies most relevant for human communication. We, however, do *not* project onto the Mel scale, since frequencies that are most relevant for humans might not be *a priori* suitable for analyzing communication of animals. In more detail, we compute the coefficients

$$c(q,t) = \left\| \mathcal{F}\left\{ \log\left(\frac{\|\mathcal{F}\{x(t)\}\|^2}{2\pi g_N} \right) \right\} \right\|^2 \qquad (1)$$

of the power cepstrum as features. Here $\mathcal{F}$ denotes the Fourier transform and the normalization factor $g_N = \max(\mathcal{F}\{x(t)\})$ where $t = 0, \ldots, T$ is given by the maximum of the power spectral density of each cut recording of length T.

For a discrete time signal $x(t_n)$ with $t_n = t_0 + n\Delta t$, with $\Delta t = 1/48\,000$ the discrete Fourier transforms are realized using a Hanning window of w time steps and an overlap of $w - 1$ time steps. For a given window length w, we obtain $w/4$ cepstral coefficients $c(q,t)$, where $q = 1,2,\ldots,w/4$ denotes the index of each *quefrency*. Adapting suitable window lengths for computing *spectrograms* (heat maps of the power spectral density with respect to frequency and time instance), we obtain a high-resolution in quefrency space, i.e., in the example shown below we work with $w/4 = 128$ cepstral coefficients. This resolution is relatively large compared to human speech processing, using 13 cepstral coefficients and their temporal derivatives. Figure 2 shows a visualization of the coefficients $c_{q,t}$, in form of a *cepstrogram* in comparison to the typical spectrograms. Due to the second Fourier transformation, peaks in the cepstrum indicate the (reciprocal) difference between peaks in the spectrum. Loosely speaking, a line in the cepstrogram [representing large values of $c(q,t)$] is

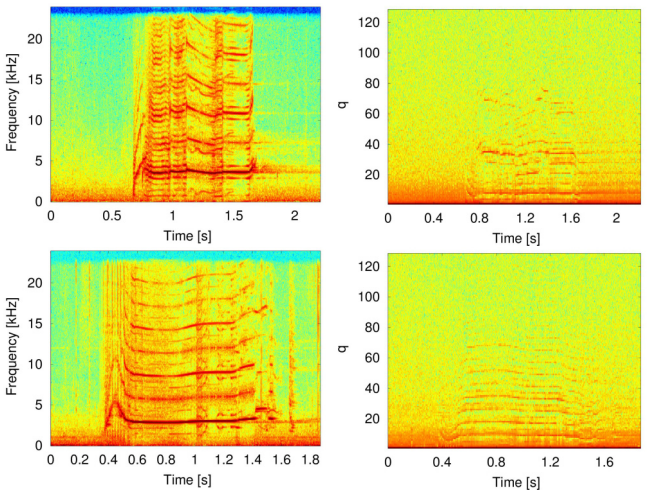

FIG. 2. Examples for spectrograms, i.e., visualizations (color coding or gray scale) of the temporal evolution of the power spectral density) (left) and cepstrograms (right) for two vocalizations of long-finned pilot whales. Loosely speaking, a line in the cepstrogram (representing large values of $c(q,t)$ is shifted upwards, whenever two lines in the spectrogram approach each other. The color coding (gray scale) of the cepstrograms refers to $\ln(c(q,t)/\max[c(q,t)])$ with $t = 0,\ldots,T$ and T being the length of the call and $q = 1,\ldots,128$ being the index of each cepstral coefficient. Spectrograms are computed with a window length of $w = 512$, allowing us to estimate $\frac{w}{2}$ spectral coefficients and $\frac{w}{4} = 128$ cepstral coefficients for each time step.

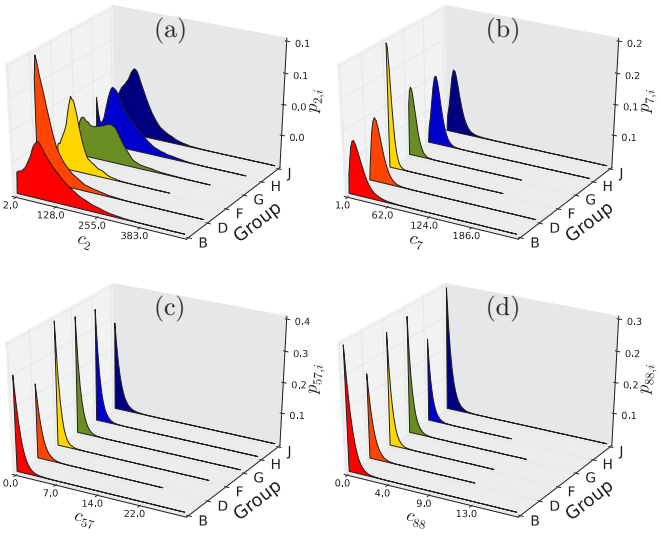

FIG. 3. Four examples for distributions of cepstral coefficients $c_i(q,t)$: (a), $p_{2,i}$ (b), $p_{7,i}$, (c) $p_{57,i}$, and (d) $p_{88,i}$ estimated for groups $i = \{B,D,F,G,H,I\}$.

shifted upwards (towards larger q), whenever two lines in the spectrogram (representing high values of the power spectral density) approach each other.

C. Comparing distributions of cepstral coefficients

For each ensemble $E_i(n)$ representing group i, we compute q time series of cepstral coefficients $\{c_i(q,t)\}$ and then estimate the distributions $p_{i,q}$ of each data set, with $q = 1,2,\ldots,128$ as specified above. Proposing that properties of the ensemble $E_i(n)$ are represented by the distribution $p_{q,i}$, we then ask in how far the 128 distributions estimated for ensemble $E_i(n)$ differ from the distributions of another ensemble $E_j(n)$. Figure 3 shows four examples for distributions of cepstral coefficients estimated for ensembles of calls recorded from different groups of whales. For some coefficients, in particular for coefficients with smaller q [e.g., Fig. 3(a)], the differences between distributions are noticeable by visual inspection, whereas many higher-order coefficients, representing small-scale fluctuations in the spectrum have more similar distributions. To quantify this observation, we use common measures for the difference in distribution, such as the (symmetrized) Kullback Leibler divergence $D(p_{q,i}||p_{q,j})$ [25], also called relative entropy and the Hellinger distance [26]. The (discrete) Kullback Leibler (relative entropy) divergence

$$D(p_{q,i}||p_{q,j}) = \sum_k \ln\left(\frac{p_{q,i,k}}{p_{q,j,k}}\right) p_{q,i,k}, \tag{2}$$

with k referring to the k-th bin of the distribution $p_{q,i}$, is a nonsymmetric measure for the difference of two distributions.. Note that $D(p_{q,i}||p_{q,j})$ is only defined if both distributions are normalized such that $\sum_k p_{q,i,k} = 1$ and if $p_{q,i,k} = 0$ implies $p_{q,j,k} = 0$ for all k. Here we use a symmetrized version of the Kullback Leibler divergence,

$$S(p_{q,i}||p_{q,j}) = D(p_{q,i}||p_{q,j}) + D(p_{q,j}||p_{q,i}), \tag{3}$$

to quantify the difference between distributions $p_{q,i}$ and $p_{q,j}$ of the q-th coefficient, representing group i and group j. Similarly, one can compute $S(p_{q,i}||\tilde{p}_{q,i})$ to measure intragroup differences by comparing the distributions $p_{q,i}$ and $\tilde{p}_{q,i}$ referring to ensembles of equal size drawn from recordings of the same group i, as explained above. We tested whether the resulting group differences are sensitive to the measure we use for comparing the distributions and obtained very similar results using the Hellinger distance [26].

To summarize and quantify the difference between inter- and intragroup comparisons we introduce the quantity

$$v_{ij} = \sum_q [S(p_{q,i}||p_{q,j}) - S(p_{q,i}||\tilde{p}_{q,i})]. \tag{4}$$

Intergroup differences are larger than intragroup differences if v_{ij} is positive and vice versa if v_{ij} is negative. Since smaller cepstral coefficients reflect large-scale structures in spectra, differences in distributions of small coefficients can be considered to be more relevant, whereas differences in higher-order coefficients can be due to small scale fluctuations and noise. Therefore, we additionally introduce a linearly weighted summary index

$$w_{ij} = \sum_q \frac{1}{q}[S(p_{q,i}||p_{q,j}) - S(p_{q,i}||\tilde{p}_{q,i})] \tag{5}$$

as a second measure for comparing inter- and intragroup differences.

To test whether calculated values of v_{ij} and w_{ij} are statistically significant we estimate confidence intervals for both coefficients by comparing randomly generated ensembles $E_r(n)$. In more detail, we construct m such ensembles by randomly drawing n recordings from all available recordings (not sorted by group) without repetition. Any 3-tuple of the resulting randomly generated not group specific ensembles E_r, $E_{r'}$, and $E_{r''}$ can serve to simulate a comparison of inter- and intragroup differences. Two random ensembles, e.g., E_r and $E_{r''}$, are interpreted as representing the same group, whereas the third one ($E_{r'}$) is assumed to represent a different group. For each 3-tuple of random ensembles, we compute the coefficients $v_{rr'}$, $v_{rr''}$ and $w_{rr'}$, $w_{rr''}$. We then use the distributions of $g(m) = \sum_{k=0}^{m-1} a(k)$ values for each coefficient [with $a(k)$ referring to triangular numbers] to estimate confidence intervals according to Student's distribution.

III. OBSERVATIONS AND RECORDINGS OF LONG-FINNED PILOT WHALES IN NORTHERN NORWAY

A. Ethics statement

All observations and recordings reported in this contribution were made in the Vestfjord in northern Norway (see GPS coordinates in Table I). In general, no permission is needed for noninvasive research on marine mammals along the Norwegian coast. To ensure that we conducted our research according to Norwegian ethical laws, we asked the Animal Test Committee (Forsksdyrutvalget) of Norway for permission, and they confirmed that our studies do not require any permission (approval paper ID 6516).

B. Encounters and recordings

We encountered six groups of long-finned pilot whales in the Vestfjord in northern Norway (Table I). Sound recordings

were made using one or two Reson TC4032 hydrophones (frequency response 5 Hz–120 kHz, omnidirectional), which were lowered directly at approximately −18 m into the water from a 7-m Zodiac boat, when in close proximity (less than 50 m) to the whales. Sound was amplified with a custom-built Etec amplifier (DK) and recorded with different mobile recording devices; in 2006–2008 we used an Edirol-R09 (Roland) with a sampling frequency of 48 kHz, and in 2009–2010 we used a Korg MR-1000 with a sampling frequency of 192 kHz. GPS coordinates were taken at the beginning and end of an encounter, and notes of the whales' behavior were continuously taken during recordings. Recordings lasted as long as the whales were within a 500 m range of the boat, and as soon as they moved out of range and the signals became weak, we stopped the recordings and moved closer to the whales. At first sign of disturbance of the whales, we ceased the studies and waited 30 min before resuming our studies. If the whales were repeatedly disturbed, then we terminated the field encounter. In most cases, however, the whales became quickly habituated to the presence of our boat and data collection was possible for longer periods.

C. Overlap in call usage

In this section we access group specificity using audiovisual (human) observer-based categorization and classification. Studying six groups of long-finned pilot whales in the Vestfjord, in northern Norway we found a complex and flexible vocal repertoire [5]. See the Supplemental Material [27] for examples of the recorded pilot whale sounds. Using a total of 32:54 h of observation time and 17:32 h of sound recordings in 99 recording sessions, 4582 recorded calls were categorized to more than 140 different types using classic audiovisual observer categorization and classification [5]. Comparing each of the six different recorded groups of pilot whales (group B, D, F, G, H, and J), we find that each group vocalized between seven (group F) and 54 (group H) different call types. Figure 4 illustrates the usage of call types among different groups of whales, based on all calls classified into 140 call types. OCU (see Fig. 4) between two groups is quite common (30 of 140 call types). Seven call types are shared among three groups and only one call type is shared between four groups. Group B and group J have the highest call type overlap ($N = 12$), and group F and group G, as well as group G and group J, have only one call type overlap. Using the OCU as a measure for similarity in vocalizations we can later compare it to the similarity of groups obtained through BOCCA (see Figs. 6 and 7).

IV. QUANTIFYING GROUP SPECIFICITY OF LONG-FINNED PILOT WHALES WITH BOCCA

A. The data set used for BOCCA

Basis of the bag-of-calls-and-coefficient analysis are short recordings that have been cut automatically or manually from continuous data, such that each of them contains one call of a long-finned pilot whale. To test whether vocalizations are group specific we only use call types of very good quality. Vocalizations of this quality originated only from utterances close to the boat and therefore ensure that only group members uttered these calls. To achieve a randomized sample of different

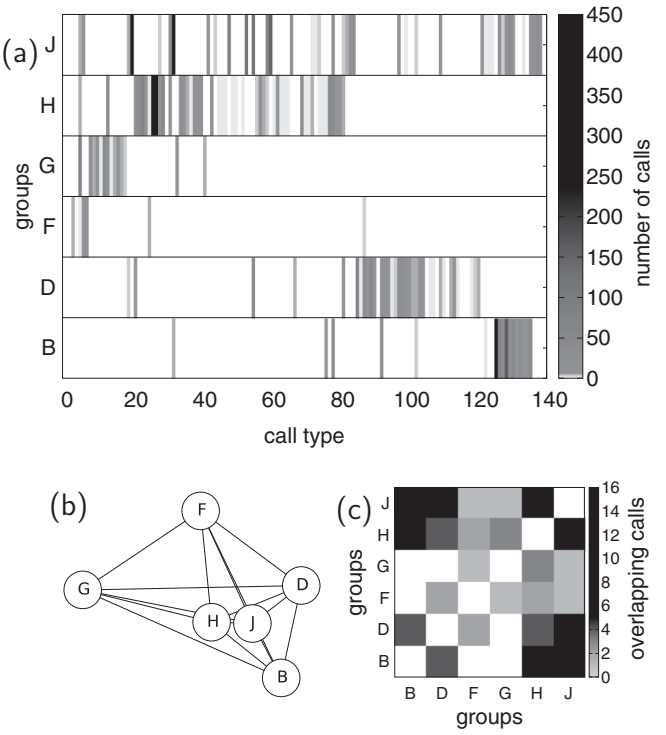

FIG. 4. Overlap in call usage. (a) Observations of calls within different groups of pilot whales. Note that call type numbering is arbitrary and does not reflect similarity in call structure, i.e., call i and $i + 1$ do not necessarily have to be similar in structure. Panels (b) and (c) show a similarity network and its adjacency matrix estimated by the OCU approach. Here the distance between edges in the network and the entries of the matrix represent the number of shared call types.

group activities and group members, we use long recordings per group (ranging from 23 min to 5:14 h, mean: 2:55 h). In total, we use 1056 nonoverlapping calls, selected from the 4582 calls according to the quality of the sound files (high signal-to-noise ratio, no boat noise). Note that this selection of mostly nonconsecutive calls (31 min long in total) is conceptually similar to a (random) sampling from all available recordings (17:32 h). The duration of the cut recordings containing one call each varies between 0.14 s and 6.27 s. The total length of all recordings used for this intra- and intergroup comparison was 31 min and 23 s. All recordings used in this part of the study have a sampling interval $\Delta t = 1/48\,000$ s.

B. Drawing ensembles of calls for inter- and intragroup comparison

Using BOCCA, we estimated group specificity on the basis of *ensembles* (or bags) of calls. Each ensemble of n calls $E_i(n)$ was drawn through random number generators from all calls that have been recorded for group i, with $i = B, D, F, G, H, J$.

For each ensemble $E_i(n)$ representing group i, with we compute time series of the q-th cepstral coefficient for each ensemble with $q = 1, 2, \ldots, 128$ and then estimate the distributions $p_{q,i}$ of each data set. Constructing ensembles $E_i(n)$ of calls and computing the features $c_{q,t}$, we can then do pairwise comparisons of ensembles of calls uttered by two different groups i and j by comparing the distributions $p_{q,i}$ and $p_{q,j}$.

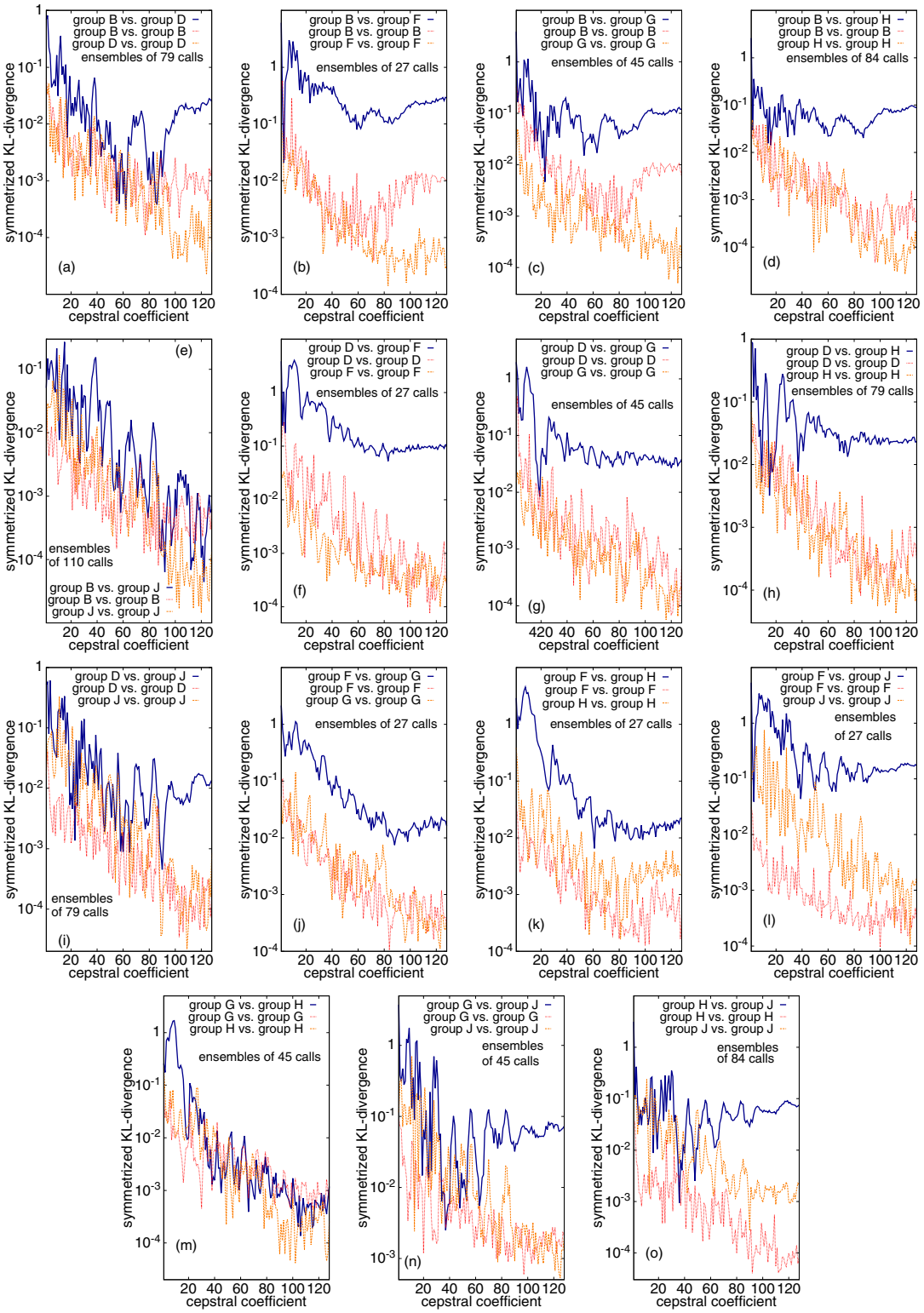

FIG. 5. Symmetrized Kullback-Leibler indicate differences in distribution of cepstral coefficients. The blue (dark) solid lines represent intragroup comparisons, whereas the red (gray) dotted and orange (light gray) dashed lines represent intergroup comparisons. Each panel shows the result for the pairwise comparison of two groups as indicated in each panel's legend. The number of calls in ensembles created for pairwise group comparisons was chosen to be the same for inter- and intragroup comparisons.

C. Comparing distributions through relative entropies

Differences between distributions are quantified by computing relative entropies, i.e., symmetrized Kullback-Leibler divergences $S(p_{q,i}||p_{q,j})$ (see Fig. 5). To relate the obtained similarity measures to a null model, we computed also intragroup comparisons. Therefore, we constructed ensembles of calls of equal size $\tilde{E}_i(n)$, $\tilde{E}_j(n)$ from the same group, and

computed distributions of coefficients for these ensembles $\tilde{p}_{q,i}$ and $\tilde{p}_{q,j}$. Intragroup comparisons are also quantified using symmetrized Kullback-Leibler divergences.

Visual inspection of KL-divergences depending on the cepstral coefficients, as displayed in Fig. 5, yields that intergroup differences are clearly larger than intragroup differences for 10 of 15 comparisons, i.e., when comparing groups B and D, groups B and D, groups B and G, groups B and H, groups D and F, groups D and G, groups D and H, groups F and G, groups F and H, and groups F and J. More attention to detail is needed to interpret the results of the comparison between groups B and J, groups D and J, groups G and H, groups G and J, and groups H and J. Comparing groups G and H [Fig. 5 (m)], we found large difference in distributions for the first coefficients than for higher-order coefficients. The smaller coefficients $(c_2, \ldots, c_{70})$ reflect large-scale structures in the spectrum, whereas higher-order coefficients capture fluctuations on a small scale, e.g., noise. Thus, the results displayed in Fig. 5(m) indicate that the intergroup difference of group G and group H is larger than intragroup differences concerning the more relevant large scale structures in the spectrum. Comparing groups B and J, there are many coefficients (especially in the vicinity of c_{40} and c_{50}) for which we found intergroup differences to be larger than intragroup differences and similarly for groups H and J. In the comparison of groups D and J, the intergroup differences [blue line in Fig. 5(i)] are about one order of magnitude larger than the intragroup differences of group D (red line). However, they are in the same order of magnitude as the intragroup differences of group J (orange line) for small q, whereas the difference was again larger for higher-order coefficients. More precisely, there are 12 of 128 coefficients for which the intragroup difference of group J exceed the intergroup difference, whereas 116 of 128 coefficients indicate that the intergroup difference is larger. Summarizing, one can therefore conclude that the overall intergroup difference between groups J and D is larger than group J's intragroup difference. Additionally, we repeat all comparisons using not the symmetrized Kullback-Leibler divergence but the Hellinger distance [26] as a measure for the similarity of distributions and obtain qualitatively and quantitatively very similar results.

D. Evaluating indices for group specificity

The example of the comparison of group J and group D indicates that the interpretation of Fig. 5 has to be done

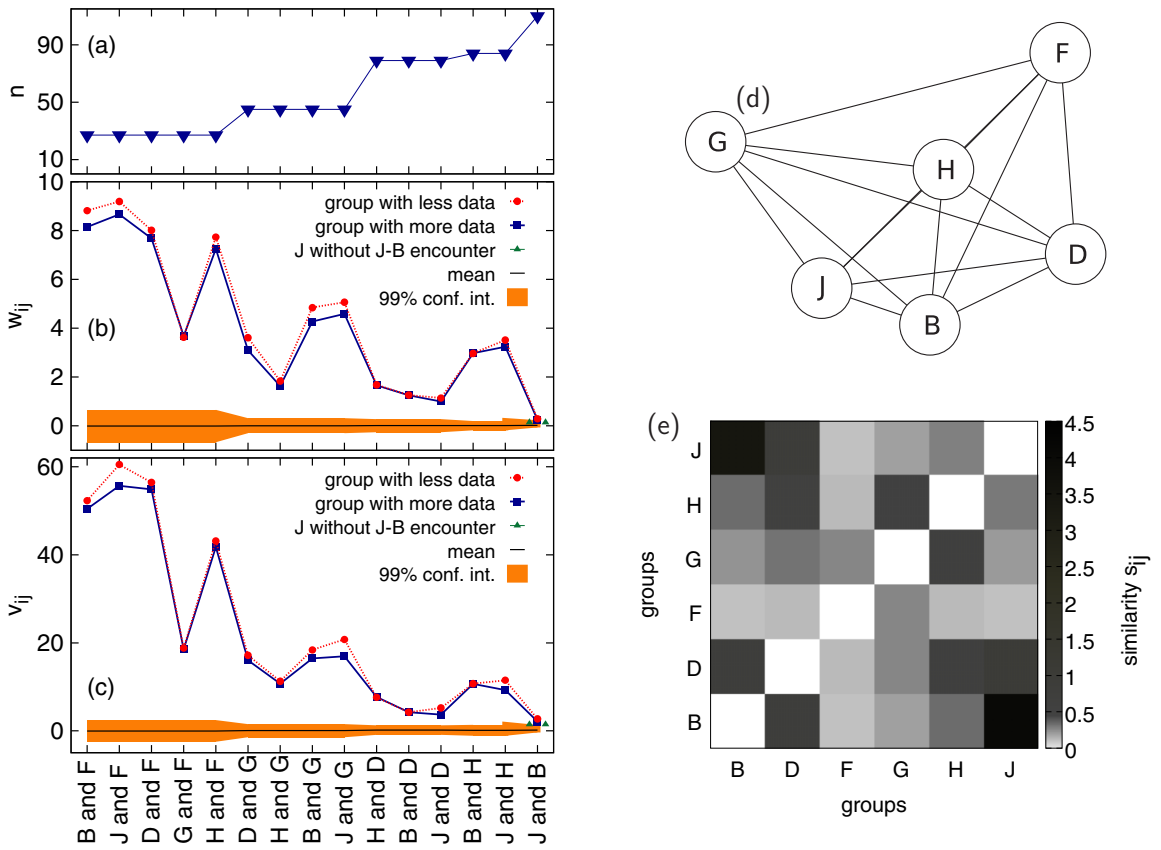

FIG. 6. Quantifying inter- and intragroup differences using BOCCA. (a) Numbers of calls in ensembles for pairwise group comparisons. Intergroup differences are larger than intragroup differences if the coefficients v_{ij} (c) and w_{ij} (b) are positive. Each comparison is done twice, calculating either intragroup difference of the group with more recordings or less recordings. Confidence Intervals [orange (light gray) shaded areas] are estimated by comparing randomly generated, not group-specific, ensembles and assuming that the resulting values follow Student's distribution. To test whether the similarity of groups B and J is due to recordings made during the common encounter of groups B and J, we repeat the analysis excluding all calls of group J that were recorded during the common encounter [plotted as green triangles in (b) and (c)]. The inverse of the weighted coefficients $s_{ij} = 1/w_{ij}$ is expressed as the distance between edges in the network (d) and as entries of the network's adjacency matrix (e).

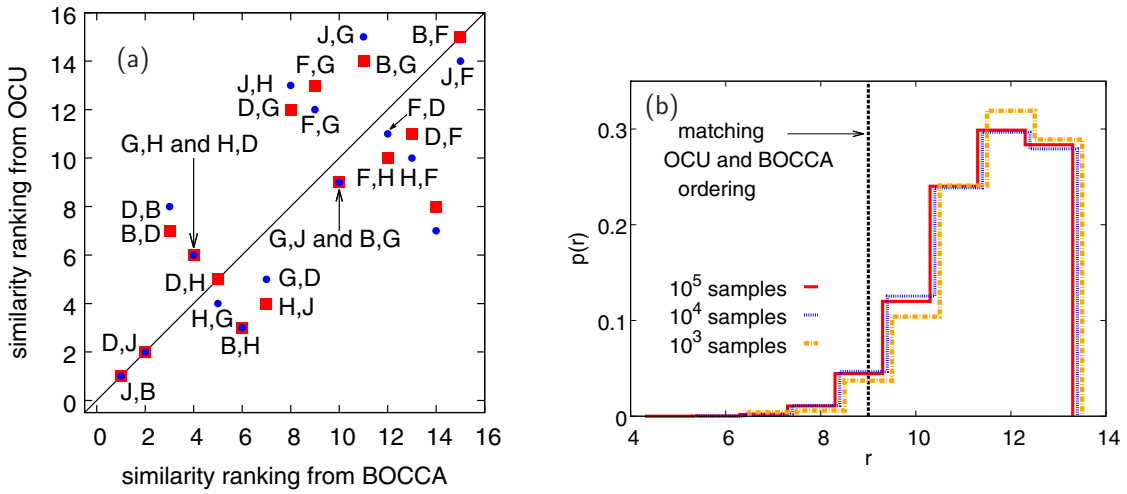

FIG. 7. Comparing the results of OCU and BOCCA. (a) Ranking pairs of groups according to their similarity allows to compare the OCU (y axis) and BOCCA (x axis). Points on the diagonal represent pairs of groups which have the same similarity ranking according to both approaches. Two results (red squares and blue points) are obtained due to the asymmetry of BOCCA with respect to the group that is chosen to estimate the intragroup similarity. (b) The minimal number of swaps needed to transform the ranking according to the OCU approach into the BOCCA ranking is 9. We compare this to r, the number of swaps needed for sorting random vectors of 15 different integers. Estimating the distributions $p(r)$ using either 10^3, 10^4, and 10^5 samples of random vectors, we find that the minimal number of swaps ($r_{\text{BOCCA-OCU}} = 9$) is lower than than 94% of all calculated values of r.

with care. Therefore, we additionally quantify inter- and intragroup differences by computing the newly introduced group difference indices v_{ij} and w_{ij} (see Fig. 6). As shown in Fig. 6, v_{ij} and w_{ij} are larger than zero, indicating that intergroup differences are larger than intragroup differences for all pairwise comparisons. Each coefficient v_{ij} and w_{ij} can be evaluated in two different ways: with respect to the intragroup difference of group i and with respect to the intragroup difference of group j. Note that the number of calls in the respective ensembles $n = \lfloor \frac{1}{2}\min(N_i, N_j) \rfloor$ is the same for inter- and intragroup comparisons.

For all but one of the 15 group comparisons (between group B and J), the values of v_{ij} and w_{ij} are also clearly larger than the confidence bounds estimated using random ensembles of sizes n. Since group B has been only observed while traveling and milling with group J (see Table I), we test whether the measured similarity in vocalization could be influenced by the fact that the recordings are made at the same location. Therefore, we repeat the comparison of group B and group J but exclude all recordings of group J that are made during the common encounter of group B and J (July 3rd, 2010). These additional results are shown as two single points (green triangles) in Fig. 6 and they are in the same order of magnitude as the previous comparison of groups J and B. Consequently, we conclude that features of the vocalizations of groups B and J are very similar, as far as we can estimate on the basis of cepstral distributions.

Using the inverse of the weighted coefficients $s_{ij} = 1/w_{ij}$ as a measure for the similarity in vocalization, we visualize the results of all comparisons in terms of a network with s_{ij} being the adjacency matrix (see Fig. 6). The distance between edges in this network corresponds to the values of s_{ij}, e.g., a high similarity in vocalizations of groups B and J is represented by a small distance between edges B and J.

E. Comparing results from BOCCA and the conventional OCU

Comparing the adjacency matrices obtained through OCU [Fig. 4(c)] and BOCCA [Fig. 6(e)], one can see that the structures of both matrices are very similar. To quantify this, we rank all pairs of groups by their similarities as estimated by OCU and BOCCA [see Fig. 7(a)], i.e., starting with the most similar groups (J and B) on rank 1 and finishing with the least similar groups (B and F) on rank 15. The most similar groups (J and B and then D and J) are ranked identically by both methods, independently of the group that is used as a reference for intragroup similarity (blue circle and red square). Moreover, the similarity between groups D and H as well as between B and F are ranked equally by OCU and BOCCA. All other group comparisons are relatively close to the diagonal, which indicates a similar ranking by both methods. We also estimated the possibility of obtaining two rankings which are so similar by chance. This is done by calculating the numbers of swaps r needed to sort random vectors of 15 integers into a specific ranking (the BOCCA ranking). We then estimate the distributions $p(r)$ from 10^3, 10^4, and 10^5 samples of random vectors [see Fig. 7(b)]. One can see that minimal number of swaps needed to transform the ranking of BOCCA to the OCU ranking $r_{\text{BOCCA-OCU}} = 9$ is smaller than the maxima of the distributions $p(r)$. Integrating the distributions of $p(r)$, we find that 94% of all randomly created rankings need more swaps than $r_{\text{BOCCA-OCU}}$ to be transformed into each other. Consequently, BOCCA produces results that are comparable to the observer-based OCU of 4582 calls, although only

1056 calls of high sound quality are included in the BOCCA analysis.

V. CONCLUSIONS

Many approaches to analyzing whale vocalizations focus on the automated categorization and classification of single sounds, such as types of calls and whistles [12–15,17,28]. Group-specific usage of vocalizations is then discussed by comparing the repertoire (which sounds are used) and whether there are variations of sounds. In the first part of this contribution we follow this well-established approach by conducting an observer-based categorization and classification of all recorded sounds.

To study group-dependent differences in vocalization, we propose and test a new automated method, the BOCCA. The main idea of this approach is that we omit separating and sorting vocalizations into call types and instead compare ensembles of vocalizations produced by each group. Investigating ensembles of calls rather than identifying individual call types is conceptually similar to the bag-of-words model [21] used in text analysis. In the original bag-of-words model a text is represented as the *bag* (multiset) of words disregarding grammar and even word order but keeping multiplicity. We investigate group-specific vocalizations by comparing ensembles i.e., *bags* of calls, that contain calls of a specific group of whales. Comparing the statistical properties of all features computed for each ensemble circumvents the necessity to establish subjective vocal categories or select specific acoustic features. Conceptually similar, using histograms of sound features, has been suggested to attribute bird songs to bird species [29]. Note that the way the ensembles of calls are constructed (choosing only high-quality sounds and additionally applying a random sampling to data from several recording sessions per group) implies that calls within an ensemble most likely originate from different behavioral contexts and that the temporal correlation of calls is destroyed due to the random sampling. Additionally, we did not select specific (hand-crafted) features but used all information contained in the cepstral coefficients [20,29] of sounds. Furthermore, we can even neglect the temporal ordering of these features and each group is well represented by their statistical distributions estimated for each ensemble of sounds. We then quantified differences in vocalization among six groups of pilot whales by computing differences in distribution. To reason whether the calculated differences in distribution were relevant, we introduced two types of coefficients that summarize the relation between inter- and intragroup differences.

Intergroup differences were significantly larger than intragroup differences for all but 1 of 15 intergroup comparisons. However, we also noticed that it is very relevant to estimate confidence intervals through randomly constructed ensembles of a given ensemble size in order to take into account potential finite-sample effects. Interestingly, groups B and J, the two groups with no significant difference in vocalizations, have also been observed traveling and milling together. One possible explanation for their similarity in vocalization is that they are related or that they are subgroups of a bigger group. The common encounter of groups B and J allowed us also to estimate the effect that a similar acoustic environment could have on the similarity of two groups: Even if calls recorded from group J during the common encounter are excluded from the analysis, we still find the same results when comparing ensembles of calls from groups B and J. Since the calls of group J used for this later comparison were recorded on a different day at a different location, we can conclude that the effect of the different acoustic environments on the computed similarity of vocalizations is rather negligible.

Both observer-based classification of calls and the bag-of-calls model yield similar results concerning group specificity. Ranking pairs of groups according to their similarity, BOCCA mostly reproduced the ordering of OCU approach which relies on observer-based classification. This is surprising, since BOCCA used less than 25% of the number of calls which were considered in the OCU analysis. Consequently, it is possible to distinguish groups of pilot whales automatically by simply comparing ensembles of calls without referring to individual properties of single calls. Knowing that quantifiable group-specific vocalization of long-finned pilot whales exist, future work might focus on testing whether clustering approaches (as successfully used in other scientific context, see, e.g., Refs. [30–32]) can confirm these findings or reveal more insight. In total, we consider the bag-of-calls-and-coefficients approach to be a valid method for specifying difference and concordance in acoustic communication in the absence of exact knowledge about signalers, as it is common observing marine mammals under natural conditions or analyzing data generated through automated acoustic monitoring.

ACKNOWLEDGMENTS

We thank Denny Fliegner, Theo Geisel, Jan Nagler, and Julia Fischer for support during project initiation. We also thank Fredrik Broms, Lotta Borg, Kerstin Haller, and Madita Zetzsche for field assistance and photoidentification. Special thanks also to Patrick Kramer for creating a literature database of many relevant articles. This study was supported by Ocean Sounds, the World Wildlife Fund Sweden, and the Max Planck Society (Germany).

APPENDIX: DETAILS ON ENCOUNTERS, RECORDINGS, AND BEHAVIOR

Table I presents details on all encounters, sound-recordings, and behavioral observations.

TABLE I. Group sizes are estimated by visual observation on site, whereas the number of photo ID's refer to animals identified *a posteriori* from pictures taken. During each encounter several recordings were made (number of recording sessions). Summing the duration of these recording sessions yields the total duration of recordings presented here. The dates in the second column of the table are given in the format 'day/month/year'.

Group	Date and location [N/E]	Estimated group size	Photo ID's	Observation time	Sound recordings [h:min]	Number of recording sessions	Behavior
B	03/07/2010 68°04.870′/14°25.130′	45	43	05:33	03:17	19	Slow traveling, socializing
D	10/08/2009 68°06.532′/14°32.771′	100	60	04:42	01:07	7	Milling, slow traveling, socializing
	11/08/2009 68°10.997′/15°29.205′	100	60	04:11	02:25	15	Milling, slow traveling, socializing, foraging, resting
F	28/06/2007 68°04.517′/14°49.436′	20	9	00:50	00:23	3	Milling, boat friendly
G	14/07/2008 68°07.612′/14°40.562′	7	4	02:10	7	00:49	Milling, socializing, boat friendly
H	22/05/2009 68°08.578′/14°31.581′	50	32	04:00	02:45	19	Milling, socializing, resting
	24/05/2009 68°01.636′/14°38.966′	50	22	02:00	01:20	9	Milling, resting
J	13/07/2009 68°01.053′/14°23.519′	60	17	05:48	03:10	19	Milling, resting socializing, boat friendly
	08/06/2010 68°08.540′/15°09.330′	60	19	01:40	01:06	9	First fast traveling -avoided boat, later calmed down
	03/07/2010 68°04.870′/14°25.130′	n/a	4	02:00	01:08	3	slow traveling, Traveling and milling with group B

[1] O. Boebel, H. Klinck, L. Kindermann, and S. E. D. E. Naggar, Bioacoustics **17**, 18 (2008).

[2] National ecological observatory network, Boulder, CO (2015) [http://data.neoninc.org/].

[3] Scaled Acoustic BIODiversity platform [http://sabiod.univ-tln.fr/].

[4] S. Van Parijs, C. Clark, R. Sousa-Lima, S. Parks, S. Rankin, D. Risch, and I. Van Opzeeland, Mar. Ecol. Progr. Ser. **395**, 21 (2009).

[5] H. Vester and K. Hammerschmidt, Vocal repertoire of long-finned pilot whales in northern norway (unpublished).

[6] J. K. B. Ford and A. B. Morton, Can. J. Zool. **69**, 1454 (1991).

[7] V. B. Deecke, L. G. Barrett-Lennard, P. Spong, and J. K. B. Ford, Naturwissenschaften **97**, 513 (2010).

[8] L. Rendell, S. L. Mesnick, M. L. Dalebout, J. Burtenshaw, and H. Whitehead, Behav. Genet. **42**, 332 (2012).

[9] L. S. Weilgart and H. Whitehead, Behav. Ecol. Sociobiol. **26**, 399 (1990).

[10] L. Nemiroff and H. Whitehead, Bioacoustics **19**, 67 (2009).

[11] H. Yurk, Vocal culture and social stability in resident killer whales (orcinus orca), Ph.D. thesis, The University of British Columbia, 2005.

[12] V. B. Deecke, L. G. Barrett-Lennard, P. Spong, and J. K. B. Ford, *The Structure of Stereotyped Calls Reflects Kinship and Social Affliation in Resident Killer Whales (Orcinus orca)* (Springer, Berlin Heidelberg, 2010), Vol. 97, pp. 513–518.

[13] V. B. Deecke, J. K. B. Ford, and P. Spong, J. Acoust. Soc. Am. **105**, 2499 (1999).

[14] V. B. Deecke and V. M. Janik, J. Acoust. Soc. Am. **119**, 645 (2006).

[15] J. C. Brown and P. J. O. Miller, J. Acoust. Soc. Am. **122**, 1201 (2007).

[16] J. C. Brown and P. Smaragdis, Brown and Smaragdis: JASA Express Lett., EL221, 2009.
[17] A. B. Kaufman, S. R. Green, A. R. Seitz, and C. Burgess, Int. J. Compar. Psychol. **25**, 237 (2012).
[18] G. Guo and S. Z. Li, IEEE Trans. Neur. Netw. **14**, 209 (2013).
[19] H. Kim, N. Moreau, and T. Sikora, IEEE Trans. Circuits Syst. Video Technol. **14**, 716 (2004).
[20] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, in *Proceedings of the Symposium on Time Series Analysis*, edited by M. Rosenblatt (Wiley, New York, 1963), pp. 209–243.
[21] Z. Harris, Word **10**, 146 (1954).
[22] L. Schrader and K. Hammerschmidt, Bioacoustics **6**, 307 (1996).
[23] R. Plomp, L. C. W. Pols, and J. P. van de Geer, J. Acoust. Soc. Am. **41**, 707 (1967).
[24] S. S. Stevens, J. Volkman, and E. B. Newman, J. Acoust. Soc. Am. **8**, 185 (1937).
[25] S. Kullback and R. A. Leibler, Ann. Math. Statist. **22**, 79 (1951).
[26] E. Hellinger, J. Crelle **136**, 210 (2009).
[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.93.022138 for examples of pilot whale sounds.
[28] A. D. Shapiro and C. Wang, J. Acoust. Soc. Am. **126**, 451 (2009).
[29] F. Briggs, R. Raich, and X. Z. Fern, in *Ninth IEEE International Conference on Data Mining* (IEEE, 2009), pp. 51–60.
[30] A. K. Charakopoulos, T. E. Karakasidis, P. N. Papanicolaou, and A. Liakopoulos, Phys. Rev. E **89**, 032913 (2014).
[31] A. Î. Charakopoulos, T. E. Karakasidis, P. N. Papanicolaou, and A. Liakopoulos, Chaos **24**, 024408 (2014).
[32] P. Wadewitz, K. Hammerschmidt, D. Battaglia, A. Witt, F. Wolf, and J. Fischer, PLoS ONE **10**, e0125785 (2015).