Marine Mammal Science

MARINE MAMMAL SCIENCE, 32(3): 911–930 (July 2016)

DOI: 10.1111/mms.12303

Automated identification and clustering of subunits within delphinid vocalizations

Kaitlin E. Frasier,[1] Scripps Institution of Oceanography, 8622 Kennel Way, La Jolla, California 92037, U.S.A.; **E. Elizabeth Henderson,** U.S. Navy Marine Mammal Program, SSC Pacific, Code 71510, 53560 Hull Street, San Diego, California 92152, U. S.A.; **Hannah R. Bassett,** Scripps Institution of Oceanography, 8622 Kennel Way, La Jolla, California 92037, U.S.A. and School of Environmental Affairs, College of the Environment, University of Washington, 3707 Brooklyn Avenue NE, Seattle, Washington, U.S.A.; **Marie A. Roch,** Scripps Institution of Oceanography, 8622 Kennel Way, La Jolla, California 92037, U.S.A. and Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, California 92115, U.S.A.

Abstract

Tonal vocalizations or whistles produced by many species of delphinids range from simple tones to complex frequency contours. Whistle structure varies in duration, frequency, and composition between delphinid species, as well as between populations and individuals. Categorization of whistles may be improved by decomposition of complex calls into simpler subunits, much like the use of phonemes in classification of human speech. We identify a potential whistle decomposition scheme and normalization process to facilitate comparison of whistle subunits derived from tonal vocalizations of bottlenose dolphins (*Tursiops truncatus*), spinner dolphins (*Stenella longirostris*), and short-beaked common dolphins (*Delphinus delphis*). Network analysis is then used to compare subunits within the vocal corpus of each species. By processing whistles through a series of steps including segmentation, normalization, and dynamic time warping, we are able to automatically cluster selected subunits by shape, regardless of differences in absolute frequency or moderate differences in duration. Using the clustered subunits, we demonstrate a preliminary species classification scheme based on rates of subunit occurrence in vocal repertoires. This provides a potential mechanism for comparing the structure of complex vocalizations within and between species.

Key words: acoustic, classification, communication, delphinid, dynamic time warping, network analysis, unsupervised learning, vocalization.

Many delphinid species produce complex, variable tonal calls or whistles, thought to have a social function (Janik 2009). Whistles can consist of numerous rises and falls with varying rates of change, inflection points, and even nonlinearities, *e.g.*, Janik *et al.* (1994), Azzolin *et al.* (2014) (Fig. 1). Whistle comparison and classification

[1]Corresponding author (e-mail: kfrasier@ucsd.edu).

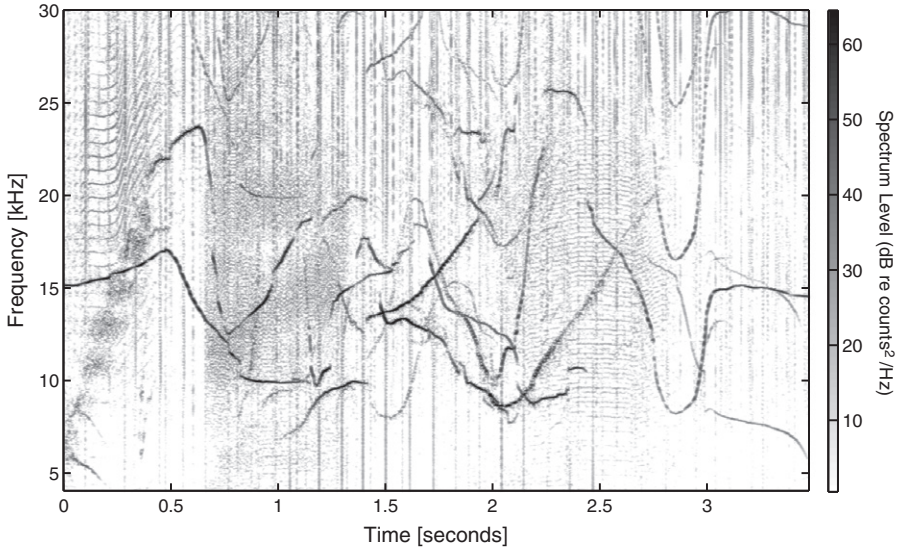

Figure 1. Spectrogram representation of whistles (dark, sweeping lines), echolocation clicks (vertical lines), and burst pulses (rapid echolocation clicks that produce spectral banding) produced by a pod of short-beaked common dolphins.

efforts are often hampered by this complexity. For instance, a pair of tonals may have similar inflection points or structure, while other aspects including frequency or duration may differ. There is a need in odontocete vocalization research for a tonal description framework capable of summarizing whistle variability in a way that facilitates shape and structural comparison for use in applications including inference of species or individual identity, or animal behavior.

Traditional analyses of delphinid whistles summarize each call by reducing it to a series of measured parameters. The earliest forms of this approach used maximum and minimum frequency, duration, and start and end frequency to characterize delphinid vocal repertoires (Steiner 1981). This method does not capture information regarding whistle frequency modulation or shape, despite the fact than human analysts (*e.g.*, Simonis *et al*. 2012) often consider these to be important features. Techniques subsequent to Steiner attempt to take whistle shape into account, such as Buck and Tyack's (1993) use of dynamic time warping (DTW, Rabiner and Juang 1993), a nonlinear technique to time-align similar vocalizations. McCowan (1995) also captured shape changes over time, by computing correlation coefficients along whistle frequency contours. These techniques improve the ability to compare tonals on the basis of shape, but do not improve the classification of more complex whistles.

Recently, techniques have been developed to simplify whistles further while preserving shape information using space transformations, such as rescaling, normalization, and other linear transformations (*e.g.*, Kershenbaum *et al*. 2013). Deecke and Janik (2006) transformed whistles into log space, because the ability of delphinids to discriminate between frequencies has been shown to decrease with rising pitch (Thompson and Herman 1975, Supin and Popov 2000). Deecke and Janik (2006) also showed that unsupervised learning methods could be used to group similarly shaped whistles within a relatively small data set ($n = 104$). However, when applied

to larger whistle sets, the number of categories identified was too great for effective clustering.[2]

An alternative to analyzing entire whistles is to split these complex signals into a series of simpler subunits. A rough analogy from the field of human speech classification is the practice of breaking words into a series of phonemes to facilitate recognition (Watrous and Shastri 1987). Like phonemes, whistle subunits are simple and more easily characterized than entire whistles, and a relatively small number of subunits can be combined in different sequences to create a wide variety of signals. Recent work on killer whale vocalizations identified common subunits across whistle-like calls and sought to characterize each call as a series of these subunits (Shapiro *et al*. 2011). There is a growing body of evidence that numerous species compose complex acoustic sequences from simpler subunits (see Kershenbaum *et al.* 2014 for a review).

The representation of whistles as series of subunits presents the possibility of applying methods used in human speech recognition to acoustic marine mammal classification. Examples of how sequences of speech subunits have been used for identifying characteristics related to humans includes language identification based on similarity to language-specific phoneme models (*e.g.*, Lamel and Gauvain 1994) and the identification of individuals based on prosodic features such as how often people use specific phrases (Doddington 2001). Similarly, a basic form of a subunit-based whistle classifier might try to distinguish species based on the relative occurrence of certain types of subunits in a vocal repertoire. A more complex classifier might build on this strategy by incorporating information about the order in which subunits occur, and other additional features of the subunits.

The goals of this work are (1) to identify a set of convenient and easily identifiable subunits within recorded delphinid whistles, (2) to develop an approach for clustering these subunits according to type, and (3) to demonstrate the potential of subunits as a starting point for whistle-based acoustic classification of delphinids. This work makes use of an annotated corpus of whistles developed for whistle detection algorithms. The whistles from three delphinid species are examined: Short-beaked common dolphin (*Delphinus delphis)*, spinner dolphin (*Stenella longirostris*), and bottlenose dolphin (*Tursiops truncatus*). Having identified whistle subunits within the corpus, we ask whether the automated clustering method produces reliable results by comparing the output to randomized clusters of species-specific subunits. To illustrate the potential of subunit-based analyses, we report preliminary species classification results based solely on the relative occurrence of subunit types. Our analyses suggest that this approach can be used to characterize and compare whistles automatically across large data sets on the basis of shape.

Methods

Data Collection and Whistle Extraction

Tonal contours were extracted from acoustic recordings compiled as part of the conference data set associated with the 5th International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals Using

[2]Personal communication from Vincent M. Janik, Scottish Oceans Institute, East Sands, University of St. Andrews, St. Andrews, Fife KY16 8LB, U.K., 13 June 2013.

Passive Acoustic Monitoring (Roch *et al*. 2011), available on MobySound (Mellinger and Clark 2006). Contours from short-beaked common dolphin, spinner dolphin and bottlenose dolphin were included in this work because of the availability and quality of labeled recordings for these species. Signals were detected automatically using the contour detector *Silbido* (Roch *et al.* 2011). An analyst manually corrected false detections (*e.g.*, echo sounder pings), incorrectly linked contours, and artificial tonal segmentation. This process generated a corpus (vocal data set) for each species consisting of paired time and frequency vectors describing the detected tonals. Signals with a maximum frequency above 30 kHz or a minimum frequency above 20 kHz were excluded from the corpus, to limit the inclusion of harmonics. Minimum signal duration required for consideration was 40 ms, below which signals lacked useful information (see *Subunit Identification and Normalization* below). All analyses were conducted using the Matlab programming environment (MATLAB version 8.0.0.783 2012, The Mathworks, Inc., Natick, MA).

Subunit Identification and Normalization

Each extracted tonal was smoothed with a Hermite cubic interpolating spline evaluated at 2 ms intervals (Matlab pchip, Fritsch and Carlson 1980), equivalent to a whistle contour sampling rate of 500 Hz. This removed any gaps associated with the detection process and smoothed small variations out of the contour, leaving a generalized shape.

Segmentation of the tonals into subunits requires the choice of a logical delimiter; in this case frequency maxima and minima (extrema) were selected (Fig. 2.; see Fig. S1, S2 for plots of subunits per whistle for each species). This choice is supported by experimental evidence that bottlenose dolphins have been shown to strongly discriminate between ascending and descending frequency contours (Ralston and Herman 1995). Inflection points along the contours were also considered as possible segment delimiters, but the ability to identify the exact location of inflection points was less robust.

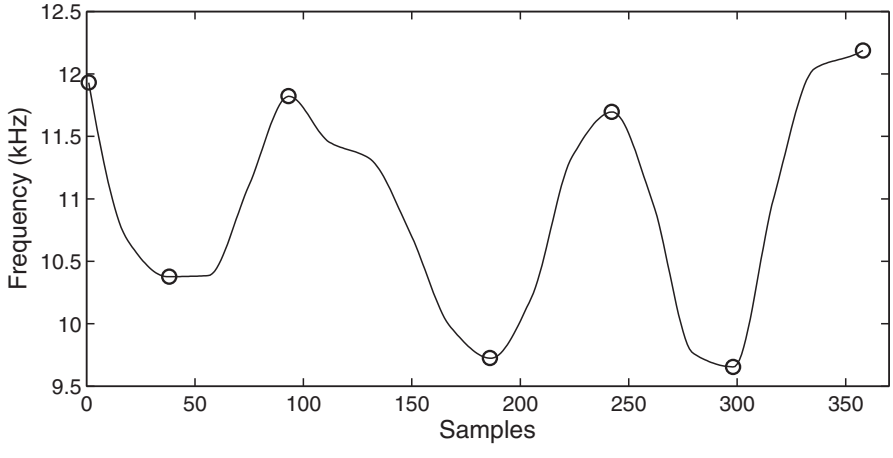

Figure 2. Example of a common dolphin whistle (black line), interpolated and resampled at 500 Hz, with subunit boundaries indicated by empty circles. Subunit boundaries were located at frequency extrema, and at whistle start and end points.

The set of extrema positions $\vec{E}$ of the interpolated tonal were identified by locating the zeros of the first derivative of the interpolated tonal with respect to time (t).

$$\vec{E} = \left(t : \frac{dT_I}{dt} = 0 \right) \tag{1}$$

A vector of boundaries $\vec{B}$ between subunits S was generated by adding whistle start and end points to the set of extrema:

$$\vec{B} = [1, \vec{E}, \text{length}(T_I)] \tag{2}$$

Subunit $\vec{S}_n$ within tonal T was defined as the whistle frequency contour between time indices $B(n)$ and $B(n + 1)$. For simplicity, we will frequently omit the subscript n and will write $\vec{S}$ with the implication that it refers to a specific subunit.

After segmentation, subunits shorter than 20 ms were excluded from further analyses, because they contained too few data points (<10) to be informative. The first derivative of each remaining subunit S was computed, and a feature vector $\vec{F}_S$ for each subunit consisting of frequency and first derivative was stored for further analysis.

$$\vec{F}_S = (\vec{S}, d\vec{S}) \tag{3}$$

The sampling rate was constant for all tonals after interpolation, therefore temporal information was not retained.

Frequency transformations were applied to subunits to improve shape-based comparability by normalizing bandwidth. Each subunit frequency vector was natural log-transformed, and then normalized using a z-score transformation $\vec{Z}_{f_S}$ (Kreyszig 1979):

$$\vec{f}_S = \ln(\vec{F}_S) \tag{4}$$

$$\vec{Z}_{f_S} = \frac{\vec{f}_S - \mu_{f_S}}{\sigma_{f_S}} \tag{5}$$

The terms μ_{f_S} and σ_{f_S} represent the mean and standard deviation of $\vec{f}_S$, respectively.

First derivatives of the subunit frequency vectors were also normalized using a modified z-score transformation ($\vec{Z}_{d_S}$)

$$\vec{Z}_{d_S} = \frac{d_S}{\sigma_S} \tag{6}$$

where d_S is the first derivative of f_S, and is the standard deviation of d_S. The mean subtraction was omitted from Equation 6 to preserve the sign of the derivative.

Warping and Distance Calculation

Pairs of subunits, S_i and S_j, were compared using a DTW algorithm (Myers *et al.* 1980), where i and $j = 1,\ldots,n$, and n is the number of subunits in the corpus. DTW

is based on the idea that two subunits may have similar shapes, but different durations. The DTW algorithm attempts to nonlinearly align pairs of feature vectors i and j such that the distortion between them is minimized. Constraints on this DTW algorithm ensure that the two sets of feature vectors are not time-reversed or expanded/compressed in unreasonable ways. A mathematical treatment of this algorithm can be found in Rabiner and Juang (1993, pp. 200–226).

An example alignment is shown in Fig. 3. The lower panel shows a heat map with the cumulative cost of associating any two feature vectors in subunits S_1 and S_2. Areas that are overly expanded or compressed have infinite warp distance costs and can be seen as the area outside the parallelogram. The DTW algorithm efficiently computes

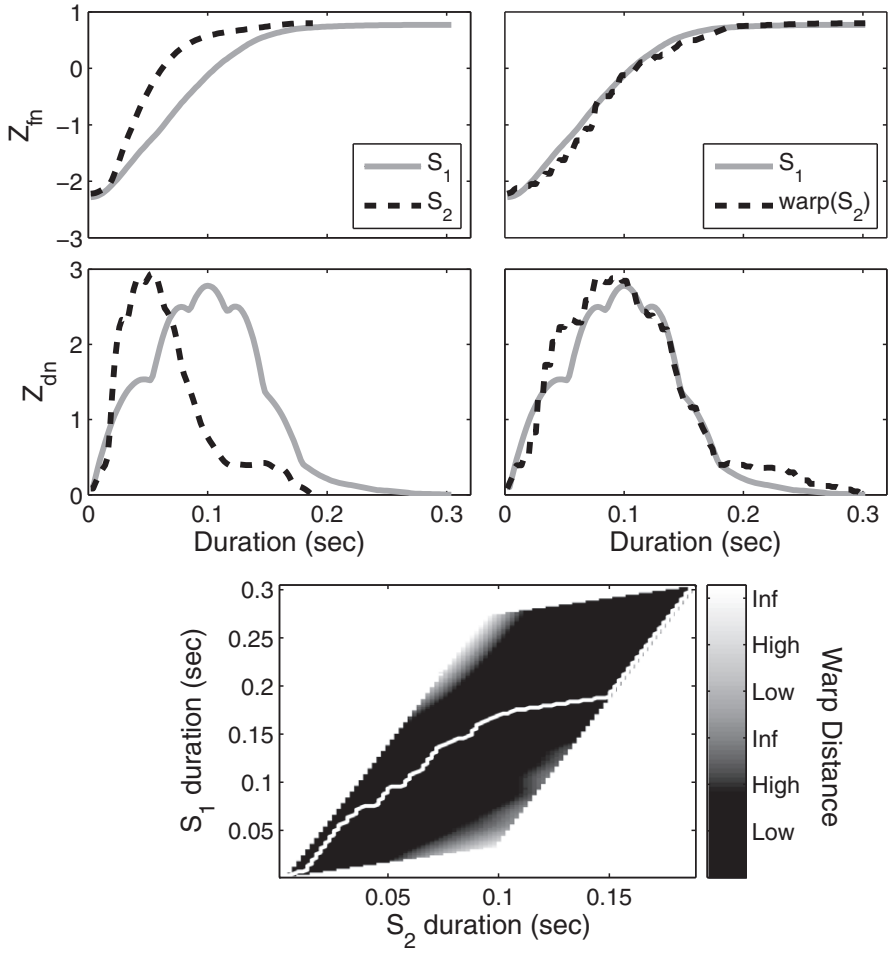

Figure 3. Dynamic time warping of a pair of whistle subunits, S_i and S_j. Top: original (left) and warped (right) z-score of frequency contours. Center: original (left) and warped (right) z-score of first derivatives of frequency. Bottom: cumulative feature vector distortion (grayscale) along the best warp path (white line) between the two components.

this map by examining pairs of feature vectors s_1 and s_2 and then searching for previously computed partial alignments that could lead to the current pair. The search uses local constraints (Type V, Rabiner and Juang 1993, p. 223) to limit the possible candidates such that unreasonable matches are not proposed. The Euclidean distance at the current point is added to the cumulative distance of the best preceding candidate. When the program terminates, there is a cumulative measure of how much the two subunits differed along the best path (white path in lower panel of Fig. 3). This process was repeated for each pair of subunits.

The feature vectors consist of the normalized frequencies and time derivatives $(\vec{Z}_{fs}, \vec{Z}_{ds})$. The use of both frequency and first derivative in the warping scheme required that the warp algorithm optimize the alignment between subunits both in terms of magnitude and slope. While the slope information can be inferred from the frequencies, the DTW algorithm compares sampled points of the frequency contour; without explicit information, the algorithm could not tell the difference between two close points that have the same or opposite trends. The upper panels of Figure 3 show the normalized frequency and frequency derivative for a sample pair of subunits. The left panels show the signals prior to warping, while the right panels show the alignment of the signals according to the optimal warp path.

Warp costs were normalized by warped path length to remove the inherent penalty on long subunits. Subunit pairs in which one member was greater than three times longer than the other were not warped, and were instead assigned an infinite warp cost. The output of the distance calculation process for each species consisted of a set of length-normalized warp distances $D_{S_iS_j}$ between subunit pairs.

Clustering

The clustering process expresses the relationships between subunits as a network in which each node represents a subunit, and weighted edges (linkages) indicate the similarity between subunits. Networks can be constructed for individual species or populations to examine subunit characteristics of the species or group. Alternatively, networks can be constructed from data pooled across multiple species, to examine general trends. In this section, preliminary experiments looked at species-specific subunits, therefore networks were constructed on a per species basis. Later, in a classification experiment (see *Classification* below), we will construct a network using data from multiple species.

Within a network, a high linkage weight indicates strong similarity between a pair of subunits. Since the warp distortion computed *via* DTW is a measure of dissimilarity, we used a sigmoid function to map the distortion to a unitless linkage weight over the interval [0, 1] indicating a poor to perfect correspondence after warping:

$$W_{S_iS_j} = \exp(-D_{S_iS_j}) \tag{7}$$

Linkage pruning threshold (p) choices from 0.1 to 0.9 were tested, such that linkages with weights less than p were removed from the network, in order to examine the effect of network pruning (Zhou *et al*. 2012) on clustering results.

A modularity-based network analysis algorithm (Blondel *et al.* 2008) implemented in the network visualization package *Gephi Toolkit* (Bastian *et al.* 2009), was used to partition the network into clusters of closely related nodes. The modularity calculation is based on the idea that a good partitioning of a network is one in which edges linking nodes within the same cluster are strong, while edges linking outward to

nodes in other clusters are weak (Newman 2006). Accordingly, the modularity Q of a network partition is a value between −1 and 1 that represents the strength or weights of the edges within clusters compared to the weights of the edges between clusters. Mathematically, modularity (Newman 2004) is defined for a network of n nodes as

$$Q = \frac{1}{u}\sum_{i,j}^{n}\left[W_{ij} - \frac{1}{m}\frac{k_i k_j}{u}\right]\delta(c_i, c_j) \tag{8}$$

where $k_i = \sum_j W_{ij}$ is the sum of the weights of the edges from all nodes j attached to node i, $k_j = \sum_i W_{ij}$ is the sum of the weights of the edges from all nodes i attached to node j, and $u = \sum_{ij} W_{ij}$ is the sum of all edge weights in the network. In the delta function $\delta(c_i,c_j)$, c_i and c_j represent the cluster to which nodes i and j have been assigned. If c_i and c_j are equal, then $\delta(c_i,c_j) = 1$, otherwise it is zero. A resolution coefficient m, added in later formulations (Lambiotte *et al.* 2008, Mucha *et al.* 2010), defaults to unity but can be adjusted as discussed below.

The best partition of a network is taken to be one that maximizes Q. In the implementation used here (Louvain method, Blondel *et al.* 2008), each node (*i.e.*, each subunit) is initially assigned to its own cluster. Clusters are iteratively merged if doing so increases Q. Iterations cease when no further merges can increase Q. The resolution coefficient m can be used to adjust clustering resolution and influence the number of modules identified (Lambiotte *et al.* 2008, Mucha *et al.* 2010). A value of m greater than one increases the positive effect of merging clusters on Q, thus favoring the identification of fewer, larger clusters. A value of m less than one decreases the positive effect of cluster merging on Q, thereby favoring the formation of more numerous, smaller clusters. Network partitions were generated using a range of resolution coefficients between 0.25 and 2 to illustrate the effects of this parameter choice. Clusters containing fewer than ten nodes were ignored.

Cluster Consistency

Clustering algorithms that provide radically different clusters for very similar data are not particularly useful. Consequently, we examine methods that subsample the subunits repeatedly, cluster them and examine how often specific pairs of subunits appearing in two subsamples are clustered together or appear in different clusters. When subunits tend to be kept in the same or different clusters across many different trials, the clustering is said to be consistent (Strehl and Ghosh 2003). High consistency indicates that the clusters retain similar structure across random samples despite variation in the subunits that were clustered. A detailed explanation of this process is given below.

Assume that two random subsamples of subunits, S_a and S_b, have been partitioned into sets of clusters, P_a and P_b, respectively (Fig. 4). Partition P_a consists of a set of k_a clusters, where n_i^a denotes the number of subunits in cluster i for $i = 1,\ldots,k_a$. Similarly, P_b consists of a set of k_b clusters, where n_j^b denotes the number of subunits in cluster j for $1,\ldots,k_b$. The number of subunits common to cluster i in P_a and j in P_b is denoted as n_{ij}^{ab}.

Cluster consistency was quantified using normalized mutual information (NMI, Strehl and Ghosh 2003), an information theory metric that is indicative of clustering sensitivity to changes in the data set. Given partitions P_a and P_b, the NMI was computed as

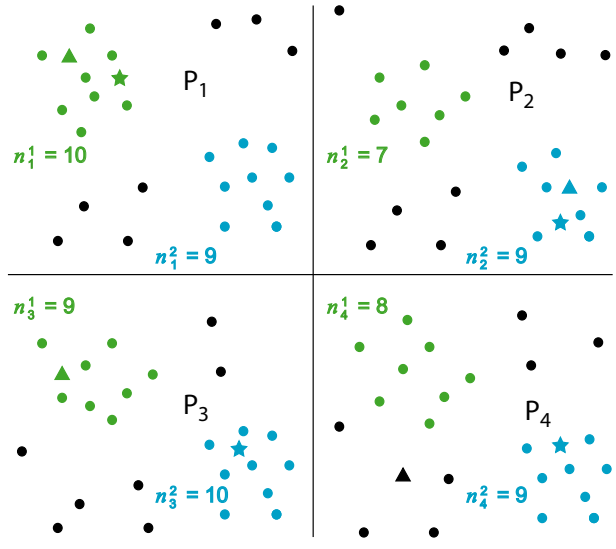

Figure 4. Four possible partitions (P1–4) of randomly selected whistle subunits. In each partition there are blue and green clusters of subunits along with a set of components that were not clustered (black). Counts in the form are given for blue and green clusters. Two specific subunits (denoted with a star and triangle) appear in each set of randomly selected data. The fact that these two subunits are clustered together in both partitions 1 and 2 would contribute positively to the normalized mutual information (NMI) of the two partitions. These same subunits would contribute to lowering the NMI when comparing P1 to P3 or P1 to P4, as the star and triangle are either in different clusters (P3) or at least one subunit was unclustered (triangle in P4), rather than together. Partitions are synthetic for illustrative purposes.

$$NMI(P_a, P_b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}^{ab}\log\left(\frac{n_{ij}^{ab}\cdot n}{n_i^a \cdot n_j^b}\right)}{\sum_{i=1}^{k_a}\log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_b}\log\left(\frac{n_j^b}{n}\right)} \tag{9}$$

where the numerator measures the number of clusters with similar composition between the two partitions, normalized by the total number of subunits in the respective clusters. The denominator represents the entropy of each partitioning (see Strehl and Ghosh 2003 for a full derivation of Eq. 9). NMI has a maximal value of one when for each cluster of subunits in P_a, there is a cluster containing the identical set of subunits in P_b (Fig. 4). NMI has a minimal value of zero when there is no cluster consistency between the partitions. The treatment of n, the number of nodes in a partition, is explained below.

In our implementation, P_a and P_b were constructed for species ω by selecting and clustering 100 different bootstrap samples (randomized with replacement) of 70% of the subunits in species ω's corpus. We then computed the mean and standard deviation of NMI between pairs of these bootstrapped partitions.

While a high NMI indicates that partitions are consistent, it does not guarantee that meaningful structure has been captured. A common method to demonstrate that some type of structure has been learned is to compare the mean NMI to that of partitions that were constructed by random assignment of items to clusters (Fred and

Jain 2005). For comparison with the clustering results, each bootstrap sample was repartitioned such that the number of clusters and their sizes remained the same but subunits were randomly assigned to clusters. The mean NMI of the cluster-based partitions was compared to the mean NMI of the randomized partitions (NMI^R). If the clustering algorithm is consistently learning the structure of the subunits, one would expect the NMI to be higher than that associated with randomly assigned clusters.

Not all whistle subunits are clustered, and thus our partition sets are likely to be a subset of the random sample. This leads to multiple possible interpretations for n. One interpretation for n is the number of whistle subunits that were selected in *both partitions* and clustered:

$$\left| \cap \left(\cup_{a \varepsilon P_a} a, \cup_{b \varepsilon P_b} b \right) \right| \tag{10}$$

where $\cup_{a \varepsilon P_a}$ denotes the set of all subunits in subset a that were included in a cluster of P_a, and $|\text{–}|$ indicates that we are looking for the number of common subunits found in both partitions of the subunits: P_a and P_b.

Alternatively, n could represent the number of whistle subunits that appear in *both subsets*, S_a and S_b, regardless of whether or not they were assigned to clusters:

$$\left| \cap \left[(a : a \in S_a), (b : b \in S_c) \right] \right| \tag{11}$$

We compute NMI for both interpretations, with the former, NMI_P, providing information about the consistency of nodes within the cluster *partition*, and the latter, NMI_S, providing an overall measure of consistency of the random *subset*. It is expected that $NMI_P \geq NMI_S$ as the n used for NMI_S penalizes the NMI for unclustered whistle subunits.

With this methodology in place, multiple trials were run while varying the pruning (p) and resolution (m) clustering parameters. Parameter choices were varied independently across trials. Each trial consisted of 100 random subsets of whistle subunits, with 100 partitions generated. The NMI statistics were calculated from the $\binom{100}{2} = 4{,}950$ possible pairwise combinations of these 100 subsets.

Classification

Clusters can have many applications for classification, most of which are beyond the scope of this paper. To demonstrate that the clusters do contain information that is relevant to classification, a trivial classifier was implemented to demonstrate the use of component type in species classification applications. Rather than clustering subunits from single species, all cross-species subunits are pooled and clustered. A cluster C can be defined as a set of subunits S_{ik}

$$C = (S_{ik}, i = 1, \ldots, n_k, k = 1, \ldots, \omega) \tag{12}$$

where n_k is the number of subunits in C associated with species k and ω is the number of species in the corpus.

For each cluster, a prior probability distribution P that a given subunit is generated by species ω is estimated by:

$$P(k|C) = \frac{n_k}{n_C} \tag{13}$$

where n_C is the total number of subunits in cluster C.

Classification decisions are made on groups of subunits with the assumption that all subunits within the group are generated by a single species. Each subunit is associated with a cluster *via* a nearest neighbor learning rule (Hastie *et al.* 2009, pp. 463–468). Assignment of a group of subunits to species is based on the joint probability of the cluster-specific prior distributions. Details on the methods follow below.

Training—To determine sensitivity of the classification results to training data, a modified bootstrap procedure evaluated 50 executions of a randomized three-fold cross validation experiment (Roch *et al.* 2015). To ensure independence of training and test data, folds were constructed such that all subunits form a single acoustic encounter with a group of animals were placed in a single fold. For each bootstrap experiment, encounters were randomly assigned to each of the three folds in a balanced manner. Three sets of models were created, holding back one fold for testing each time. Within the training folds, a bootstrap sample (randomized with replacement) of 75% of the subunits for each species were selected. The selected training subunits for all species were combined into one set, which was clustered using the previously described algorithm. Resulting clusters could contain subunits from multiple species (Fig. 5). Subunit probabilities were computed for each species according to Equation 13.

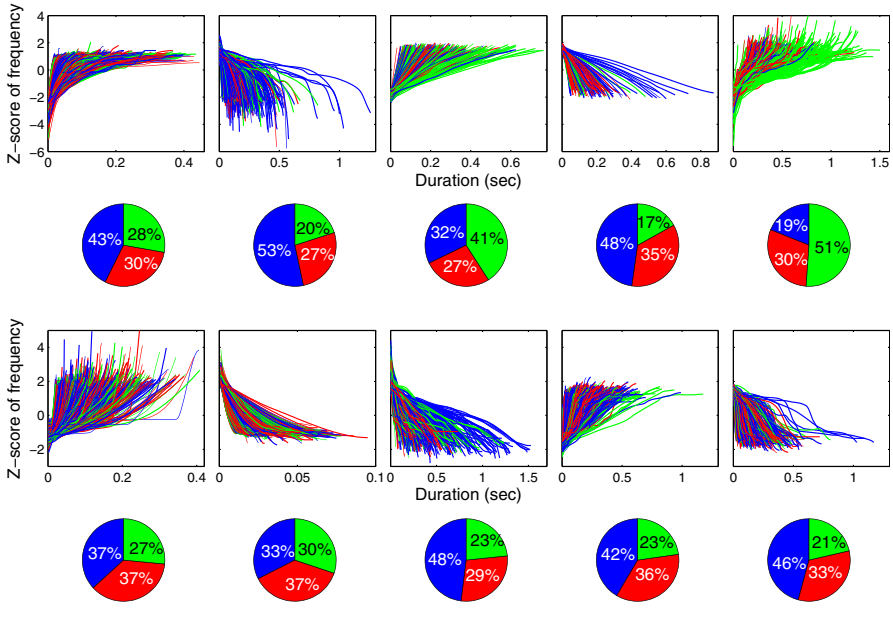

Figure 5. Example training set in which subunits from all training species are clustered as one set. In the square panels, normalized subunits (curved lines) associated with each training cluster are color-coded according to species. Blue: bottlenose dolphin, red: short-beaked common dolphin, green: spinner dolphin. A pie chart below each panel indicates the percentage of subunits in the training cluster associated with each species. These ratios are used to estimate the probability that a test subunit resembling a particular cluster is associated with each training species.

Testing—Classification of a set of subunits consisted of finding the cluster associated with each subunit by nearest neighbor search and then considering the probability that the cluster was produced by a specific species. The joint probabilities of the entire set belonging to a specific species were evaluated and the species hypothesis producing the highest joint probability was selected.

The nearest neighbor search assigned cluster labels to test subunits by selecting the cluster with a minimum mean similarity μ_C measure between a test subunit S_i, and the subunits $S_j (j = 1,\ldots,n_C)$ in training cluster C:

$$\mu_C = \frac{1}{n_C} \sum_{S_j \in C} DTW(S_i, S_j) \tag{14}$$

where DTW denotes the dynamic time warping distance. Subunit pairs with an infinite warp distance were assigned a similarity of zero. Subunit S_i was assigned to the cluster for which μ_C was highest, and was assigned a probability of belonging to each of the training species based on the probabilities associated with that cluster. Classification decisions were made on groups of 100 sequential subunits. A simplifying assumption was made to treat the subunits as if they were independent, permitting easy computation of the joint probability that the group was produced by species ω:

$$\log P(\omega|S) = \sum_{i=1}^{100} \log\{P[\omega|NN(S_i)]\} \tag{15}$$

where $NN(S_i)$ denotes the cluster associated with subunit S_i by nearest neighbor search. Species assignment for the group of subunits was based on the maximum joint probability of Equation 15:

$$\underset{(\omega \in species)}{\arg\max} \log p(\omega|S)$$

Results

Corpus size varied between species according to the amount of recorded data containing detectable vocalizations, with spinner dolphin data having the smallest sample size (Table 1). In all cases, the mean number of components per whistle was biased toward a single component. The number of subunits per encounter varied from 123 to 2,055.

Clustering

Clustering results varied between trials in response to the combined effects of the clustering parameters m and p (Table 2, Fig. 6). Higher pruning values of p were generally correlated with increased NMI for both clustered and randomized partitions. Increased p also decreased the percentage of nodes retained in clusters, which accounts for the moderate effect of pruning threshold on the NMI metric that ignores unclustered subunits (NMI_S) compared to the stronger effect seen on the NMI metric that penalizes unclustered subunits (NMI_P). The effect of counting nonclustered nodes against cluster consistency scores was to reduce NMI_S relative to NMI_P. Both resolution m and pruning p affected the number of clusters per partition; as expected, lower

Table 1. Corpus details and component breakdown by species.

Species	Number of encounters	Number of whistles detected	Number of subunits identified	Number of subunits per whistle mean (SD)
Short-beaked common dolphin	4	1,959	3,573	1.82 (1.08)
Spinner dolphin	4	1,608	2,610	1.62 (0.99)
Bottlenose dolphin	4	2,280	4,090	1.79 (1.08)

Table 2. Comparison of mean ± standard deviation of NMI and cluster summary statistics by species. Two sets of parameter pairs (high m and low p, *vs*. low m and high p) are shown, to illustrate the influence of parameter choice on clustering results.

	Short-beaked common dolphin		Spinner dolphin		Bottlenose dolphin	
	High $m = 1.5$ Low $p = 0.3$	Low $m = 0.5$ High $p = 0.8$	High $m = 1.5$ Low $p = 0.3$	Low $m = 0.5$ High $p = 0.8$	High $m = 1.5$ Low $p = 0.3$	Low $m = 0.5$ High $p = 0.8$
NMI_P	0.833 ± 0.070	0.876 ± 0.024	0.835 ± 0.056	0.865 ± 0.027	0.785 ± 0.061	0.906 ± 0.029
NMI_S	0.471 ± 0.038	0.527 ± 0.019	0.482 ± 0.034	0.523 ± 0.026	0.452 ± 0.036	0.537 ± 0.039
NMI_P^R	0.004 ± 0.002	0.040 ± 0.006	0.008 ± 0.003	0.072 ± 0.010	0.014 ± 0.006	0.118 ± 0.024
NMI_S^R	0.004 ± 0.002	0.037 ± 0.005	0.008 ± 0.003	0.063 ± 0.009	0.014 ± 0.006	0.094 ± 0.020
% of nodes clustered	96.61 ± 0.32	76.83 ± 0.79	94.42 ± 0.45	67.73 ± 1.11	90.74 ± 1.14	57.06 ± 2.10
Number of clusters	5.35 ± 0.58	15.08 ± 1.52	5.77 ± 0.65	14.07 ± 1.28	5.23 ± 0.80	9.71 ± 1.26

values of m favored the formation of more clusters, as did higher values of p. In general, NMI was maximized at high p and low m. In all cases, clustered partition NMI scores were over an order of magnitude higher than those for randomly generated partitions (NMI^R), indicating that clusters were not random. As expected, the contrasting NMI metrics from randomized clusters (NMI_S^R, NMI_P^R) have values near zero indicating that there is very little consistency between the clusterings (Table 2).

Species Classification

Correct species classification rates varied as a function of the network pruning parameter p. Low values of p increased correct classification rates (Fig. 7). Classifications were not significantly affected by the choice of resolution coefficient m. Within the range of parameters tested, the mean correct classification rate was maximized using $p = 0.3$ and $m = 0.5$. Using these parameters, the average error rate was 27% (SD = 17%) across 50 runs of three experimental folds. For comparison, classification

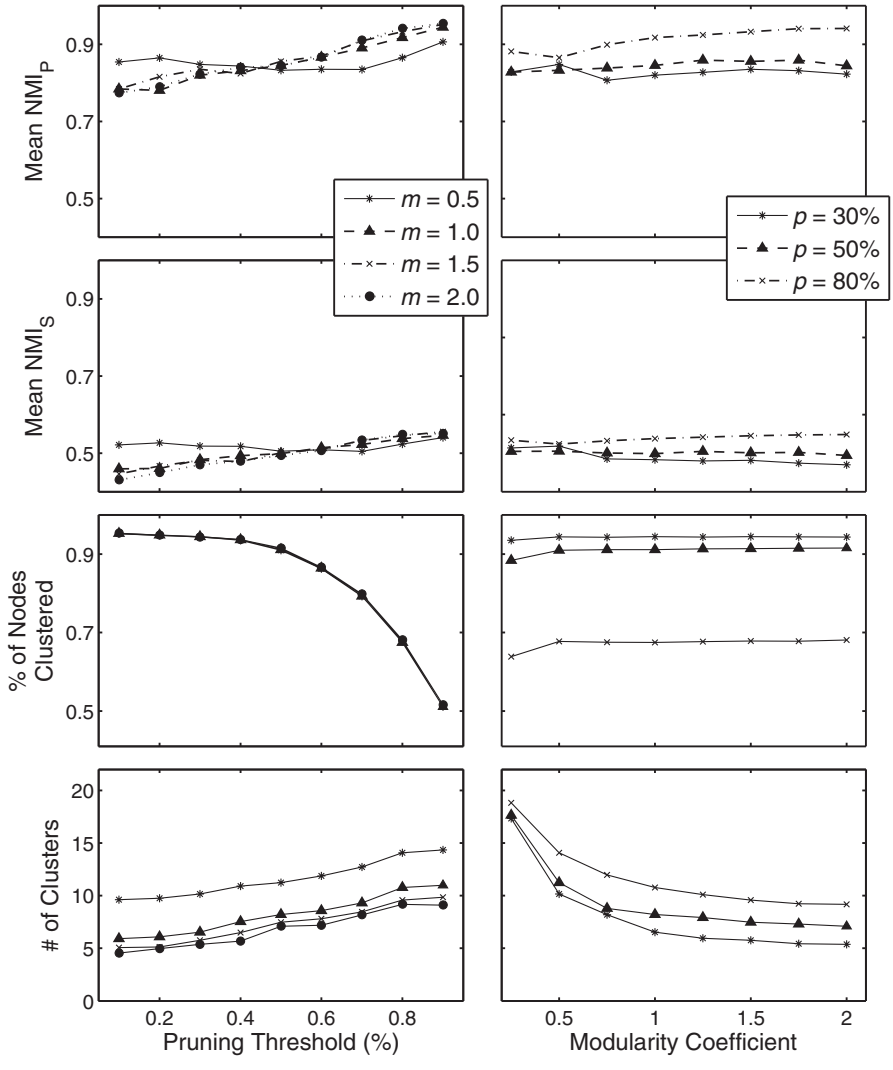

Figure 6. Clustering results for spinner dolphin corpus as a function of pruning threshold (p) and resolution coefficient (m) variation. Each point represents the mean of 100 partitions. Left column: for a constant value of m (see legend), the effects of changes in p on NMI and clustering results are shown. Right column: for a constant value of p, the effects of changes in m on NMI and clustering results are shown.

by random assignment is expected to have an average error rate of 66%. Confusion was highest between common and spinner dolphin subunits (Table 3).

Discussion

Identifying and categorizing whistle subunits provides a framework for summarizing complex tonal calls as a series of elements that can be automatically recognized

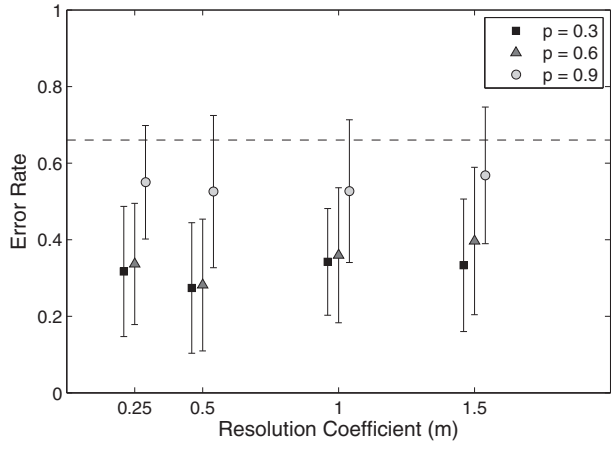

Figure 7. Mean species classification error rates using a simple, subunit shape-based classifier. Mean error rates are computed across 50 randomized trials, each consisting of three experimental folds. Error bars indicate one standard deviation from the mean. The horizontal line indicates the error rate expected by random assignment (66%). Within the range of parameters tested, error rates decrease with lower pruning thresholds (p) but did not vary significantly as a function of resolution coefficient (m).

Table 3. Confusion matrix showing correct species classifications as a percentage of total subunit sets classified, across 50 runs of three randomized folds using clustering parameters m = 0.5 and p = 0.3. Whistle subunits were classified in sequential sets of 100. Percentages are rounded to the nearest integer. Bold font indicates correct classifications.

		Produced by		
		Bottlenose dolphin	Common dolphin	Spinner dolphin
Classified as	Bottlenose dolphin	**80%**	21%	18%
	Common dolphin	14%	**65%**	27%
	Spinner dolphin	6%	14%	**55%**
	Column total	100%	100%	100%

and categorized by type. Although entire tonals can be compared to one another on the basis of shape (Deecke and Janik 2006), the wide variability of whistle shapes quickly leads to a very large number of clusters as corpus size increases. As a result, clusters are no longer meaningful for categorization. A useful whistle categorization system needs enough categories to fully describe whistle shape variability but few enough categories that each type is seen repeatedly, across different encounters. By decomposing whistles into subunits, we can reduce the number of shape categories while retaining descriptive power, thereby facilitating shape-based comparisons within a large data set.

In this work, similarly shaped whistle subunits were grouped automatically using a network-based clustering approach. Normalization using a z-score transformation allowed subunits to be compared on the basis of shape, regardless of frequency content. The effects of z-score normalization can be seen in cluster 2 of network B (Fig. 8). This cluster includes subunits of very different bandwidths (some are nearly flat) because they have similar shapes postnormlization. The use of DTW allowed

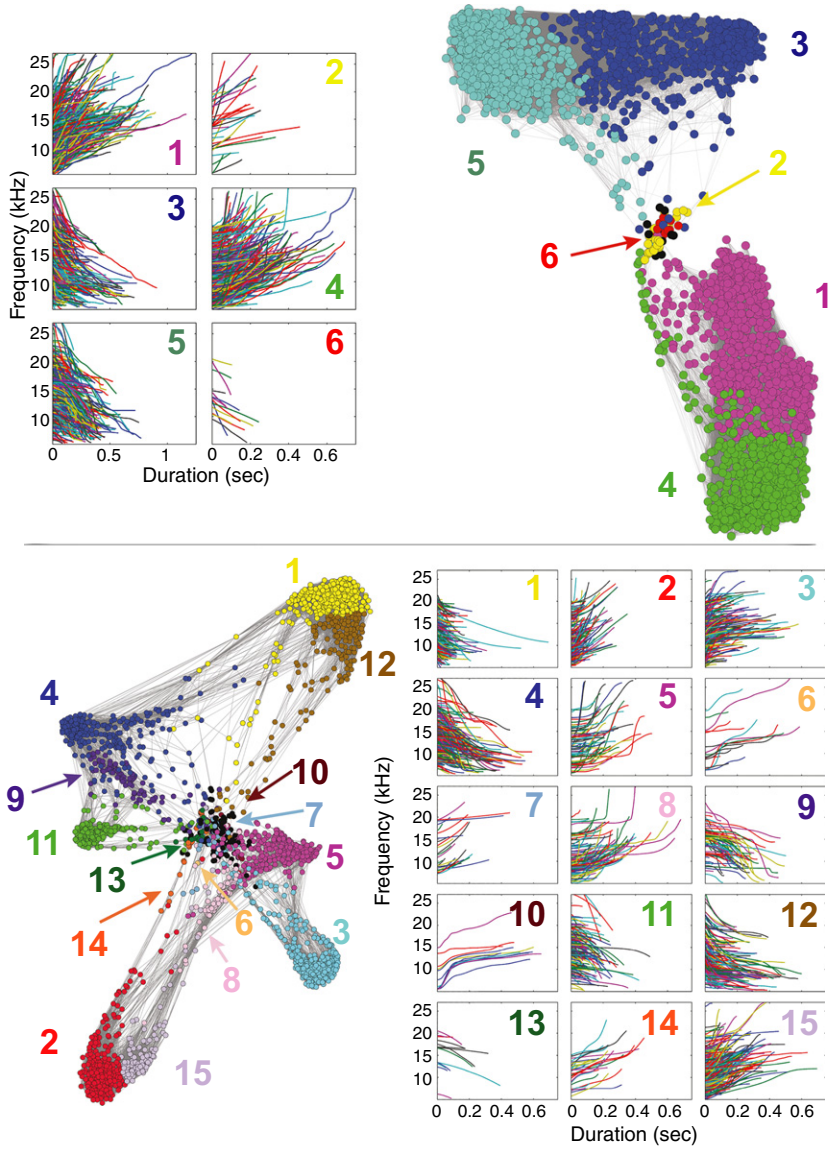

Figure 8. Visualization of node clusters from the short-beaked common dolphin corpus obtained using two different clustering parameter pairs. A: $p = 0.3$, $m = 1.5$; B: $p = 0.8$, $m = 0.5$. Network: colors indicate identified clusters. Each sphere represents a whistle component in the corpus. Gray lines between nodes represent linkages, with longer lines indicating less similarity. Black nodes are not assigned to a cluster. Network images generated using a force-directed layout routine *Force Atlas 2* (Jacomy *et al.* 2014) of the graph visualization tool *Gephi* (Bastian *et al.* 2009). Grid: each pane displays the original subunits contained in each cluster. Colored numbers identify the cluster associated with each pane.

comparison of subunit shapes despite differences in duration. Similarly shaped contours could be clustered together despite fairly large differences in duration, as in cluster 4 of network B (Fig. 8).

The desired level of cluster purity, *i.e.*, the similarity among subunits within a cluster, likely depends on the application of the user. If the goal is to describe whistle complexity as a function of behavior, for example, a few generalized categories might suffice, while efforts to understand signature whistle variability might require a larger number of detailed types. The algorithm outlined here can be optimized for either case (Fig. 8, short-beaked common dolphin whistle subunits; see Fig. S1, S2 for other species that are similar in nature) by adjusting the clustering parameters m and p.

Pruning weak connections between nodes by raising the pruning threshold p reduces computation time and improves visualization readability, while retaining important information about the network structure. Prior to pruning, every node in the network is linked to every other node, resulting in linkages for an node network. In large networks, a high percentage of the weakest linkages may be pruned. However, as p is increased, more nodes will become completely isolated from the network. These nodes often represent more complex or unique subunits, which will then be excluded from further analysis. High pruning thresholds werefound to negatively affect correct classification rates, presumably because informative nodes were removed from the training sets, thus indicating that complex or unique subunits contain useful information.

The choice of resolution coefficient m primarily affects number of clusters and individual cluster purity. High values of m will typically group subunits into four general categories: concave up, concave down, convex up, and convex down. As m is reduced, large clusters become subdivided into smaller, more self-similar groupings. A modularity-based clustering algorithm was used here because it is widely used and easily implemented using an existing code base. However, other clustering algorithms including hierarchical clustering (Bron and Kerbosch 1973), and clique-identification methods (Johnson 1967) could be used.

Reliability is critical if this method is to be used for comparison and classification of vocalizations. Our analyses demonstrate that this clustering approach is reasonably consistent in its categorizations across changes in training data as evidenced by the NMI metrics. These metrics provide a sense of the repeatability of the clustering results, rather than a measure of cluster quality. Both NMI metrics (NMI_P and NMI_S) have merit. Restricting the number of common whistle subunits n to those that were clustered in both partitions (NMI_P) gives a good interpretation of consistency amongst clustered whistle subunits, but does not tell the story of subunits that remain unclustered. Alternatively, penalizing whistle subunits that were not clustered (NMI_S) could be interpreted as providing a better indication of the algorithm's performance in general, although it should be recognized that some unclustered whistle subunits may simply be outliers with respect to the random sample and cannot reasonably expect to be clustered.

The classification results indicate that subunit shapes are useful for classifying tonal calls to species. No clear trend in classification performance occurred with respect to the resolution coefficient (m). In contrast, high values of the pruning coefficient (p) resulted in poorer classification performance. High values of the pruning coefficient result in the exclusion of subunits with shapes that are dissimilar to the rest of the training data, and it is possible that the exclusion of outliers contributes to overtraining of the classifier.

The trivial classifier implemented here looks only at relative occurrence of different subunit types. Variability in classification success was high in part due to the limited number of independent encounters in the data set. In some randomized folds, the encounters selected as the training set contained very few whistles, therefore classification success suffered. As several known parameters known to help distinguish whistles (*e.g.*, frequency; Oswald *et al.* 2003) are not part of the clustering process, it is likely that classification performance could be improved by incorporating these features into the classifier system. An alternative classification method in which each species' subunits were clustered independently was explored, however results were biased by relative training set sizes for each species, as well as cluster numbers and sizes, therefore further exploration normalization techniques for such alternative classification methods is needed.

Subunit clustering has the potential to inform several types of classification questions, ranging from species identification as shown here to other types of analyses such as social cooperation, behavioral state, *etc*. that may examine the sequences of subunits. Consequently, we see a variety of questions that could be addressed using subunit clustering as part of a broader methodology.

Conclusion

Subunits were identified within recorded whistles of the three delphinid species using automated methods. Network analysis was then used to cluster subunits according to their shape. Normalization and dynamic time warping of the whistle subunits allowed for categorization of distinct contour shapes rather than categorization based on time and frequency. Cluster composition remained similar across experiments despite variation in the training data, indicating that the clusters formed were nonrandom. Using the clustered subunits, a preliminary species classification scheme was implemented based on rates of subunit occurrence in vocal repertoires. This work suggests that segmentation of whistles into subunits may facilitate shape-based whistle categorization and comparison efforts.

Acknowledgments

This project was supported by Dr. Michael Weise at the Office of Naval Research, and by Dr. Frank Stone and Dr. Ernie Young at the Chief of Naval Operations division N45. Thanks to Dr. Simone Baumann-Pickering at Scripps Institution of Oceanography, Dr. Melissa Soldevilla at the NOAA Southeast Fisheries Science Center, and the Cascadia Research Collective for contributing the recordings used in this study. This data set was assembled for the 5th International Workshop on Detection, Classification, Localization, and Density Estimation, (DCLDE) of Marine Mammals using Passive Acoustics. Special thanks to John Hildebrand and the Scripps Institution of Oceanography Whale Acoustics Lab for support.

Literature Cited

Azzolin, M., A. Gannier, M. O. Lammers, *et al*. 2014. Combining whistle acoustic parameters to discriminate Mediterranean odontocetes during passive acoustic monitoring. The Journal of the Acoustical Society of America 135:502–512.

Bastian, M., S. Heymann and M. Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. Pages 361–362 *in* Proceedings International AAAI Conference on Weblogs and Social Media, San Jose, CA.

Blondel, V., J. Guillaume, R. Lambiotte and E. Lefebvre. 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 10:P10008.

Bron, C., and J. Kerbosch. 1973. Algorithm 457: Finding all cliques of an undirected graph. Communications of the ACM 16:575–577.

Buck, J. R., and P. L. Tyack. 1993. A quantitative measure of similarity for *Tursiops truncatus* signature whistles. The Journal of the Acoustical Society of America 94:2497–2506.

Deecke, V., and V. Janik. 2006. Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls. The Journal of the Acoustical Society of America 119:645–653.

Doddington, G. R. 2001. Speaker recognition based on idiolectal differences between speakers. Pages 2521–2524 *in* Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark.

Fred, A., and A. Jain. 2005. Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence 27:835–850.

Fritsch, F. N., and R. E. Carlson. 1980. Monotone piecewise cubic interpolation. SIAM Journal on Numerical Analysis 17:238–246.

Hastie, T., R. Tibshirani and J. H. Friedman. 2009. The elements of statistical learning data mining, inference, and prediction. Springer, New York, NY.

Jacomy, M., T. Venturini, S. Heymann and M. Bastian. 2014. ForceAtlas2, A continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLOS ONE 9:e98679.

Janik, V. M. 2009. Acoustic communication in delphinids. Advances in the study of behavior 40:123–157.

Janik, V. M., D. Todt and G. Dehnhardt. 1994. Signature whistle variations in a bottlenosed dolphin, *Tursiops truncatus*. Behavioral Ecology and Sociobiology 35:243–248.

Johnson, S. C. 1967. Hierarchical clustering schemes. Psychometrika 32:241–254.

Kershenbaum, A., L. S. Sayigh and V. M. Janik. 2013. The encoding of individual identity in dolphin signature whistles: How much information is needed? PLOS ONE 8:e77671.

Kershenbaum, A., D. T. Blumstein, M. A. Roch, *et al*. 2014. Acoustic sequences in non human animals: A tutorial review and prospectus. Biological Reviews. doi:10.1111/brv.12160.

Kreyszig, E. 1979. Advanced engineering mathematics. John Wiley & Sons Inc., Hoboken, NJ.

Lambiotte, R., J.-C. Delvenne and M. Barahona. 2008. Laplacian dynamics and multiscale modular structure in networks. arXiv preprint arXiv:0812.1770.

Lamel, L. F., and J.-L. Gauvain. 1994. Language identification using phone-based acoustic likelihoods. Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 94:I/293–I/296.

McCowan, B. 1995. A new quantitative technique for categorizing whistles using simulated signals and whistles from captive bottlenose dolphins (Delphinidae, *Tursiops truncatus*). Ethology 100:177–193.

Mellinger, D., and C. Clark. 2006. MobySound: A reference archive for studying automatic recognition of marine mammal sounds. Applied Acoustics 67:1226–1242.

Mucha, P. J., T. Richardson, K. Macon, M. A. Porter and J. P. Onnela. 2010. Community structure in time-dependent, multiscale, and multiplex networks. Science 328 (5980):876–878.

Myers, C., L. Rabiner and A. E. Rosenberg. 1980. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. Acoustics, Speech and Signal Processing, IEEE Transactions 28:623–635.

Newman, M. E. J. 2004. Detecting community structure in networks. The European Physical Journal B - Condensed Matter and Complex Systems 38:321–330.

Newman, M. E. J. 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America 103:8577–8582.
Oswald, J. N., J. Barlow and T. F. Norris. 2003. Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean. Marine Mammal Science 19:20–37.
Rabiner, L. R., and B.-H. Juang. 1993. Fundamentals of speech recognition. PTR Prentice Hall, Englewood Cliffs, NJ.
Ralston, J. V., and L. M. Herman. 1995. Perception and generalization of frequency contours by a bottlenose dolphin (*Tursiops truncatus*). Journal of Comparative Psychology 109:268–277.
Roch, M., T. Brandes, B. Patel, Y. Barkley, S. Baumann-Pickering and M. Soldevilla. 2011. Automated extraction of odontocete whistle contours. The Journal of the Acoustical Society of America 130:2212–2223.
Roch, M. A., J. Stinner-Sloan, S. Baumann-Pickering and S. M. Wiggins. 2015. Compensating for the effects of site and equipment variation on delphinid species identification from their echolocation clicks. The Journal of the Acoustical Society of America 137:22–29.
Shapiro, A., P. Tyack and S. Seneff. 2011. Comparing call-based versus subunit-based methods for categorizing Norwegian killer whale, *Orcinus orca*, vocalizations. Animal Behaviour 81:377–386.
Simonis, A., S. Baumann-Pickering, E. Oleson, M. Melcon, M. Gassmann, S. Wiggins and J. Hildebrand. 2012. High-frequency modulated signals of killer whales (*Orcinus orca*) in the North Pacific. The Journal of the Acoustical Society of America 131:EL295–EL301.
Steiner, W. W. 1981. Species-specific differences in pure tonal whistle vocalizations of five western North Atlantic dolphin species. Behavioral Ecology and Sociobiology 9:241–246.
Strehl, A., and J. Ghosh. 2003. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research 3:583–517.
Supin, A. Y., and V. Popov. 2000. Frequency-modulation sensitivity in bottlenose dolphins, *Tursiops truncatus*: Evoked potential study. Aquatic Mammals 26:83–94.
Thompson, R. K., and L. M. Herman. 1975. Underwater frequency discrimination in the bottlenosed dolphin (1–140 kHz) and the human (1–8 kHz). The Journal of the Acoustical Society of America 57:943–948.
Watrous, R. L., and L. Shastri. 1987. Learning phonetic features using connectionist networks. The Journal of the Acoustical Society of America 81:S93–S94.
Zhou, F., S. Mahler and H. Toivonen. 2012. Simplification of networks by edge pruning. Pages 179–198 *in* M. R. Berthold, ed. Bisociative knowledge discovery. Springer, Heidelberg, Germany.

Received: 5 November 2014
Accepted: 27 November 2015

Supporting Information

The following supporting information is available for this article online at http://onlinelibrary.wiley.com/doi/10.1111/mms.12303/suppinfo.

Figure S1. Visualization of node clusters from the spinner dolphin corpus obtained using two different clustering parameter pairs. Upper: $p = 0.3$, $m = 1.5$; Lower: $p = 0.8$, $m = 0.5$.

Figure S2. Visualization of node clusters from the bottlenose dolphin corpus obtained using two different clustering parameter pairs. Upper: $p = 0.3$, $m = 1.5$; Lower: $p = 0.8$, $m = 0.5$.