Using self-organizing maps to classify humpback whale song units and quantify their similarity

Jenny A. Allen[1], Anita Murray, Michael J. Noad, and Rebecca A. Dunlop
Cetacean Ecology and Acoustics Laboratory, School of Veterinary Science, University of Queensland, Gatton, QLD, 4343, Australia

Ellen C. Garland
School of Biology, University of St Andrews, St Andrews, Fife, KY16 9TH, UK

Author to whom correspondence should be addressed. Electronic mail: j.allen3@uq.edu.au

Running title: Quantitative classification using signal similarity

Keywords: animal communication; sequence analysis; neural networks; humpback whale

PACS: 43.80.Ev, 43.80.-n

Word Count: ~ 6640 (Supplemental Materials excluded)

Date uploaded on JASA manuscript submission system: 20/3/2017

ABSTRACT

Classification of vocal signals can be undertaken using a wide variety of qualitative and quantitative techniques. Using east Australian humpback whale song from 2002-2014, a subset of vocal signals were acoustically measured and then classified using a self-organizing map (SOM). The SOM created 1) an acoustic dictionary of units representing the song's repertoire, and 2) Cartesian distance measurements among all unit types (SOM nodes). Utilizing the SOM dictionary as a guide, additional song recordings from east Australia were rapidly (manually) transcribed. To assess the similarity in song sequences, the Cartesian distance output from the SOM was applied in Levenshtein distance similarity analyses as a weighting factor to better incorporate unit similarity in the calculation (previously a qualitative process). SOMs provide a more robust and repeatable means of categorizing acoustic signals along with a clear quantitative measurement of sound type similarity based on acoustic features. This method can be utilized for a wide variety of acoustic databases especially those containing very large datasets, and be applied across the vocalization research community to help address concerns surrounding inconsistency in manual classification.

I. INTRODUCTION

Acoustic signals are commonly used for communication in a variety of species and signals typically convey different kinds of information. Information can range from simple species identification (Gerhardt, 2001) to complicated ideas such as foraging (Slocombe and Zuberbühler, 2006) or social hierarchy (Catchpole and Slater, 2008). Vocal studies are therefore imperative to understanding a broad range of concepts such as species distribution, signal information content, or vocal learning. One major hurdle for any vocalization study is a precise means to analyze data (Kershenbaum *et al.*, 2014). Acoustic features such as duration or frequency can be quantified (Tchernichovski *et al.*, 2000; Cerchio *et al.*, 2001),yet these features do not always provide complete signal representation (Janik, 1999). As a result signals are often classified into categories qualitatively by a human observer (Janik, 1999; Kershenbaum *et al.*, 2014).

Manual classifications can be corroborated by several means. Naïve matching tests compare agreement between independent observers (e.g., Garland *et al.*, 2011). Quantitative testing can also assess manual classification, including multivariate statistics such as discriminant function analysis (DFA) (e.g., Dunlop *et al*., 2007), Classification And Regression Trees (CART) (e.g., Melendez *et al.*, 2006, Rekdahl *et al.*, 2013) or Random Forest analysis (e.g., Risch *et al.*, 2013, Garland *et al.*, 2015). Despite quantitative support, classifying signals remains largely qualitative. Automated methods provide more objectivity, but cannot always be implemented if signals are too varied or complex (Janik, 1999). Subjectivity is a key weakness in vocalization studies: it impedes standardized classification across studies of the same vocal display, and there is no reliable way to determine if classifications are biologically relevant to the study species. Different methods are therefore required that can move classification towards a more repeatable and objective approach.

One such technique is an artificial neural network called a Self-Organizing Map (SOM) (Kohonen, 1990). What makes the SOM such a beneficial tool is that it uses an “unsupervised” learning algorithm: there is no parameter selection of the data’s variables or user feedback involved in the target classification outputs (Suzuki *et al.*, 2006; Green *et al.*, 2007; Kohonen, 2014). Unsupervised learning removes a degree of the subjectivity that can come from predetermining how to group information, which occurs in “supervised” learning (Kohonen, 1990; Green *et al.*, 2007). It also allows for the possibility of recognizing patterns that may not be apparent to a human observer (Green *et al.*, 2007). This is advantageous given the aforementioned difficulty with determining a feature’s biological relevance.

SOMs organize information into a 2-dimensional “output space” (Bauer and Pawelzik, 1992), made up of ‘nodes’ which serve as the categories into which data will be grouped. Before this can happen, the map must learn to classify the dataset in question. Acoustic signals within the dataset are each represented by an input vector of values (i.e. each vector is the list of measured variables). Training occurs by repeatedly presenting the map with each of the input vectors. Each node contains a weight vector of the same length as the input vectors, and the nodes learn to respond to the data during training (Kohonen, 1990). A principal component analysis on the input vectors provides initial values for the weight vectors (Hagan *et al.*, 1996; Kohonen, 2014). SOMs can then place a signal into whichever node has the weight vector that best matches its input vector (Kohonen, 1990; Walker *et al.*, 1996). The spatial arrangement of the nodes is dictated by two parameters: neighborhood size and learning rate. Learning rate controls the extent to which a node is altered, while neighborhood size determines how many surrounding nodes are affected by those alterations (Hagan *et al.*, 1996; Callan *et al.*, 1999). The result is that more similar nodes are arranged to have closer proximity to one another within the map. An added advantage of this spatial arrangement is that the distance between nodes can be measured in either Euclidean or Cartesian space. These measures serve as a means of quantifying similarity between sound types, which can then be utilized in subsequent analyses (Garland *et al.*, 2017). SOMs have been used as a method for analyzing vocal signals in species such as domestic pigs (*Sus scrofa)* (Schön *et al.*, 2001), white-crowned sparrows (*Zonotrichia leucophrys pugetensis*) (Ranjard and Ross, 2008), and humans (Callan *et al.*, 1999).

SOMs appear particularly useful in the classification of humpback whale song units (Walker *et al.*, 1996; Mercado and Kuh, 1998; Suzuki *et al.*, 2006; Green *et al.*, 2007; Kaufman *et al.*, 2012; Murray *et al.*, 2016). Humpback whale song has a hierarchical structure consisting of sound units repeating in a set pattern to make up a phrase. Phrases then repeat a number of times to form a theme. Themes are repeated sequentially to make up a song cycle (Payne and McVay, 1971; Payne and Payne, 1985; Cholewiak *et al.*, 2013). Although all males in a population typically sing the same song pattern at any given time, the song tends to changes progressively (Payne *et al.*, 1983; Payne and Payne, 1985). Recent work by (Murray *et al.*, 2016) expanded on the use of acoustic features for song unit classification by measuring the frequency contours of tonal sounds, and including them as variables in the SOM classification. Classification results were then used to transcribe phrases into numeric strings to represent the unit sequences of those phrases. The Levenshtein distance, a similarity analysis that is highly suited to comparing vocal sequences (Kershenbaum *et al.*, 2014), was then used between transcribed sequences along with cluster analyses to quantitatively identify themes.

The degree of complexity and rapid evolutionary change found in humpback whale song make it an ideal model to test the robustness and repeatability of this methodology in highly complex vocal displays. While similar prototypes have been generated before (Walker *et al.*, 1996; Mercado and Kuh, 1998), the current study expands on this by creating an acoustic dictionary, a task that has yet to be undertaken in vocalization research (Placer *et al.*, 2006). The size of many acoustic datasets often makes it impractical to measure every signal required to generate large sample sizes of vocal sequences. A dictionary can serve as a guide for the rapid transcription of new, unmeasured recordings into numeric sequences, bolstering sample size. Additionally, by applying SOM distance measurements that provide a quantitative measure of unit similarity in higher-level (sequence) analyses, the utility and repeatability of transcription using this dictionary is apparent. The relative efficiency of SOM classification is also investigated in comparison to the manual classification method when based on the same input data. Use of the SOM method described here provides a more repeatable and robust means of classifying acoustic signals, along with the application of quantified signal similarity in higher-level analyses in the complex song hierarchy. The current study aims to 1) to create an acoustic dictionary of humpback song units for one population over multiple years, 2) extract a means of quantifying similarity between those song units, 3) test the classification of sounds by the SOM against qualitative classification using CART and RF analyses, and 4) use sequence analysis to demonstrate the utility of applying both the acoustic dictionary and quantitative similarity measures to new recordings.

II. METHODS

A. Study Sites

Data used in the current study were collected off the coast of Peregian Beach (26°30’ S, 153°05’ E), located on the Sunshine Coast in Queensland, Australia (Fig. 1a) as well as Point Lookout (27°43’ S, 153°53’ E), located on North Stradbroke Island, Queensland, Australia (Fig. 1b). Both locations are along the migratory corridor of east Australian humpback whales where the whales often swim within a few kilometers of the shoreline (Paterson and Paterson, 1984; Noad and Cato, 2001).

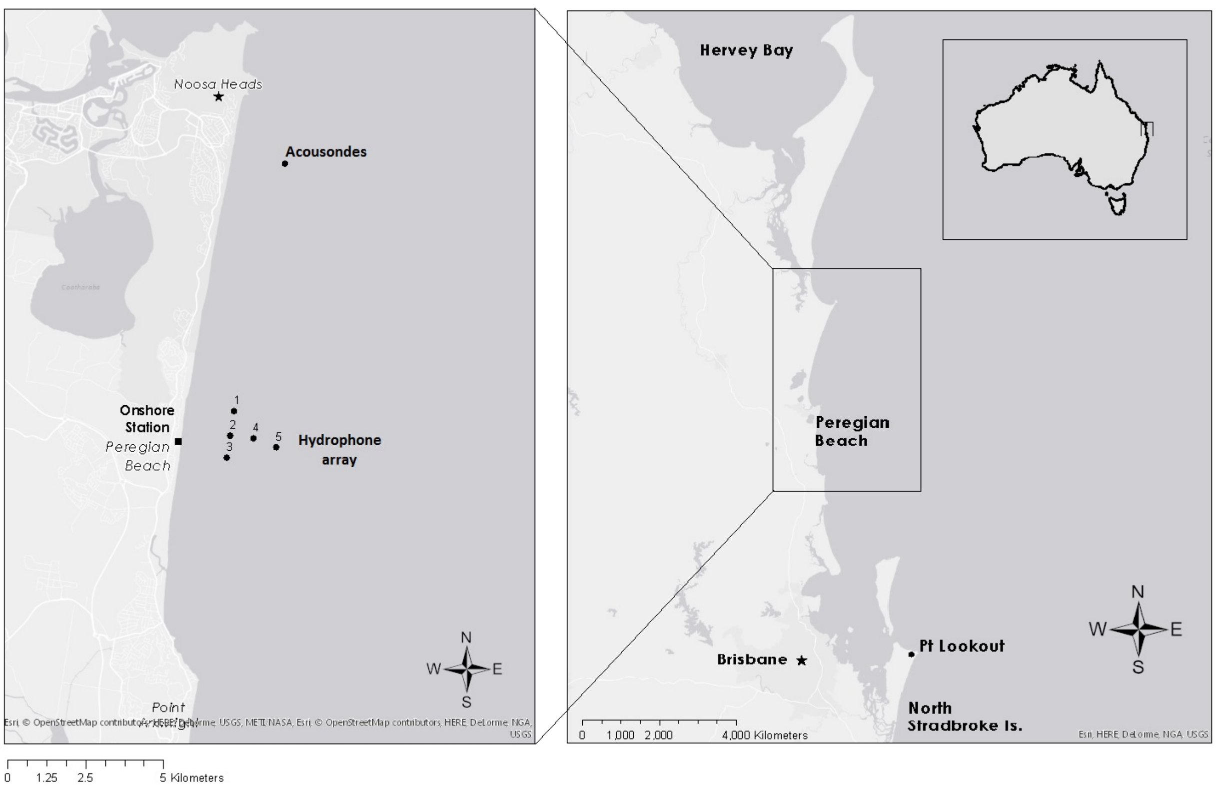


FIG. 1. East Australia study sites: Peregian Beach and Point Lookout. The panel on the left shows the placement of the hydrophone array (hydrophone buoys are numbered 1-5) and the autonomous recorder deployments. The panel on the right shows the relative distance between the two study sites.

B. Data Collection

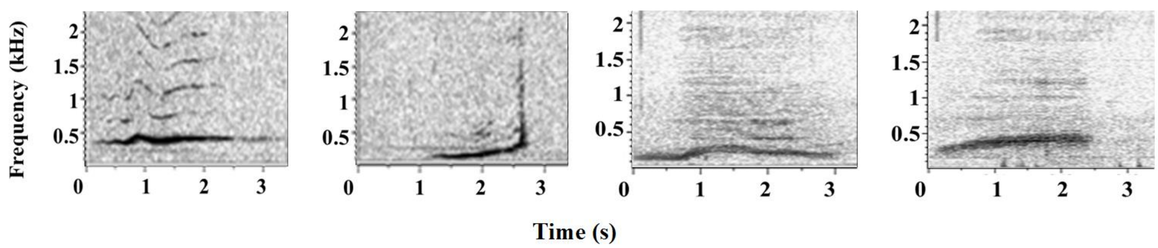
Recordings from 2002-2014 were made using several platforms. A moored hydrophone array consisting of five buoys was deployed off of Peregian Beach in 2002-2004, 2008-2011, and 2014 (Fig. 1a). Each buoy had a High Tech HTI-96-MIN hydrophone with a built-in pre-amplifier (+40 dB), a customized amplifier (+20 dB), and a VHF radio transmitter (AN/SSQ-47A). They were set up 1.5 -2.5 km from shore, spaced approximately 750 m apart at depths of 18-28 m. Buoy signals were received at an onshore base station using a four-channel type 8101 Sonobuoy VHF receiver (buoys 1-4), or a single channel Sonobuoy frequency converter connected to a commercial FM radio receiver (buoy 5). Signals were digitized using a National Instruments E-series data acquisition card and recorded to a desktop computer with *Ishmael* acoustic software (Mellinger, 2001) at a sampling rate of 22 kHz, 16 bit depth, and stored as multi-channel WAV files. These recordings were supplemented with boat-based recordings using Cleavite CH17, GEC Marconi SH101X, or High Tech Inc. HTI-96-MIN hydrophones connected to Sony DAT, Microtrack, or Zoom digital recorders (generally using 44.1 kHz sampling rate, 16 bit depth, frequency response 30 Hz-20 kHz). Boat based recordings were the sole source of data in 2005-2007.

Autonomous underwater acoustic recorders were placed off the coast of Peregian Beach in 2012-2014. Each of the two recorders (Acousonde 3A with external battery housings, Greenridge Sciences, www.acousonde.com) had a sampling rate of 25,818 Hz with a 9 kHz low pass filter and a gain of 20 dB. Both Acousondes were placed in the same location, approximately 1.5 km from the shoreline (Fig. 1a). Each was set on alternate 12 hour duty cycles, resulting in essentially continuous recording for the duration of each deployment. All recordings covered the frequency range of humpback whale song.

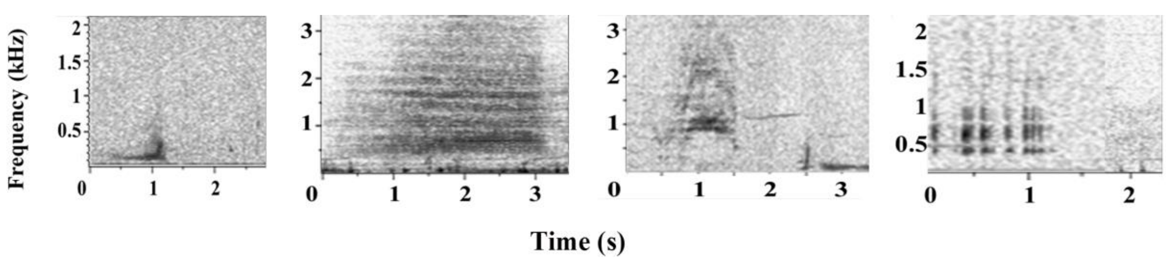
C. Measurement of Acoustic Features of Sound Units

Recordings of songs were visualized as spectrograms in Raven Pro 1.4 (www.birds.cornell.edu/raven) using a Fast Fourier Transforms with Hann window, and 90% overlap. Good quality spectrograms were defined by a signal-to-noise ratio (SNR) of at least 10 dB above the background noise. Six complete song cycles from a singer in each year (2002-2014) were selected for measurement. Themes, phrases, and units were identified based on the accepted hierarchical structure of humpback whale song as described in Payne and McVay (1971). The exception was 2007, in which only four song cycles were selected due to a lack of available, high quality recordings. This resulted in 76 complete song cycles from 13 individuals being selected for acoustic measurement. From each of the six song cycles in a given year, three phrase repetitions of each theme were selected for measurement based on the highest quality repetitions within the recording (high SNR). The aim of the current study was to create a set of general representative sound types, and thus every atypical signal need not be represented. A subsample of phrase repetitions addresses variability found within themes while preventing overrepresentation of themes whose phrases are repeated with disproportionately high frequency. Further, the three phrase repetitions were taken from the beginning, middle, and the end of the theme to account for shifting themes that change subtly over multiple repetitions (Payne and Payne, 1985). A total of 3720 phrases from the 76 complete song cycles were selected and utilized for acoustic measurement.

Sound units were separated into two groups prior to measurement, contoured and non-contoured, which have distinctly different feature profiles (Dunlop *et al.*, 2007; Murray *et al.*, 2016). Separate methods were used in order to measure the acoustic features of each sound type in more detail (following Murray *et al.*, 2016). Contoured units have a definitive and traceable shape, such as tonal and harmonic units, as well as complex units containing both broadband and harmonic elements (see examples in Fig. 2a) (Dunlop *et al.*, 2007). Non-contoured units have no traceable shape or harmonic elements, such as purely broadband and pulsed calls (see examples in Fig. 2b). The decision to separate units allows for the use of contour tracing software, which provides multiple frequency measurements along the contour of a sound. This results in a more comprehensive representation of tonal and complex sounds by quantifying a signal's shape. A frequency contour cannot be generated for non-contoured units due to the lack of a traceable shape, necessitating the use of two different methods of measurement.



a)



b)

FIG. 2: Spectrogram examples of a subset of the a) contoured and b) non-contoured units. All spectrograms were generated in Raven Pro 1.4 using 2048 FFT, Hann window, 90% overlap.

1. Contoured Feature Measurement

Contoured sound units were measured using the frequency contour tracing program *Beluga* (http://biology.standrews.ac.uk/soundAnalysis/), within *MATLAB 2014b* (The

MathWorks Inc, 2014). Recordings were imported into *Beluga* as WAV files. A spectrogram was calculated using an FFT of 2048, frame length of 1024, 93.75% overlap between frames, and Hanning window function. A tracing box was placed around the entire signal (Fig. 3a), and the recording was filtered to remove the average noise spectrum. The frequency contour was extracted using the "peaks" method without harmonics, measuring peak frequency every 0.03 seconds along the signal and creating a vector with a length analogous to the unit's duration (Fig. 3b). SOMs require vectors of equal length; therefore, contour vectors were truncated by extracting fifty equally spaced points along the vector. Each point was treated as a separate variable, similar to the computations method of classification developed by McCowan (1995).

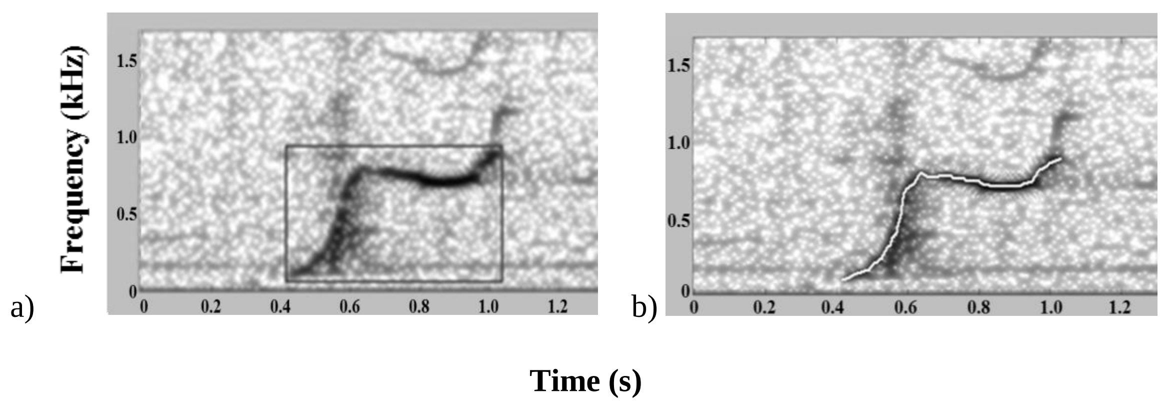


FIG. 3: Spectrogram example of the *Beluga* contour tracing method, showing a) the tracing box around the signal and b) frequency contour trace

Additional measurements extracted from *Beluga* were: minimum frequency, maximum frequency, start frequency, stop frequency, duration, trend, and bandwidth (see Table I for full descriptions). Inflections, defined as changes in the slope of the frequency contour, were counted based on the extracted contour of the sound (following Dunlop *et al.*, 2007). Pulse repetition rate (PRR) was counted (per second) using the Raven spectrograms and corresponding waveforms from which these units were originally transcribed.

2. Non-Contoured Feature Measurement

Non-contoured units were measured using the robust measurements available in Raven Pro 1.4 (Charif *et al.*, 2010). Recordings were imported into Raven as WAV files. Spectrograms of recordings were loaded with an FFT of 2048, Hann window, and 90% overlap. A tracing box was placed around units and the following features were extracted: duration, center frequency, peak frequency, frequency 5%, frequency 95%, and bandwidth 90% (Table II). Inflection and pulse repetition rate (PRR) were counted visually based on the spectrogram and corresponding waveform.

D. Creating a Self-Organizing Map

Self-organizing maps (SOM) were created using the *selforgmap* function of the Neural Network Toolbox in *MATLAB* 2014b. There were 59 acoustic features (9 variables and 50 frequency contour points) in the contoured input vectors, and 8 acoustic features in the non-contoured input vectors. Z-scores were used to standardize the data in order to account for the variety of different variable scales. Separate maps were created for the two types of signals due to the different methods of acoustic feature measurement described above (following Murray *et al.*, 2016). Map sizes that divide data too coarsely over-simplify differences, while dividing it too finely creates categories with superfluous detail (Walker *et al.*, 1996; Céréghino and Park, 2009). Map dimensions were therefore determined using trial and error (Kohonen, 2014). Due to the current study's aim of creating generalized sound types, 'lumping' signals into fewer broad groups was favored over 'splitting' them into many smaller ones that would not represent generalized categories (Mercado and Kuh, 1998). The resulting dimensions were a 10 x 10 map (100 nodes) for contoured units and a 7 x 7 map (49 nodes) for non-contoured units. Once dimensions were established, the SOM was trained and created using the dataset, with neighborhood size and learning rate kept at the default MATLAB settings of 3 and 0.01 respectively (Demuth *et al.*, 2014). The chosen dimensions determined the number of nodes, or groupings into which the data were placed. Each measured signal was placed into a single node.

E. Comparison of SOM and Qualitative Classification

Classification and Regression Tree (CART) (Breiman *et al.*, 1984) and Random Forest (Breiman, 2001) analyses were used to assess the relative consistency between SOM and manual classification techniques when given the same set of data and input variables. Prior to the formation of the map, the measured sounds were also qualitatively assessed and classified by JA resulting in 261 contoured sound types and 42 non-contoured sounds. Agreement between the method of classification and the decision tree analyses were calculated for each classifying technique separately. Contoured and non-contoured units also had to be evaluated separately due to the differences in their acoustic variables. Multivariate PCA and DFA are commonly used analysis methods for corroboration of qualitative data categorization, particularly for animal vocalization (Boisseau, 2005; Dunlop *et al.*, 2007; Rekdahl *et al.*, 2013). However, CART analysis addresses assumptions made by these analyses; data can be non-parametric, non-normal, and have correlated variables (Van Opzeeland and Van Parijs, 2004; Melendez *et al.*, 2006; Garland *et al.*, 2012; Rekdahl *et al.*, 2013). CART decision trees split data into branches based on the Gini Index, a commonly used measure of "goodness of split" which reduces heterogeneity within the groups (Breiman *et al.*, 1984). At each split of the tree, all possible divisions to the data (by variable) are considered. This allows division of data to be based on a different splitting criterion at each branch (e.g., is start frequency > 500 Hz). The criterion chosen represents the highest reduction in heterogeneity in the data (Karels *et al.*, 2004). CART was implemented here with cross-validation using the *rpart* package in R (Therneau *et al.*, 2014), with each terminal branch of the CART (analogous to a node or a category) set to a minimum size of 10 (Table III). Each of the resulting decision trees were pruned to prevent overfitting of the data using the 1-standard deviation rule (see Breiman *et al.*, 1984). CART provides information on the ability of the analysis to classify calls (root node error) and also the agreement in classification between CART and the classification technique it is evaluating.

Random Forest is a more robust expansion of CART, where a forest of CART trees is created to allow an internal estimate of uncertainty. By applying a bootstrapping technique known as 'tree bagging' to the process of creating decision trees, Random Forests can randomly sample combinations of the variables available to produce the lowest out-of-bag (OOB) error rate. This allows an estimate of classification error per call type and the overall OOB error rate of the forest, from which classification agreement can be determined. Random Forest was implemented here using the *randomForest* package in *R* (Table III) (Liaw and Wiener, 2002), with 1000 trees grown for each forest and the predictor variables that were randomly selected set to 3. The Gini Index was also used here to indicate the importance of each of the predictor variables. Gini values indicate order of relative variable importance in the splitting decisions and are not directly comparable across separate analyses.

CART and Random Forest analyses were each used to evaluate the two classification techniques: 1) manual, or qualitative description (**Q**), and 2) SOM node placement (**SOM**).

Contoured (**C**) and non-contoured (**NC**) units were analyzed separately given that they were measured differently. The dataset of contoured units was classified independently by both the SOM (**C-SOM**) and qualitatively (**C-Q**). The dataset of non-contoured units was also classified by both the SOM (**NC-SOM**) and qualitatively (**NC-Q**). Each of the four classifications was treated as a separate subset of the data. Each subset was evaluated separately for classification agreement with a CART analysis, as well as with a Random Forest analysis, for a total of eight analyses. A non-parametric Mann-Whitney/Wilcoxon test was used to compare the degree of classification agreement found for each method.

F. Utilizing SOM Cartesian distances to quantify song similarity

To quantify the relative acoustic similarities between prototype units, the distance between the nodes was measured on the Cartesian plane as arranged by the SOM spatial layouts (Fig. 4). Each SOM was placed on a two-dimensional plane and every node was assigned an (X,Y) coordinate with all adjacent nodes having a distance of 1. Based on these coordinates, a matrix was generated of all the relative Cartesian distances between the nodes in the SOM layout. This matrix provided a quantitative measurement of relative similarity among unit types based on their spatial arrangement in the SOM.

To demonstrate the utility of SOMs in combination with the similarity weightings, song cycles from the East Australian population in 2008 were transcribed following the prototype units generated from the SOM classification as a guide. Qualitatively identified themes within the 2008 song were then validated using Levenshtein distance analysis of the phrase repetitions transcribed using the SOM classifications. The Levenshtein distance is a similarity measurement that calculates the minimum number of insertions, deletions, and substitutions needed to convert one string of data into another. This score can then be normalized to account for differences in string length, creating an index of similarity known as the Levenshtein distance similarity index (LSI) (Helweg *et al.*, 1998; Garland *et al.*, 2012; Murray *et al.*, 2016). Here, a weighted LSI analysis was implemented where the cost matrix for substituting units was based on the matrix of Cartesian distances extracted from the SOM, exponentially scaled between 0 and 1. This allowed the cost of substituting similar units to be a direct measure of acoustic similarity and the cost of insertions or deletions remained as cost=1 (see Garland *et al.* (2017) for detailed methodology and rationale). In essence, substitutions between highly similar units were considered to be less costly (based on SOM distances), while insertions, deletions, and substitution of units from separate maps were assigned a maximum penalty of cost=1. If themes that were qualitatively identified within the 2008 song could also be identified through the Levenshtein Distance analysis, it would demonstrate the repeatability of the transcriptions made using the acoustic dictionary. Average-linkage hierarchical cluster analysis and bootstrapping (using *pvclust* and *bootstrap* in *R*) were run to assess the similarity between all data strings. The cophenetic correlation coefficient (CCC) was also calculated as a measure of how accurately the above analyses represented the true similarity associations within the data, with a CCC>0.8 indicating a good representation of the data (Sokal and Rohlf, 1962).

III. RESULTS

A. Creation of Prototype Units

From 76 song cycles and 3720 phrases, 6409 sound units were measured and placed in 149 SOM nodes, 100 nodes within a 10x10 contoured SOM and 49 nodes within a 7x7 non-contoured SOM. For each node, the average of each acoustic feature was calculated using all of the units placed in that particular node, creating feature vectors for a set of prototype units (Supplementary Materials, Table VI and VII). For the contoured SOM, each of the 50 frequency contour points within a node was averaged and graphed, creating a visual representation of the prototype unit for each node (Fig. 4). A visual representation of the non-contoured prototype units was not possible because there was no frequency contours to extract. Nodes were numbered from left to right, starting from the upper left node and ending with the lower right node. Prototype units were numbered 1-100 for contoured units based on their SOM node position, and from 101-149 for the non-contoured units. These units comprise an acoustic dictionary of sound units which represents the song repertoire from 2002-2014 for the East Australian humpback whale population.

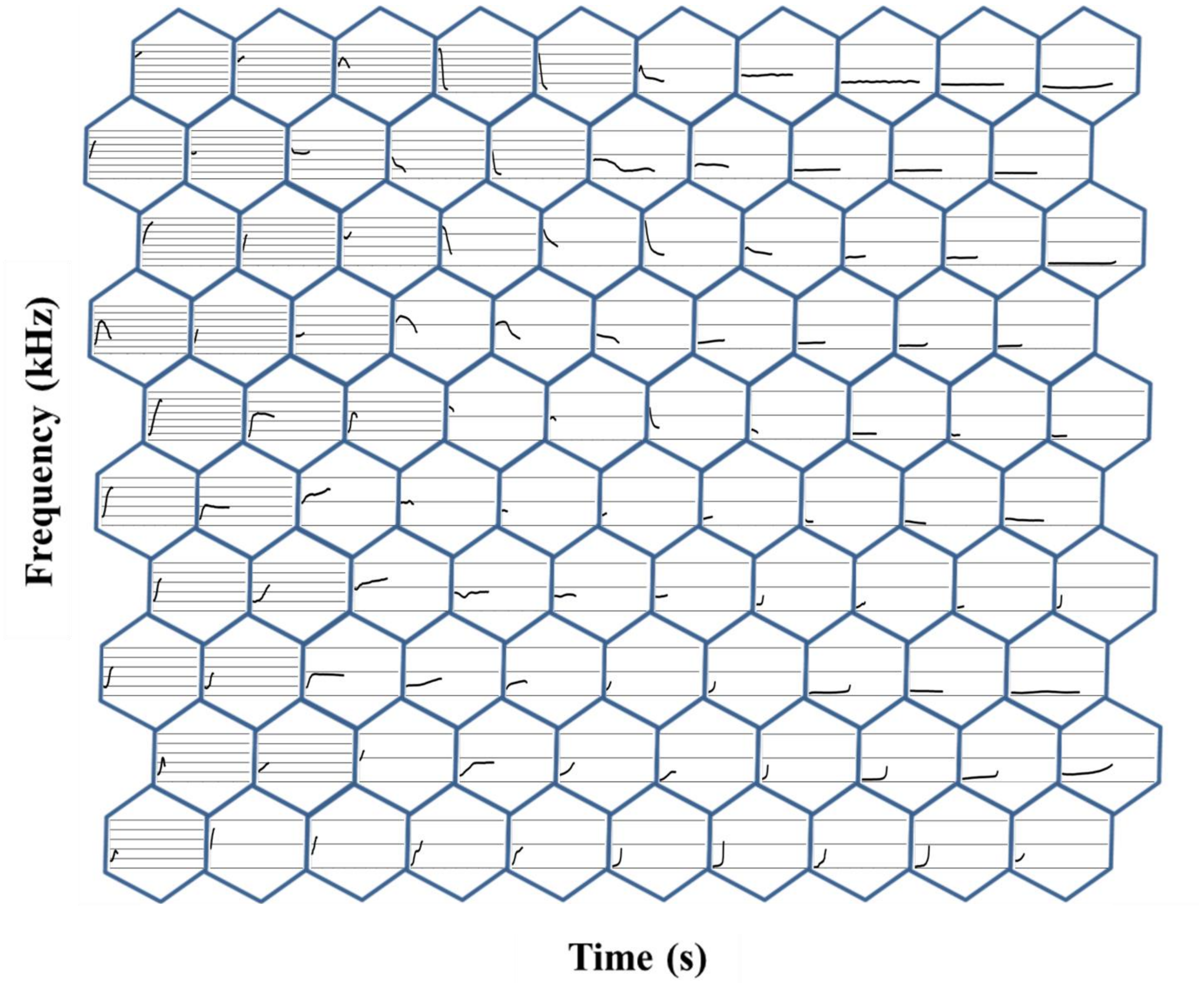


FIG. 4. Visual representations of prototypical unit contours generated from the contoured unit 10 x 10 SOM, based on the 50 frequency contour points extracted using Beluga. All visual representations have time on the x-axis (5 seconds for all nodes) and frequency on the y-axis (gridlines represents one kilohertz intervals). Adjacent nodes are more similar to each other than those that are not adjacent.

B. CART Analyses

For each of the CART analyses, a proportion of variables provided a root node error. This resulted in an agreement of classification between the classification technique (either qualitative or SOM) and the CART analysis. A summary of the classification agreements for each analysis can be found in Table III. The top five variables used by the analyses and their respective Gini Index values in each analysis can be found in Table IV.

C. Random Forest Analyses

For each of the Random Forest analyses, agreement in classification between the classification technique (either qualitative or SOM) and the Random Forest analysis was reported, as well as the most important variables as assessed by the Gini Index. A summary of classification agreements for each analysis can be found in Table III. The top five variables used by the analyses and their respective Gini Index values in each analysis can be found in Table V.

D. Comparison of SOM and Qualitative Classification

Results of the comparison between SOM and qualitative classifications are summarized in Table III. Classification agreement with the CART analysis was found to be significantly higher with the SOM technique (73%) as compared to the manual method (58%; Mann-Whitney/Wilcoxon, $W=4770.7$, $p<0.01$) for contoured units, but there was no significant difference in non-contoured units (Mann-Whitney/Wilcoxon, $W=918$, $p=0.48$). Classification agreement with the Random Forest analysis was found to be significantly higher with the SOM technique for both contoured (89% vs 73%; Mann-Whitney/Wilcoxon, $W=3987.5$, $p<0.01$) and non-contoured units (91% vs 83%; Mann-Whitney/Wilcoxon, $W=685$, $p<0.01$).

E. Utilizing the SOM prototypes and Cartesian distances to quantify song similarity

Using the SOM classifications, 36 complete song cycles of the 2008 song were transcribed from nine singers, comprising 7847 sound units arranged into 1864 phrases. No song cycles measured for the original SOM analyses were used in this analysis to ensure independent sampling. A dendrogram was generated based on LSI values using both hierarchical cluster analysis and bootstrapping to display similarity between phrases (Fig. 5). The cophenetic correlation coefficient (CCC) of 0.97 verified that the dendrogram was a very good representation of the associations within the dataset. Most phrase repetitions of a given qualitatively-identified theme were clustered together on the same major branch: therefore, each major branch represented a different theme. The exception was Theme D, which contained three phrase variants based on different phrase lengths (D1: two units, D2: three units, and D3: five units). A qualitative examination of these variants (Fig. 5) showed that all three variants contained the same two starting units. For example, to create D2, the three-unit variant, one unit was inserted at the end of D1, the two-unit sequence. To create D3, the five-unit variant, two additional units were inserted to the end of D2 (the three-unit sequence). Differences in length are reflected in the LSI analysis, as insertions and deletions which lengthen or shorten a string were more heavily penalized in this weighted LSI framework than substitutions (Garland *et al.*, 2017).

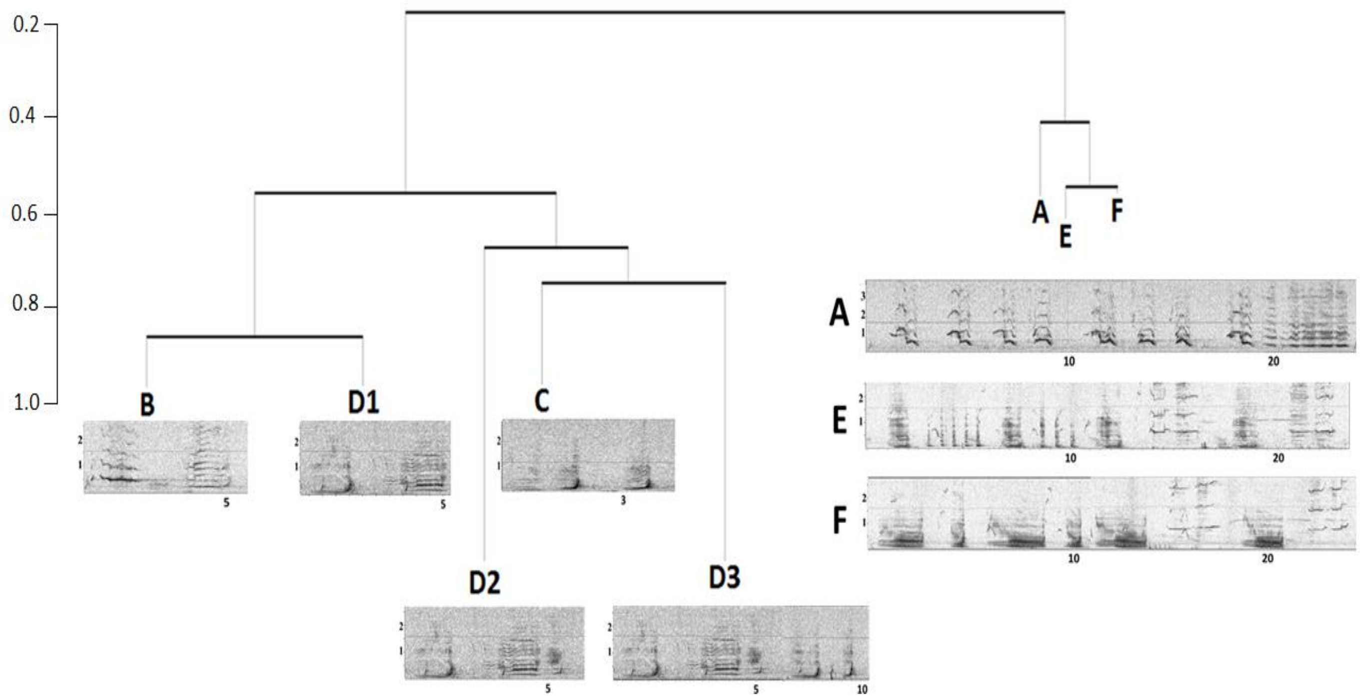


FIG. 5: Average-linkage hierarchically bootstrapped dendrogram of the East Australian 2008 song based on the Levenshtein Similarity Index (LSI), which was weighted for substitutions using the Cartesian distances between units in the SOM. Horizontal lines correspond to the proportion of similarity, shown on the y-axis, between two branches. Each letter represents a qualitatively identified theme. Phrase repetitions of every theme, with the exception of Theme D, were clustered onto separate major branches. Spectrogram figures provide a visual representation of each theme, with time (s) on the x-axis and frequency (kHz) on the y-axis. Note that only major branches are shown; terminal branches representing individual phrase repetitions were excluded for clarity.

IV. DISCUSSION

SOM classification enabled the creation of an acoustic dictionary of prototypical units, which represents the repertoire of the east Australian humpback whale population's songs from 2002-2014. The Cartesian distances between those units, a valuable product of the SOM classification, provided a means of quantifying the similarity between all units across the entire dictionary, which can be utilized in higher-level sequence analyses (Garland *et al.* (2017). This dictionary can serve as a guide by which vocal sequences from new recordings can be manually transcribed in a rapid, repeatable, and efficient manner. While prototypical units have been created to represent humpback whale song before (Walker *et al.*, 1996; Mercado and Kuh, 1998), small sample size in many of these studies limited their ability to be representative of an entire repertoire over multiple years. Furthermore, none quantified the acoustic similarities between their units. Cartesian distances as unit similarity weightings were instrumental to the repeatability of the dictionary's application to a dataset. There will inevitably be variation in signal classification for manual transcriptions for sequences. Quantifying similarity across units allowed the Levenshtein Distance analysis to identify and cluster repetitions of a specific theme despite those variations. The splitting of one theme's variations onto several branches based on length and unit types reveals the important role that qualitative judgment still plays in the classification and analysis of sequences. Ultimately a dictionary can minimize the amount of work needed to analyze large volumes of data; it requires only a relatively small subset of acoustic signals to be individually measured. Given that acoustic databases can contain hundreds of hours of recordings, comprehensive analyses can be difficult if every unit must be measured. Measuring a representative subsample to create a dictionary should increase the sample size of recordings that could ultimately be used for further analysis in many types of vocalization studies.

Precedence exists for SOM signal classification in a number of species, and it has some advantages over the manual technique. Although entirely automatic techniques would be the most objective, vocal signals often have too much variation for these to be effective (Janik, 1999). SOM classification eliminates one of the many qualitative steps within the study of vocalizations by placing signals into categories through quantitative and repeatable means. Map size is subjectively derived, but an advantage of this is that it allows for flexibility in studies of vocalizations at different resolutions. Small maps can be used for broad-scale contexts like territories or inter-population variation, while larger maps can be used for fine-scale detail such as individual variation. When implementing the dictionary on new, unmeasured recordings, the prototype unit that is ultimately selected as the best match for a signal is still manually decided. The similarity weightings derived from the SOM account for the variations in manual classification that occur due to subtle differences or similarities in unit types that may be identified by the human observer.

CART and Random Forest analyses provided a quantitative means of directly comparing between SOM and manual classification techniques. Both analyses found significantly higher classification agreement when contoured units were classified by the SOM method as compared to being classified manually. While Random Forest also found significantly higher agreement when non-contoured units were classified by the SOM, there was no significant difference in classification agreement when non-contoured units were classified either SOM or manually. This implies that the SOM method is more effective for contoured sounds. Acoustic characteristics can impact which technique might be better suited to each signal type. Subtle differences in the contour of tonal sounds may be obscured to a human observer, particularly in cases of repetitive sequences with gradually changing units. Conversely, acoustic measurements of non-contoured units may not necessarily create a comprehensive description of the signal. It should be noted, however, that biological relevance of these differences in either signal type is unclear. A disadvantage of the SOM is that human observers can often detect nuanced differences not captured by measurement alone, which is why automatic classification has typically been less accurate (Janik, 1999). This could explain why CART found the SOM and manual techniques to be equivalent for non-contoured units. Manual classification has the advantage of recognizing and addressing these nuanced differences, while SOM has the advantage of being a more repeatable and robust approach.

The methods described here are only applicable to high-quality recordings from which acoustic features can be measured accurately. The subset of recordings measured must also be representative of the dataset under analysis. Additionally, the use of a single singer in each year does not consider individual variations. This represents a limitation of the method as applied to this dataset, and should be taken into account whenever appropriate during use in future studies. Using data that fit the described criteria, acoustic similarity and structure of vocal signals can be quantified for any number of vocal databases. Furthermore, an acoustic dictionary could also be generated for these databases, filling a current gap in the body of knowledge (Placer *et al.*, 2006; Kaufman *et al.*, 2012). This dictionary could then be used as a guide to transcribe sequences in new recordings from the respective population or database. Quantifiable similarity between these prototypical units can enhance the repeatability of the dictionary's application when used in subsequent sequential analyses. While this method by no means eliminates the limitations of the traditional approaches to acoustic signal categorization and analysis, it does provide a key step in the process towards a more quantitative, robust, and repeatable approach.

ACKNOWLEDGEMENTS

The authors would like to thank the staff and volunteers of the Cetacean Ecology and Acoustics Laboratory for data collection, maintenance, and storage over the course of the study. Data were collected during the HARC project for 2002-2004 and 2008-2009, and during the BRAHSS project for 2010-2012 and 2014. HARC was funded by the US Office of Naval Research, the Australian Defense Science and Technology Organization, and the Australian Marine Mammal Center. BRAHSS was funded by the E&P Sound and Marine Life Joint Industry Programme and the US Bureau of Ocean Energy Management. Additional funding was provided to M.J.N. and J.A.A. by the Sea World Research and Rescue Foundation, Inc., and J.A.A. was funded by an Australian Government Research Training Program Scholarship. E.C.G. was funded by a Royal Society Newton International Fellowship. Thanks to Douglas Cato for his significant role in instigating and participating in all the field programs, and to Alycia Rajendran for compilation of study site maps.

***See supplementary material at [URL will be inserted by AIP] for Tables VI and VII, which provide the averages for each of the acoustic feature variables used in the 10x10 contoured SOM (Table VI) and the 7x7 non-contoured SOM (Table VII).**

V. REFERENCES

Bauer, H.-U., and Pawelzik, K. R. (**1992**). "Quantifying the neighborhood preservation of self-organizing feature maps," IEEE Transactions on neural networks **3**, 570-579.

Boisseau, O. (**2005**). "Quantifying the acoustic repertoire of a population: The vocalizations of free-ranging bottlenose dolphins in Fiordland, New Zealand," The Journal of the Acoustical Society of America **117**, 2318-2329.

Breiman, L. (**2001**). "Random forests," Machine learning **45**, 5-32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (**1984**). *Classification and regression trees* (Chapman and Hall, London, UK), pp. 1-359.

Callan, D. E., Kent, R. D., Roy, N., and Tasko, S. M. (**1999**). "Self-organizing map for the classification of normal and disordered female voices," Journal of Speech, Language, and Hearing Research **42**, 355-366.

Catchpole, C. K., and Slater, P. J. (**2008**). *Bird song: biological themes and variations* (Cambridge University Press, Cambridge, UK), pp. 1-348.

Cerchio, S., Jacobsen, J. K., and Norris, T. F. (**2001**). "Temporal and geographical variation in songs of humpback whales, Megaptera novaeangliae: synchronous change in Hawaiian and Mexican breeding assemblages," Animal Behaviour **62**, 313-329.

Céréghino, R., and Park, Y.-S. (**2009**). "Review of the self-organizing map (SOM) approach in water resources: commentary," Environmental Modelling & Software **24**, 945-947.

Charif, R. A., Waack, A. M., and Strickman, L. M. (**2010**). "Raven Pro 1.4 User’s Manual," (Cornell Lab of Ornithology, Ithaca, NY).

Cholewiak, D. M., Sousa-Lima, R. S., and Cerchio, S. (**2013**). "Humpback whale song hierarchical structure: Historical context and discussion of current classification issues," Marine Mammal Science **29**, E312-E332.

Demuth, H. B., Beale, M. H., De Jess, O., and Hagan, M. T. (**2014**). *Neural network design* (Martin Hagan, Boulder, CO, USA), pp. 1-800.

Dunlop, R. A., Noad, M. J., Cato, D. H., and Stokes, D. (**2007**). "The social vocalization repertoire of east Australian migrating humpback whales (Megaptera novaeangliae)," The Journal of the Acoustical Society of America **122**, 2893-2905.

Garland, E. C., Goldizen, A. W., Rekdahl, M. L., Constantine, R., Garrigue, C., Hauser, N. D., Poole, M. M., Robbins, J., and Noad, M. J. (**2011**). "Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale," Current Biology **21**, 687-691.

Garland, E. C., Lilley, M. S., Goldizen, A. W., Rekdahl, M. L., Garrigue, C., and Noad, M. J. (**2012**). "Improved versions of the Levenshtein distance method for comparing sequence information in animals’ vocalisations: tests using humpback whale song," Behaviour **149**, 1413-1441.

Garland, E. C., Rendell, L., Lilley, M. S., Poole, M. M., Allen, J. A., and Noad, M. J. (**2017**). "The devil is in the detail: quantifying vocal variation in a complex, multileveled, and rapidly evolving display.," Journal for the Acoustical Society of America **In Press**.

Gerhardt, H. C. (**2001**). "Acoustic communication in two groups of closely related treefrogs," Advances in the Study of Behavior **30**, 99-167.

Green, S. R., Mercado, E., Pack, A. A., and Herman, L. M. (**2007**). "Characterizing patterns within humpback whale (Megaptera novaeangliae) songs," Aquatic Mammals **33**, 202-213.

Hagan, M. T., Demuth, H. B., Beale, M. H., and De Jesús, O. (**1996**). *Neural network design* (Campus Publishing Service, University of Colorado Bookstore. , Boulder, CO, USA), pp. 1-800.

Helweg, D. A., Cato, D. H., Jenkins, P. F., Garrigue, C., and McCauley, R. D. (**1998**). "Geographic variation in South Pacific humpback whale songs," Behaviour, 1-27.

Janik, V. M. (**1999**). "Pitfalls in the categorization of behaviour: a comparison of dolphin whistle classification methods," Animal Behaviour **57**, 133-143.

Karels, T. J., Bryant, A. A., and Hik, D. S. (**2004**). "Comparison of discriminant function and classification tree analyses for age classification of marmots," Oikos **105**, 575-587.

Kaufman, A. B., Green, S. R., Seitz, A. R., and Burgess, C. (**2012**). "Using a self-organizing map (SOM) and the hyperspace analog to language (HAL) model to identify patterns of syntax and structure in the songs of humpback whales," International Journal of Comparative Psychology **25**, 237-275.

Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., Bohn, K., Cao, Y., Carter, G., and Cäsar, C. (**2014**). "Acoustic sequences in non-human animals: a tutorial review and prospectus," Biological Reviews.

Kohonen, T. (**1990**). "The self-organizing map," Proceedings of the IEEE **78**, 1464-1480.

Kohonen, T. (**2014**). "MATLAB Implementations and Applications of the Self-Organizing Map," Unigrafia Oy, Helsinki.

Liaw, A., and Wiener, M. (**2002**). "Classification and regression by randomForest," R news **2**, 18-22.

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.

Melendez, K. V., Jones, D. L., and Feng, A. S. (**2006**). "Classification of communication signals of the little brown bat," The Journal of the Acoustical Society of America **120**, 1095-1102.

Mellinger, D. K. (**2001**). "Ishmael 1.0 user's guide," NOAA Technical Memorandum OAR PMEL **120**, 98115-96349.

Mercado, E., and Kuh, A. (**1998**). "Classification of humpback whale vocalizations using a self-organizing neural network," in *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on* (IEEE), pp. 1584-1589.

Murray, A., Dunlop, R. A., Goldizen, A. W., and Noad, M. J. (**2016**). "Stereotypy and variability differ between humpback whale (Megaptera novaeangliae) phrase types offering structural support for the hypothesis that song is a mult-message display," Journal for the Acoustical Society of America **140**.

Murray, A., Dunlop, R. A., Goldizen, A. W., and Noad, M. J. (**In Review**). "Stereotypic and complex phrase types provide structural evidence for a multi-message display in humpback whales (Megaptera novaeangliae)," Journal for the Acoustical Society of America.

Noad, M., and Cato, D. (**2001**). "A combined acoustic and visual survey of humpback whales off southeast Queensland," Memoirs of the Queensland Museum **47**, 507-523.

Paterson, R., and Paterson, P. (**1984**). "A study of the past and present status of humpback whales in east Australian waters," Biological conservation **29**, 321-343.

Payne, K., and Payne, R. (**1985**). "Large scale changes over 19 years in songs of humpback whales in Bermuda," Zeitschrift für Tierpsychologie **68**, 89-114.

Payne, K., Tyack, P., and Payne, R. (**1983**). "Progressive changes in the songs of humpback whales (Megaptera novaeangliae): a detailed analysis of two seasons in Hawaii," Communication and behavior of whales, 9-57.

Payne, R. S., and McVay, S. (**1971**). "Songs of humpback whales," Science **173**, 585-597.

Placer, J., Slobodchikoff, C., Burns, J., Placer, J., and Middleton, R. (**2006**). "Using self-organizing maps to recognize acoustic units associated with information content in animal vocalizations," The Journal of the Acoustical Society of America **119**, 3140-3146.

Ranjard, L., and Ross, H. A. (**2008**). "Unsupervised bird song syllable classification using evolving neural networks," The Journal of the Acoustical Society of America **123**, 4358-4368.

Rekdahl, M. L., Dunlop, R. A., Noad, M. J., and Goldizen, A. W. (**2013**). "Temporal stability and change in the social call repertoire of migrating humpback whales," The Journal of the Acoustical Society of America **133**, 1785-1795.

Risch, D., Clark, C. W., Dugan, P. J., Popescu, M., Siebert, U., and Van Parijs, S. M. (**2013**). "Minke whale acoustic behavior and multi-year seasonal and diel vocalization patterns in Massachusetts Bay, USA," Marine Ecology Progress Series **489**, 279-295.

Schön, P.-C., Puppe, B., and Manteuffel, G. (**2001**). "Linear prediction coding analysis and self-organizing feature map as tools to classify stress calls of domestic pigs (Sus scrofa)," The Journal of the Acoustical Society of America **110**, 1425-1431.

Slocombe, K. E., and Zuberbühler, K. (**2006**). "Food-associated calls in chimpanzees: responses to food types or food preferences?," Animal Behaviour **72**, 989-999.

Suzuki, R., Buck, J. R., and Tyack, P. L. (**2006**). "Information entropy of humpback whale songs," Journal of the Acoustical Society of America **119**, 1849.

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., and Mitra, P. P. (**2000**). "A procedure for an automated measurement of song similarity," Animal Behaviour **59**, 1167-1176.

Therneau, T., Atkinson, B., and Ripley, B. (**2014**). "rpart: Recursive Partitioning and Regression Trees. R package version 4.1-8."

Van Opzeeland, I. C., and Van Parijs, S. M. (**2004**). "Individuality in harp seal, Phoca groenlandica, pup vocalizations," Animal behaviour **68**, 1115-1123.

Walker, A., Fisher, R., and Mitsakakis, N. (**1996**). "Singing maps: Classification of whalesong units using a self-organizing feature mapping algorithm," in *Department of Artificial Intelligence* (University of Edinburgh), pp. 1-12.

VI. TABLES

TABLE I. Acoustic features measured for contoured units in *Beluga*.

Acoustic Feature	Definition
Max frequency (Hz)	Highest peak frequency extracted from the frequency contour
Min frequency (Hz)	Lowest peak frequency extracted from the frequency contour
Start frequency (Hz)	The first peak frequency extracted from the frequency contour
End frequency (Hz)	The last peak frequency extracted from the frequency contour
Trend	Start frequency/end frequency. Values >1 indicate a sound that decreases in frequency, while values <1 indicate a sound that increases in frequency
Duration (s)	Length of the unit based on the extracted frequency contour
Bandwidth (Hz)	Maximum frequency – minimum frequency
Inflection	Number of changes in the slope of the frequency contour
Pulse repetition rate (/s)	The number of pulses in sounds that are contoured but have a pulsative element
Contour point (x50) (Hz)	Subsamples of the peak frequency measurements taken every 0.03 seconds to create the frequency contour. 50 samples were taken, evenly spaced along the frequency contour. Each subsample was treated as its own acoustic feature

TABLE II. Acoustic features of non-contoured units measured using robust measurements in *Raven*

Acoustic Feature	Definition
Center frequency (Hz)	Frequency at which the sound is divided into two intervals of equal energy
Peak frequency (Hz)	Frequency at which the sound has maximum amplitude.
Frequency 5% (Hz)	Frequency at which the sound is divided into intervals containing 5% and 95% of its energy
Frequency 95% (Hz)	Frequency at which the sound is divided into intervals containing 95% and 5% of its energy
Duration (s)	Length of the unit based on the spectrogram visualization
Bandwidth 90% (Hz)	Frequency 95% - Frequency 5%
Inflection	Number of changes in the slope of the frequency contour
Pulse repetition rate (/s)	Number of pulses in sounds that have a pulsative element

TABLE III. Classification agreements between method of classification and decision tree analysis (both CART and Random Forest) used to evaluate classification techniques. Root node errors, determined for CART only, represents the percentage of classification of call types. Significantly higher agreements based on Mann-Whitney/Wilcoxon tests are shown in bold.

Corroborating Method	**Unit Types**	**Qualitative Agreement**	**SOM Agreement**
CART	Contoured	57.55% (95.02% root node error)	**73.03%** (95.35% root node error)
CART	Non-contoured	78.97% (93.20% root node error)	74.24% (81.11% root node error)
Random Forest	Contoured	73.01%	**89.21%**
Random Forest	Non-contoured	83.31%	**90.93%**

TABLE IV. Variables used in the CART analyses and mean decrease in Gini index. C-SOM = contoured units classified by SOM, C-Q = contoured units classified by qualitative naming, NC-SOM = non-contoured units classified by SOM, NC-Q = non-contoured units classified by qualitative naming.

CART							
C-SOM		**C-Q**		**NC-SOM**		**NC-Q**	
Variables	**Gini**	**Variables**	**Gini**	**Variables**	**Gini**	**Variables**	**Gini**
Duration	823	Duration	630	Freq. 95%	578	Duration	597
Trend	628	Trend	360	Bandwidth 90%	553	Center	439
Inflection	622	Start	344	Freq. 5%	434	Freq. 95%	413
End	469	Inflection	326	Center	418	Peak	410
Max	443	Contour Point 2	319	Peak	392	Freq. 5%	373

TABLE V. Variables used in the Random Forest analyses and mean decreasing Gini index. C-SOM = contoured units classified by SOM, C-Q = contoured units classified by qualitative naming, NC-SOM = non-contoured units classified by SOM, NC-Q = non-contoured units classified by qualitative naming.

RANDOM FOREST							
C-SOM		**C-Q**		**NC-SOM**		**NC-Q**	
Variables	**Gini**	**Variables**	**Gini**	**Variables**	**Gini**	**Variables**	**Gini**
Duration	805	Duration	821	Bandwidth 90%	330	Duration	416
Inflection	595	Trend	477	Freq. 95%	241	PRR	210
Trend	559	Inflection	342	PRR	224	Peak	197
Max	221	Max	188	Freq. 5%	218	Center	186
PRR	207	Bandwidth	179	Duration	217	Freq. 95%	155