The devil is in the detail: quantifying vocal variation in a complex, multi-levelled, and rapidly evolving display.

Ellen C. Garland[1] and Luke Rendell

School of Biology, University of St Andrews, St Andrews, Fife, KY16 9TH, UK

Matthew S. Lilley

SecuritEase International, Level 8, IBM Tower, 25 Victoria Street, Petone, 5012, New Zealand

M. Michael Poole

Marine Mammal Research Program, BP 698, Maharepa, 98728, Mo'orea, French Polynesia

Jenny Allen and Michael J. Noad

Cetacean Ecology and Acoustics Laboratory, School of Veterinary Science, University of Queensland, Gatton, QLD, 4343, Australia

Running title: Quantifying multi-levelled vocal variation

Keywords: song; sequence; cultural evolution; Levenshtein distance; humpback whale

[1]ellen.garland@gmail.com

ABSTRACT

Identifying and quantifying variation in vocalizations is fundamental to advancing our understanding of processes such as speciation, sexual selection, and cultural evolution. The song of the humpback whale (*Megaptera novaeangliae*) presents an extreme example of complexity and cultural evolution. It is a long, hierarchically structured vocal display that undergoes constant evolutionary change. Obtaining robust metrics to quantify song variation at multiple scales (from a sound through to population variation across the seascape) is a substantial challenge. Here, we present a method to quantify song similarity at multiple levels within the hierarchy. To incorporate the complexity of these multiple levels, the calculation of similarity is weighted by measurements of sound units (lower levels within the display) to bridge the gap in information between upper and lower levels. Results demonstrate that the inclusion of weighting provides a more realistic and robust representation of song similarity at multiple levels within the display. Our method permits robust quantification of cultural patterns and processes that will also contribute to the conservation management of endangered humpback whale populations, and is applicable to any hierarchically structured signal sequence.

PACS number(s): 43.80.Ka, 43.80.Ev

I. INTRODUCTION

Identifying and quantifying variation in vocalizations is fundamental to advancing our understanding of processes such as speciation (Riesch *et al.*, 2012), sexual selection (Catchpole & Slater, 2008), and cultural evolution (Rendell & Whitehead, 2001; Noad *et al.*, 2000; Janik, 2014). For example, variations in the group-specific calls of killer whales (*Orcinus orca*) are believed to be leading to speciation (Riesch *et al.*, 2012) potentially through culture-genome coevolution (Foote *et al.*, 2016), while the vocal displays of song birds are driven by both sexual selection and cultural evolution (Catchpole & Slater, 2008). Understanding variation within and between habitats can also support conservation and management by revealing details of population structure. Therefore, robust metrics to quantify vocal variation at multiple scales (from single utterances through to variation across the land and seascape) are essential to address what defines a "dialect", how dialects may correspond to populations, and how this information is incorporated into the management of populations or species.

Substantial research has been conducted at comparing the population repertoires of many species, including our own, to identify and quantify dialect variation (*e.g.,* human language: Wieling & Nerbonne, 2015; bird song: Catchpole & Slater, 2008; whale song: Payne and Guinee, 1983; rock hyrax, *Procavia capensis*: Kershenbaum *et al.*, 2012). Studies on non-human animals typically compare call types, and how the parameters of each call and frequencies with which they are used vary geographically. This can become complicated when vocalizations are grouped together into bouts or displays. Songbird dialects are a well-established means of defining groupings (Catchpole & Slater, 2008). Dialects are defined as song differences between neighboring populations of potentially interbreeding individuals (Connor, 1982). Bird songs typically last for a few seconds and are composed of a few to tens of syllables. In contrast, humpback whale (*Megaptera novaeangliae*) songs can last in excess of 20 minutes and commonly comprise thousands of units (individual sounds). This male-only vocal display is long, complex, and highly stereotyped (Payne and McVay, 1971).

Humpback whale song is divided into multiple levels that are stacked on top of each other (*i.e.,* it is a nested hierarchy; Payne and McVay, 1971; Herman and Tavolga, 1980). The shortest, continuous sound to our ear is called a 'unit' (Payne and McVay, 1971; Fig. 1[i]). Several units are arranged in a stereotyped sequence that is termed a 'phrase'. A phrase is repeated multiple times and this is called a 'theme'. A few different themes, each comprised of repeats of a different stereotyped phrase, are sung in a particular order to make a 'song'. Songs are repeated multiple times by an individual whale to comprise a 'song session'. Different versions of the song (comprised of different themes and phrases) are termed 'song types' (Garland *et al.*, 2011). For context, humpback whale phrases and bird songs are considered analogous (see Cholewiak *et al.*, 2012). There is a clear challenge in incorporating all of this variation into a quantitative analysis that includes as much information as possible without abstracting from the data.

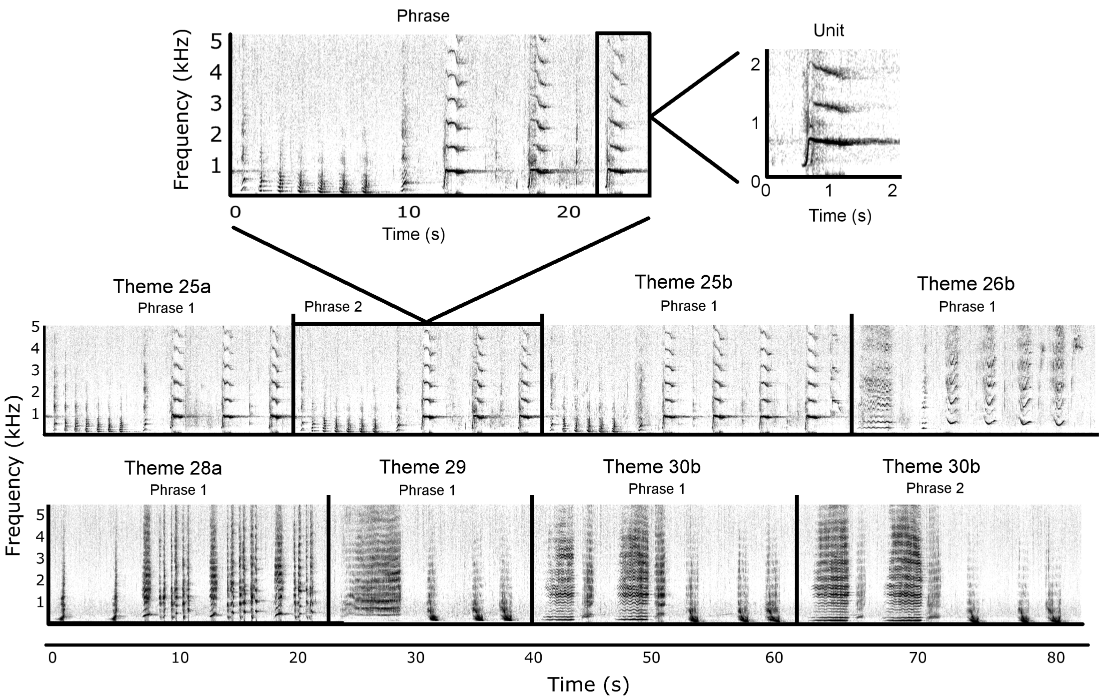

FIG. 1. Spectrograms illustrating the hierarchical structure of humpback whale song. A single unit (‘trumpet’) and a single phrase from Theme 25a are shown in the top panel. Theme 25a units from the single phrase in the top panel are as follows: short ascending moan, grunt, grunt, grunt, grunt, grunt, grunt, short ascending moan, trumpet, squeak, trumpet, squeak, trumpet. The repetition of phrases and the sequential singing of themes are shown in each of the subsequent panels (corresponding audio: SuppPubmm1.wav). Spectrograms were 2048 point fast Fourier transform (FFT), Hann window, 31 Hz resolution, and 75% overlap, generated in Raven Pro 1.4.

Within a population, most males conform to the current arrangement and content of the song (Winn and Winn, 1978; Payne *et al.*, 1983). The song progressively evolves through time (Payne and Payne, 1985), with all males incorporating these changes to maintain the observed similarity. Across an ocean basin, populations that are geographically closer to each other display a higher degree of song similarity (Payne and Guinee, 1983; Helweg *et al.*, 1990, 1998; Cerchio *et al.*, 2001). However, song sharing within the western and central South Pacific is very dynamic as songs can be directionally transmitted eastward across the region from eastern Australia to French Polynesia, usually over a period of two years (Garland *et al.*, 2011, 2013). The underlying drivers for this unidirectionality in song transmission are not well understood, but have been suggested to be a result of differences in population sizes within the region (Garland *et al.*, 2011). Despite this transmission of different versions of the display across the region, it is possible to use differences in the song to identify different dialects and also populations at any point in time (Garland *et al.*, 2015). Songs and the stereotyped sequences of units therein are used to define geographic dialects (Payne and Guinee, 1983; Garland *et al.*, 2015). Since variation can occur at all levels of the song structure, it is a substantial analysis challenge to incorporate variation at all these levels into a single metric.

Many studies have undertaken quantification of humpback whale sounds (units) to allow comparison, typically involving the measurement of time and frequency parameters (*e.g.,* Dunlop *et al.*, 2007; Stimpert *et al.*, 2011; Rekdahl *et al.*, 2013). Previous work has also compared multiple metrics to establish which of a variety of commonly employed sequence analysis techniques performs best for comparing humpback whale song (Kershenbaum and Garland, 2015). The string edit or Levenshtein distance (LD) metric outperformed all other metrics in comparing humpback whale song sequences. The LD is a robust metric that should be employed in the comparison of song in preference to other commonly utilized techniques (such as Markov chains, hidden Markov models or Shannon entropy). The LD is a basic technique in computer science and information theory which has been used in genetics for analyzing the sequence of nucleotides in DNA (*e.g.,* Altschul *et al.*, 1990) and has also found favour in linguistics (*e.g.*, Wieling and Nerbonne, 2015) and animal bioacoustics (*e.g.,* Margoliash *et al.,* 1991; Kershenbaum *et al.*, 2012). More advanced applications of the LD have been undertaken to investigate bird song dialects (e.g., Ranjard and Ross, 2007, 2008) and language relatedness (see Wieling and Nerbonne, 2015), where the cost of substitution was reduced based on the proportional similarity of acoustic features or phonetic similarity. The LD has also previously been used to quantify song similarity in humpback whales (Helweg *et al.*, 1998; Eriksen *et al.*, 2005; Tougaard and Eriksen, 2006; Garland *et al.*, 2012, 2013, 2015). These studies have compared song similarity among individuals and populations in the South Pacific to understand dialect grouping; however, none have employed a weighting system to better represent the complexities in song structure.

Here, we present a straightforward LD-based analysis method to quantify stereotyped sequences of sounds that vary geographically (*i.e.,* song dialects) at multiple levels within the display. To incorporate the complexity of these multiple levels, the calculation is weighted by sound unit measurements taken from lower levels within the display. We use humpback whale song as an example due to its inherent complexity and constant evolution. Instead of qualitatively judging unit similarity as is commonly undertaken, the quantitative level of similarity as calculated using a suite of variables taken directly from each unit type is an important step towards a robust, reportable and repeatable quantification of humpback whale song.

II. METHODS

A. Calculating the Levenshtein distance (LD) and its derivatives

Both the conceptual understanding of the LD and its calculation is straightforward. The LD measures the similarity between any two strings (sequences) of data by calculating the minimum number of changes (insertions, deletions and substitutions) needed to convert one string into another (Levenshtein, 1966; Kohonen, 1985). The Levenshtein distance (LD) is calculated by:

$$LD(a, b) = \min(i + d + s) \quad (1)$$

where string (*a*) is converted into string (*b*) by the minimum number of insertions (*i*), deletions (*d*) and substitutions (*s*). To ensure the output is comparable to more then a single pair of strings, the LD is standardised by the length of the longest string within the pair to give the Levenshtein distance similarity index (LSI), defined as:

$$LSI(a, b) = 1 - \frac{LD(a,b)}{\max(len(a), len(b))} \quad (2)$$

where the LD between strings *a* and *b* is divided by the length of the longer string of the pair (see Garland *et al.*, 2012, 2013). This produces a measure of similarity among multiple sequences of varying lengths, and an overall understanding of the similarity of all sequences (Helweg *et al.*, 1998; Eriksen *et al.*, 2005; Tougaard and Eriksen, 2006; Garland *et al.*, 2012, 2013, 2015).

Within any set of sequences, a median, or most representative sequence, for that set can be calculated. Examples of a set (or group) include all of the songs from a population, all songs from a population in a particular year, repeated songs from an individual, or all examples of a particular theme from all individuals within a population. The string with the highest overall similarity to all other strings within the group or set is found by summing all LSI scores per string. The string or sequence with the highest summed LSI and thus highest similarity to all other members within the group is then assigned as the 'set median string' (Kohonen, 1985). This provides a representative string for the set that can then be used to compare among sets without losing substantial amounts of information.

As noted in Kershenbaum and Garland (2015), the LD relies more on the straight sequence of sound units and does not account for any hierarchy in the overall structural pattern. To address this gap we propose a method of weighting changes in higher levels within the song hierarchy using measurements taken directly from lower levels.

B. Calculating weightings

1. Song recordings

Recordings of humpback whale song were made in Mo'orea, French Polynesia in 2005 using a Sony DAT TCD-D100 recorder and a hydrophone designed by John and Beverly Ford of Vancouver, Canada (recorded digitally but then transferred to computer by digital to analog conversion followed by re-digitizing at 44.1 kHz and 16 bit). Two different song types (Blue and Dark Red) were identified in the recordings based on previously described songs (Garland *et al.*, 2011, 2012, 2013). Given that songs are constantly evolving through changes in the arrangement and content of phrases and themes (Payne and Payne, 1985), and these differences can then be transmitted to another population (Noad *et al.*, 2000; Garland *et al.*, 2011), identifying differences between song types is essential to identify the underlying dynamics and track dynamic dialect boundaries.

2. Unit measurements

Units, the shortest continuous sound to our ear delineated by silence (Payne and McVay, 1971), were initially categorized into sound types by a human classifier (E.C.G.; following Dunlop *et al.,* 2007 classification system) as is common in humpback whale studies (see Cholewiak *et al.*, 2012; Fig. 1). Units were named as they sound (*e.g.,* moan, groan, squeak) and included information on the slope (*e.g.,* ascending, modulated) and length of the call (*e.g.,* short, long). This resulted in a fine-scale classification of units instead of large, variable unit categories (for example the unit category 'purr' could be further subdivided into 'long purr' or 'short purr' based on length). All units were coded for each recording. As a single song can contain upwards of 1,000 units, a subset of units from each recording is measured. All units in the first, full phrase of each theme in the recording were measured to provide a variety of units from different themes in the song, and from different individuals for comparison. This resulted in 750 measured units, a set containing multiple examples of 96 unique unit types. All measured units were taken from a subset (described above) of the 636 available phrases. Units were measured in Raven Pro 1.4 for 11 frequency and duration variables (Table I) following those outlined in Dunlop *et al.* (2007). These measurements were taken from a spectrogram made with a 2048 point fast Fourier transform (FFT), Hann window, 16 bit, 31 Hz resolution, and 75% overlap. In R (R Development Core Team, 2015), this subset of measured units (N=750, 96 unit types) was subjected to both Classification And Regression Tree analysis (CART) and Random Forest classification. Of the 96 unit types classified by CART and Random Forest, 77% and 73% (respectively) were classified in the same way by the human classifier, inferring repeatability in the naming of units. Therefore, all 636 phrases (which included both the qualitatively assigned units and the 750 measured units) were included in further analysis.

3. Turning unit measurements into a weighting system

To create a weighting cost or penalty between every pair of unit types (*e.g.,* a moan or a whoop) based on the distance among units to allow a quantification of similarity, the mean of each variable (*e.g.,* maximum frequency) for each unit type was calculated. These were taken from the 750 measured units. The mean unit type values for each variable were then transformed into z-scores to ensure all the variables were comparable on the same scale. Given that we do not currently know what sound features are most important to humpback whales, all variables were included in the analysis in preference to reducing these to a small number of factors (*e.g.,* through Principal Components Analysis). The Euclidian distance was computed for all unit types creating a single measure of distance between each pair of unit types in *n*-dimensional acoustic feature space (here, *n*=11 as there were 11 variables). The Euclidian distance was normalized to the maximum pairwise distance (*i.e.,* linearly) to represent a value between 0 and 1, where 1 represented the largest distance (or highest dissimilarity) between unit types in *n*-dimensional space. The linear normalized cost *d(x,y)* is simply the Euclidian distance between the z-scores of units x_i and y_i, divided by the maximum value of *d*:

$$d(x,y) = \frac{\sqrt{\sum_i (z(x_i) - z(y_i))^2}}{\max(d)} \quad (3)$$

This linear normalized Euclidian distance between every unit type was used as a weighting penalty for substitutions in subsequent LD calculations (Fig. 2). However, preliminary tests indicated a linear scale was inadequate at capturing the differences among units as the majority of penalty scores were aggregated at one end of the scale due to a small number of very different units (Fig. 3).

To account for this, a non-linear transformation that compressed the range of Euclidian distances that represent the most variation in the normalised scale was undertaken. An exponential scale was able to capture the small but important differences among very similar units, while also ensuring a high penalty score for the very different units (Fig. 3). The exponential normalized cost is given by:

$$exp\_cost(x,y) = 1 - e^{-\beta d(x,y)} \tag{4}$$

where β is the exponential coefficient. The exponential coefficient β could be altered to relax the penalty slope, which resulted in a reduction in the cost for substitution (Fig. 2 & 3). All initial weighting tests were run at $\beta = 1$, and then the coefficient was reduced to $\beta = 0.5$ and $\beta = 0.25$ to allow the effects of weighting to be explored (Fig. 2). $\beta = 1$ represents the closest distribution of penalty scores to the un-weighted analysis (with all scores = 1), while the relaxing of the slope to $\beta = 0.5$ and $\beta = 0.25$ pushes the distribution to the left (Fig. 3) into lower penalty scores. A linear distribution represents the other extreme with a large number of very low substitution costs (see Results for the consequences of such a situation). An alternative to our weighting system not explored here would be to use a penalty matrix based on the output of node weights, Euclidian distances, or Cartesian distances from a self-organizing map (SOM; Placer *et al.*, 2006; Green *et al.,* 2011).

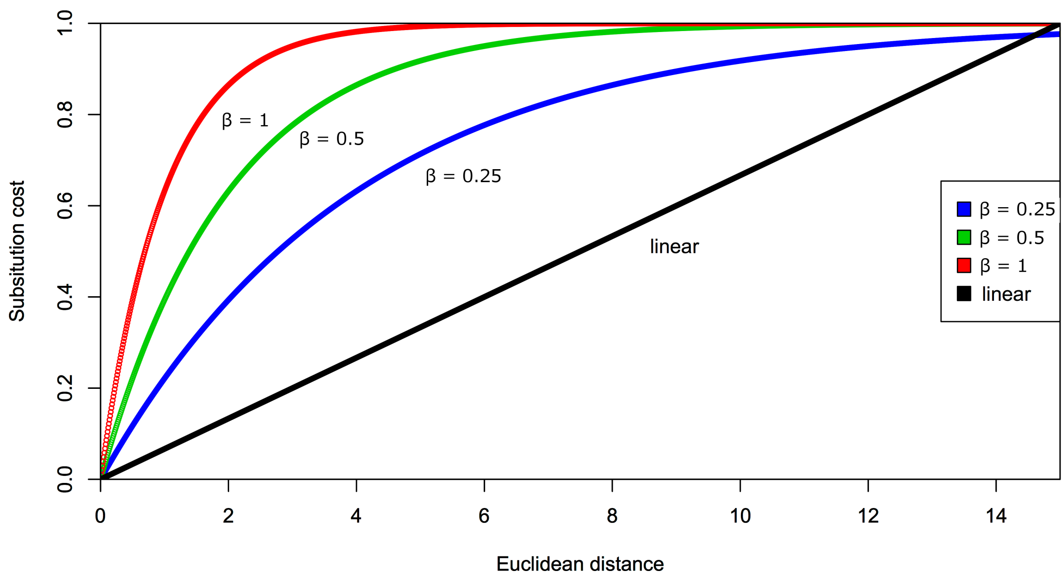

FIG. 2. Substitution costs with different exponential coefficients ($\beta = 1$, $\beta = 0.5$ and $\beta = 0.25$) and linear scaling on the Euclidian distances calculated from sound unit measurements (color online).

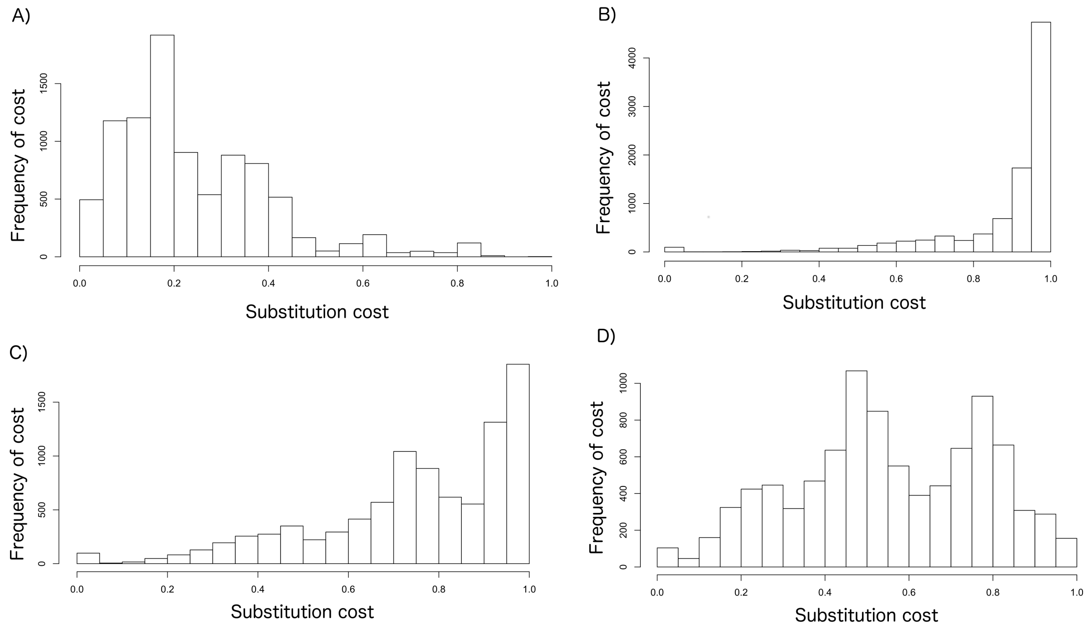

FIG. 3. Histogram of the frequency of normalized substitution costs with A) linear scaling, and exponential coefficients B) $\beta = 1$, C) $\beta = 0.5$, and D) $\beta = 0.25$. Note the difference in the y-axis scale.

C. Applying weightings to better capture hierarchical complexity

The cost of any change (insertion, deletion or substitution) was initially set to 1 (cost of 1 for a change, cost of 0 for no change *i.e.*, *exactly* the same unit in the same position) following the traditional application of the metric. Previous qualitative analyses of song variation have not been so categorical; instead, substituting a unit with a similar unit was considered a less important change relative to substituting it with a less similar unit (Helweg *et al.*, 1998). This is inherently sensible as there are a number of sound units that are indeed very similar. However, the quantitative level of similarity as calculated using a suite of variables taken directly from each unit type is used here instead of qualitatively judging this similarity to move towards a robust, reportable and repeatable quantification of similarity. The penalty or cost of substitution is therefore assigned based on the Euclidian distance between sound units and the exponential coefficient, β. Previous studies have shown that phrase duration is one of the most stable components of humpback whale song (Cholewiak *et al.* 2012). Therefore the cost of insertion or deletion of sounds resulting in the lengthening or shortening of a phrase remains unaltered (cost remains as 1). Insertions and deletions are therefore more heavily penalized than substitutions in this framework.

D. Tests using humpback whale song sequences

Three different analyses were undertaken to demonstrate the utility of this weighted analysis in capturing the inherent multi-levelled structure and complexity within the display. These can be viewed as the major steps in song quantification from lower to upper levels. In each analysis, the strings used for calculating the LSI represent different levels in the hierarchical song structure:

A. Assigning a sequence of units to a known phrase and by extension a theme. In this analysis, a string represents a sequence of units.
B. Identifying a median unit sequence per phrase/theme. Here, a string also represents a sequence of units.
C. Assigning a song to a song type based on the sequence of phrases (as quantified from analyses A and B). In this final analysis, a string represents a sequence of phrases.

The upper level of analysis (C.) of assigning songs to song types is run solely un-weighted in this instance. Weightings could be utilized to trace evolving themes (none are present in the current dataset; Garland *et al.*, 2011) by including the LSI dissimilarity score for those particular themes as the penalty score. The analysis was run in R (R Development Core Team, 2015) utilizing custom written code (available at https://github.com/ellengarland/leven). The code calculates the LSI similarity matrix, creates median strings per group (as specified by the user; see below), calculates the average LSI score within and between groups to investigate average similarity and also within theme variability, and calls the *hclust*, *pvclust* and *pvrect* packages (see Suzuki and Shimodaira, 2004) to cluster strings and calculate bootstrap errors. Examples of a group include all of the songs from a population, all songs from a population in a particular year, repeated songs from an individual, or all examples of a particular theme from all individuals within a population. The percentage theme similarity function calculates the average LSI similarity of all strings within a group (*e.g.,* population, individual, theme, etc.) to provide an understanding of the variability in similarity within that group. This is also calculated among groups; pairwise LSI scores calculated between all strings from two groups are averaged to find the average % theme similarity between those particular groups. This complements the single LSI score calculated between set medians from each group. Clustering was conducted using either single or average-linkage (UPGMA) clustering. Each cluster matrix was bootstrapped with multi-scale bootstrap resampling (AU) and normal bootstrap probability (BP) 1,000 times to establish p-values (significance for AU at $p > 95\%$ and for BP at $p > 70\%$) and SE for each split in the tree (see Garland *et al.*, 2012 for detailed methods). Branches with high AU and BP values are strongly supported by the data while lower values suggest variability in their division. As a further test of how well a dendrogram represented the data, the Cophenetic Correlation Coefficient (CCC) was calculated. A CCC score of over 0.8 is considered high and thus a good representation of the associations within the data (Sokal and Rohlf, 1962).

III. RESULTS

From 19 recordings containing three hours and 24 minutes of song, a total of 636 phrases (*i.e.,* a sequence of individual sound units) were transcribed. Similar phrases were qualitatively assigned to themes and song types for ease of understanding (following previous analyses that qualitatively matched themes and/or assigned song types using un-weighted LSI analyses; Garland *et al.*, 2011, 2012). Sixteen themes were identified; the Blue song type (Table II) contained nine themes (labelled 23 to 30b) with 212 phrases, and the Dark Red song type contained seven themes (labelled 31a to 37b) with 424 phrases. Previous qualitative assignment of these themes (presented in Garland *et al.*, 2011) provides a direct comparison of this quantitative method to naïve matching tests.

A. Assigning a sequence of units to a phrase and, by extension, a theme

The aim of this test was to assign multiple strings of units to a phrase (and therefore a theme, which represents the repetition of a stereotyped set of similar phrases). The clustering of phrases into themes using both un-weighted and weighted analyses was conducted for all themes for both the Blue and Dark Red song types (data not shown), with similar results to those reported below. To demonstrate this, three themes were chosen from the Blue song type to ensure a complex task that could also be visually presented without requiring a magnifying glass. All strings from each of the chosen themes were included in the analysis (N=72 phrases). Theme 28a (N=19 phrases) was a long phrase that contained between nine and 20 units, made up of a possible 11 unique unit types (Table III). The length of a 28a phrase depended on the number of repetitions of a sub-phrase (a sequence of one or more units that is sometimes repeated in a series; Cholewiak *et al.* 2012) comprising the 'ascending moan' and 'violin' units (see Table III). Theme 30b (N=33 phrases) was shorter then Theme 28a with between four and seven units, and was made up of six possible unit types (Table III). None of the unit types were shared between the two themes. Theme 25a (N=20 phrases) contained between 11 and 20 units, and was made up of seven possible unit types (Table III). The length of a 25a phrase primarily depended on the number of 'grunts' (*gt*; a short, low frequency unit that was repeated multiple times) sung in the first sub-phrase, and whether this first sub-phrase was itself repeated (Table III). Theme 25a and Theme 28a shared two unit types (*ba*: 'bark', and *sq*: 'squeak'), while a number of other units were very similar in their acoustic features (*i.e.,* frequency and duration measures). However, these themes are clearly different in the arrangement of their units (see Fig. 1), and the selection of these two themes was intentional in an attempt to confuse and identify shortcomings in the weighted analysis.

When the analysis was run un-weighted (*i.e.,* every substitution cost=1), bootstrapping indicated three general clusters corresponding to the three themes (Fig. 4a). The CCC of 0.974 indicated a very good representation of the associations within the data, despite some of the branches in the tree not reaching AU or BP significance. The average % similarity between Themes 25a and 28a was 4%, with 0% similarity between either of these themes and 30b. The analysis was then run as a weighted analysis with $\beta = 1$. Average-linkage hierarchical clustering and bootstrapping indicated two major branches and four general clusters were present (Fig. 4b), and the dendrogram was again a very good representation of the data (CCC=0.982). The average % similarity between Themes 25a and 28a rose to 33%, with similarity between either of these themes and Theme 30b ranging from 4 to 6%. The weighting allowed similar units to be less costly for substitution. Two clusters within the left branch (Fig. 4b) were present after bootstrapping and clustering of the weighted data, as Themes 25a and 28a were subdivided at a higher level of similarity than 30b. This relates to the length of strings as the LD attempts to find the *minimum* number of changes (which is weighted towards less costly substitutions). Theme 30b contained two versions based on length and thus two clusters within the overall theme: a single (short) or repeated (long) 'groan' and 'purr'. Given this variation is permitted and considered the *same* Theme in qualitative assessment, this provides a guide for understanding the impact of length on weighting. Alternatively, it may indicate that Theme 30b should be split into two finer-scale groupings based on length (*i.e.*, 30b short and 30b long).

To understand the overall variability in sequences within a phrase/theme, the average similarity score to all other strings within the theme set was calculated (Table II, % Theme similarity). While visually the difference introduced by weighting ($\beta = 1$) is subtle among these three themes, weighting has a profound effect on stabilising and reducing variability within a theme. This is best seen in the increase in within theme similarity for each theme (Table II, column 5). The difference between un-weighted and weighted ($\beta = 1$) analyses was clear. Theme 25a increased in similarity to itself (from 73% to 79%), as did Theme 28a (from 60% to 70%) and Theme 30b (from 44% to 53%) from un-weighted to weighted analyses, respectively. For example, the cost of substituting between two units, a 'bark' (*ba*) and a 'long bark' (*lb*), was significantly reduced from cost = 1 (un-weighted analysis) to cost = 0.506 in the weighted analysis ($\beta = 1$), as a long bark represents a longer duration version of a bark (> 1 sec). There is a trade-off, however, between reducing variability within a theme and increasing the similarity among themes.

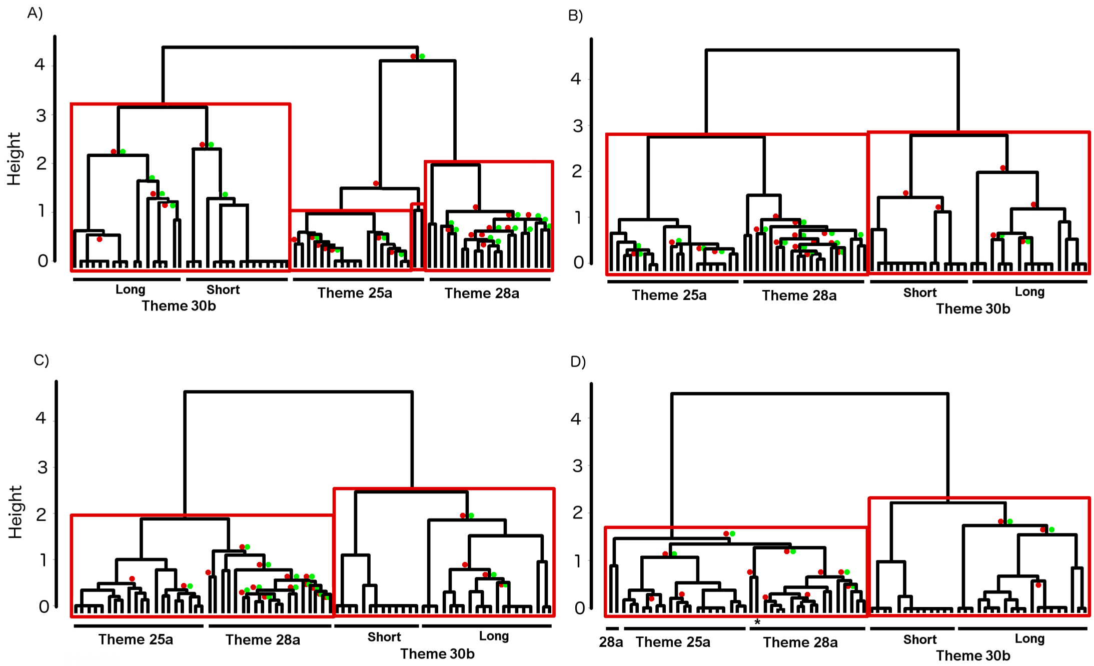

FIG. 4. Dendrograms of bootstrapped (1000) LSI average-linkage hierarchical clustered individual unit strings from Themes 25a, 28a and 30b (N=72) for A) un-weighted, B) $\beta = 1$, C) $\beta = 0.5$, and D) $\beta = 0.25$ analyses. Where multi-scale bootstrap resampling (AU; left, red •) p-values and normal bootstrap probability (BP; right, green •) p-values did not meet significance ($p<0.95$, $p<0.7$, respectively), these are displayed (color online). Red boxes indicate clusters that are strongly supported by the data. Theme 30b is split into two versions: 'Long' had four starting units, while 'short' contained two starting units. Note the confusion of Theme 25a and 28a in D (*) indicating the process of relaxing the coefficient value has gone too far.

To further explore the impact of weighting and this trade-off, the exponential coefficient was relaxed from $\beta = 1$ to $\beta = 0.5$ and $\beta = 0.25$. This reduces the steepness and relaxes the penalty slope, drawing similar units closer together (Fig. 2 & 3). For example, substituting from a bark to a long bark had an initial penalty of 0.506 when $\beta = 1$. This decreased to a penalty of 0.297 for $\beta = 0.5$, and to 0.162 when $\beta = 0.25$.

This resulted in all themes increasing their self-similarity at each change in scale (Table II). For example, Theme 30b increased its within theme similarity to 64% at β = 0.25 (from 53% at β = 1, and 59% at β = 0.5). Relaxing the slope continues to reduce the penalty of substitution. However, there is an obvious limit to relaxing the penalty for substitution as a threshold was reached in this case where similarity in phrase length overrode content of the phrase. It was less costly to substitute all units then undertake any insertion or deletion operations. Using the bark/long bark example above, a substitution penalty of 0.162 may allow up to six substitution operations being equivalent to one insertion operation (insertion penalty cost=1). This threshold was reached at β = 0.25; phrases from Theme 25a and 28a start to be mixed together in a single cluster at this level of weighting (Fig. 4d). To balance the trade-off between reducing within-theme variability and increasing among theme similarity in the current study, the majority of substitution penalty scores needed be above 0.6 (*i.e.,* Fig. 3b & 3c) to ensure a small number of very similar sounds could be substituted while the majority of sounds were costly. Investigating the distribution of penalty scores (Fig. 3) allowed a visualization of the potential skew in distribution that was particularly exacerbated by linear scaling (where there were a high number of extremely low [<0.2] penalty scores).

B. Assigning a median unit sequence (set median) per phrase

Utilising all Blue song strings (N = 212 phrases, each containing a string of units), the most representative unit sequence (string) for each theme was identified with and without weighting. This became the set median for each theme as this string had the highest summed % similarity of all strings within the theme (Table II). As analyses were run four times (*i.e.,* un-weighted, $\beta = 1$, $\beta = 0.5$, and $\beta = 0.25$), four set medians were calculated for each theme.

The analysis was first run un-weighted to provide the initial set medians, followed by weighted analyses. This provides a distinction between changes in set medians arising as a result of weighting (un-weighted *vs.* weighted), or as a result of changing the level of beta coefficient (*e.g.,* $\beta = 1$ *vs.* $\beta = 0.5$). Within a theme, including weighting ($\beta = 1$) resulted in a single set median string changing arrangement from the un-weighted set median: Theme 27 (Table II). This theme had the highest sample size (N=79), and it was also particularly variable in unit choice. Weighting allowed similar units (*i.e.,* 'ascending' and 'n-shaped trills', *ti(a)* and *ti(n)*) to be substituted with a reduced penalty. Therefore, the similarity within the theme increased by 19%, from 42% to 61%.

As above, the exponential coefficient was relaxed from $\beta = 1$ to $\beta = 0.5$ and $\beta = 0.25$ to explore the impact of weighting on set median string assignment. Weighting at $\beta = 0.5$ resulted in two additional themes, Themes 26b and 30b, changing their set medians (Table II). Both themes were lengthened by two units, instead of being represented by the more condensed version of the theme. Theme 27 did not change its set median sequence from $\beta = 1$ to $\beta = 0.5$ (Table II). Themes 30b and 26b had the second and third largest sample sizes in the study, respectively. When $\beta = 0.25$, Theme 25a included a sixth grunt (*gt*) in its set median, and increased its within theme similarity to 82% (from 81% at $\beta = 0.5$; Table II). Once a set median changed through weighting, it remained in the new form as the exponential coefficient was further relaxed.

Cluster analysis of the un-weighted set median sequences indicated the similarity in arrangement among themes (Fig. 5a). Including weighting in the analysis (*β* = 1, *β* = 0.5 and *β* = 0.25; Fig. 5b-d) increased the similarity among themes, as it was less costly to substitute between phrases of a similar length.

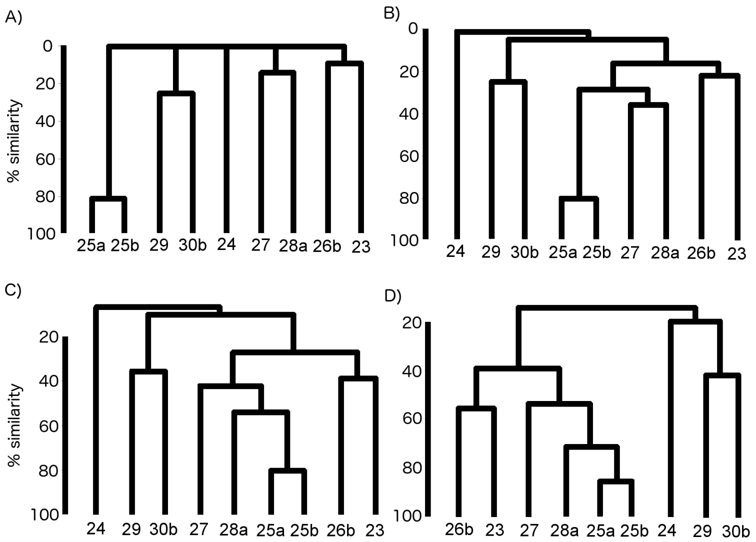

FIG. 5. Dendrograms of bootstrapped (1000) LSI similarity average-linkage hierarchical clustered set medians for Blue song themes for A) un-weighted, B) *β* = 1, C) *β* = 0.5, and D) *β* = 0.25 analyses.

C. Assigning a song to a song type based on the sequence of phrases

The above analyses grouped similar strings of units together to represent a theme. These theme groupings can themselves be assessed at the next level in the hierarchy: assigning songs to song types. This top level in the analysis was run un-weighted. From 18 strings of phrases (including all of the phrase repetitions, *e.g.,* 27, 27, 27, 27, 28a, 28a, 28a, 29, *etc.*) that ranged in length from four to 134 phrases, two significant clusters were formed (Fig. 6). These corresponded to the two different song types, Blue and Dark Red, identified in the data (and previously classified using un-weighted LSI of theme sequences in Garland *et al.*, 2012, 2013). The CCC for the resulting un-weighted average-linkage dendrogram was 0.892, indicating a good representation of the structure within the data despite some branches not reaching AU or BP significance.

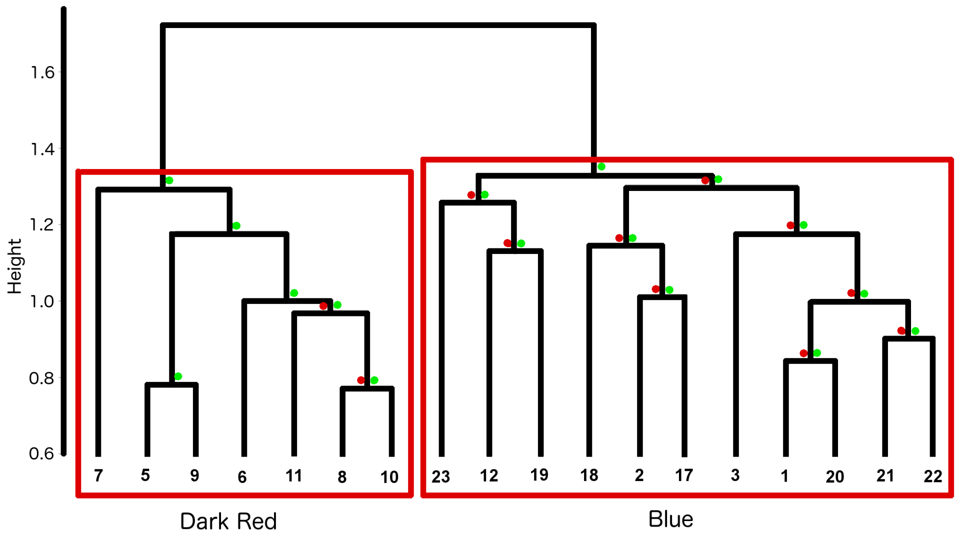

FIG. 6. Dendrogram of bootstrapped (1000) LSI average-linkage hierarchical clustered strings of phrases (*i.e.,* a song) from all recordings. Terminal node numbers refer to recording number. The two clusters correspond to the two different song types, Dark Red and Blue. Where multi-scale bootstrap resampling (AU; left, red •) p-values and normal bootstrap probability (BP; right, green •) p-values did not meet significance (p<0.95, p<0.7, respectively), these are displayed (color online).

IV. DISCUSSION

Here, we have shown how weighting unit substitutions when calculating sequence similarities can better represent the biological reality that some sound units are more similar than others in the quantitative analysis of humpback whale song. We did this by incorporating direct acoustic measurements from lower levels in the song hierarchy into sequence similarity calculations focussed on upper levels. There is no perfect solution to such analytical challenges and no weighting scheme that will be optimal in all situations, but this does nonetheless represent a step forward by reducing the abstract nature of sequence comparisons relative to the empirical system under study. We suggest that researchers think carefully about the research question at hand before employing a weighting scheme. Each of the three analytical tests was affected differently when weighted, resulting in varying levels of ‘success’. Given the extensive previous quantification of these two song types and themes by a number of different researchers (Miksis-Olds *et al.*, 2008; Smith *et al.*, 2008; Garland *et al.*, 2011, 2012, 2013, 2015; Rekdahl *et al.*, 2013), we considered ‘success’ in this context as agreement with those previous studies – but such studies will not be available in most cases. Below we review the impact of weighting on each analysis and outline some potential implications and avenues for improvement.

A. Assigning a sequence of units to a phrase and, by extension, a theme

The clustering of the un-weighted unit sequences mirrored the previous qualitative assignment of unit sequences to phrases/themes. When weighting was applied, however, clustering was more defined at a higher level before reaching a tipping point where different themes were merged together. Weighting will favor substitution (with a cost of <1) over insertion or deletion (both cost 1), as the LD algorithm strives to find the lowest cost to turn string one into string two. Therefore, phrases of similar length are artificially going to be considered closer together (as was evident between Themes 25a and 28a). The inclusion of two themes that were closely aligned in length with a suite of potentially similar units was intentional. However, weighting continued to divide these themes into two distinct clusters with no mixing of themes until the coefficient was significantly relaxed ($\beta = 0.25$; Fig. 4d). This corresponded to the majority of substitution costs being below 0.6 (Fig. 3), indicating a tipping point where length may override theme content. The different location and arrangement of themes in the song should guide the researcher in interpreting this structure in the context of the research question at hand.

Utilizing all strings from the Blue song type, weighting resulted in clear groupings of strings into phrases and themes. While un-weighted analyses do represent the structure of song and should always be undertaken in the first instance, weighting provides a quantitative way of making and reporting decisions about 'similar units in similar locations' to differentiate between themes more subtlety.

Here, we have not binned the substitution costs (*e.g.,* 0.25 to 0.5 = cost 0.5) or included a cut-off value within the cost matrix where the cost will automatically change to 1. One could modify our approach by deciding that any calculated Euclidian distance cost above 0.25 or 0.5, for example, represented a very different suite of sounds, and thus should have a penalty of 1. Alternative cost matrices generated from other analyses, such as output Euclidian or Cartesian distances among nodes from a self-organising map (SOM), could also provide a representative cost matrix if sound types were assigned using the SOM.

B. Assigning a median unit sequence (set median) per phrase

The utility of weighting is clear in this task. Here we are moving from assigning unit strings (phrases) to a theme, to finding the most representative unit string for the theme. If all strings are not going to be included in upper level analyses, this data-condensing task to find their representative is extremely important. Weighting significantly increased the average within theme % of similarity, as highly similar units (*e.g.,* bark vs. long bark) could be better incorporated into the analysis. This results in the analysis treating the barks as longer or shorter duration versions of another similar sound type, rather than simply as separate novel types of sound. As β

decreased, no set median string reverted back to the un-weighted set median. There was an interaction with sample size (N) as larger sample sizes in terms of number of strings, and more variable themes (*i.e.*, 27) switched to a new set median first, followed by themes with a moderate sample size. This indicates that larger sample sizes allow the underlying variability in arrangement to be captured and longer phrases allow for more variability in unit sequences, and both provide more options for set medians. Increasing within-theme similarity to reduce this variability is desirable.

As β was decreased, set medians increased in length (Table II, Themes 25a, 26b and 30b). Weighting appears to better incorporate both the ability to quantify similar units and differences in length. However, the increase in unit similarity (through relaxing β) also resulted in the 'incorrect' placement of phrases into different themes as β passed a tipping point where similarity in phrase length appeared to be more important then similarity in content. This tipping point corresponded to the majority of substitution costs being below 0.6. There was less and less discrimination between units resulting in phrases with the same number of units being hard to differentiate. It became less costly to substitute all units then undertake any insertion or deletion operations. Continuing the bark/long bark example, a substitution penalty of 0.162 may allow up to six substitution operations to equal one insertion operation (insertion penalty cost=1). Therefore, caution and common sense is warranted when applying a weighting system.

One application of this set median analysis is to construct median strings per individual. A researcher can calculate the most representative phrase for each theme (intra-individual), and then these can be put forward into comparisons among individuals to understand any differences in the cultural diversity within a population.

This could be further explored in a way analogous to genetic studies by using AMOVA type techniques (Meirmans, 2012) to compare diversity within and between populations. This could also be used in intra- and inter-group comparisons to quantitatively assign song (dialects).

C. Assigning a song to a song type based on the sequence of phrases

Phrases and themes were labelled using the assignments from lower levels. The sequence or string of phrases could then be compared to assign song types. Here, we utilized the raw sequence of phrases without condensing the repeated phrases down to a single theme label (as in previous work; Garland *et al.*, 2012, 2013). For example, the sequence of phrases 27, 27, 27, 27, 28a, 28a, 28a, 29, 29, 30b, 30b, 30b, and so on, was used instead of removing phrase repeats and condensing the sequence to theme headings (*e.g.,* 27, 28a, 29, 30b, *etc*.). The aim of the exercise was to assign songs to song types, therefore having a variable number of repeats solely impacted the strength of similarity and not the assignment to clusters in this instance (as there were no shared themes). The question at hand should dictate whether phrase repeats should be included or not, as the number of repeats may be impacted by behavioral context (Smith, 2009). The relative strength of similarity within a song type varied due to the number of phrase repeats. There was no impact to the 'correct' assignment of songs to song types.

The LSI calculation at this step was un-weighted; however, a researcher interested in tracing the evolution of a theme through time may assign weightings to different evolutionary stages of a theme based on LSI scores. The utility to trace songs as they naturally evolve through time is extremely desirable. In the current example representing a snapshot in time from a single year, we had no evolving themes but instead had two very different song types.

As very few species rapidly change their songs through time, establishing differences between two different versions of a display (*i.e.,* two 'dialects') was the initial aim of this exercise to allow the technique to be widely applicable. Within a season, differences in humpback whale song types can be used to identify dialect boundaries and populations (Garland *et al.,* 2015). However, the dynamic transmission of song among populations results in a complex task to assign dialect boundaries through time as multiple song types transit a region (see Garland *et al.,* 2015). Weighting of the LD analysis will further assist in clarifying fine-scale differences in songs to assign dialect and population boundaries for conservation measures.

V. CONCLUSIONS

Here we have demonstrated that weighting the LSI analysis better incorporates the variability of unit choice in the song, allowing a suite of similar units to pose little penalty for substitution. The quantification of a previously qualitative process, and the merging of hierarchical levels through weightings from lower levels is an important step towards a robust, reportable and repeatable quantification of humpback whale song. Given that humpback whale song variation among populations can be used to both identify populations and assess connectivity between them (Payne and Guinee, 1983; Helweg *et al.*, 1990, 1998; Cerchio *et al.*, 2001; Garland *et al.*, 2015), having robust metrics to quantify dialect differences is essential. Understanding variation and how this occurs across the seascape also underpins the application of conservation measures to manage populations such as the endangered Oceania (South Pacific) humpback whale subpopulations (Childerhouse *et al.*, 2008), from which these data were sourced. Identifying and quantifying variation in vocalizations is also fundamental to advancing our understanding of processes such as speciation, sexual selection, and cultural evolution.

Humpback whale song presents an extreme example in complexity and cultural evolution. It can serve as a model for complex animal vocalizations; ensuring metrics that incorporate as much information with the least amount of abstraction can only strengthen outcomes. The use of such sequence comparisons and weighting systems using acoustic feature space are nonetheless applicable to other singing species such as bowhead and fin whales, song birds, mice, and hyrax, to name a few. Humpback song shows complete population-wide changes which are replicated in multiple populations at a vast geographical scale (Garland *et al.*, 2011). The level and rate of this cultural transmission remains unparalleled in any other non-human animal. Accurately and quantitatively tracing these changes will help in uncovering the underlying drivers of these processes and thereby contribute to our understanding of animal culture, vocal learning and cultural evolution, and also the roots of human language and culture.

ACKNOWLEDGEMENTS

We thank Emma Carroll for providing valuable comments on a previous version of this manuscript. Song recording in French Polynesia was conducted under permits issued to M.M.P. by the Ministry of the Environment, French Polynesia. E.C.G. was funded by a Royal Society Newton International Fellowship. L.R. was supported by the MASTS pooling initiative (The Marine Alliance for Science and Technology for Scotland) and their support is gratefully acknowledged. MASTS is funded by the

Scottish Funding Council (grant reference HR09011) and contributing institutions. Some funding and logistical support was provided to M.M.P. by the National Oceanic Society (USA), Dolphin & Whale Watching Expeditions (French Polynesia), Vista Press (USA), and the International Fund for Animal Welfare (via the South Pacific Whale Research Consortium).

[i]See supplementary material at [] for audio file (SuppPubmm1.wav) corresponding to Fig. 1.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (**1990**). "Basic local alignment search tool," J. Mol. Biol. **215**, 403–410.

Catchpole, C. K., and Slater, P. J. B. (**2008**). *Bird Song: Biological Themes and Variations*, 2nd ed. (Cambridge University Press, Cambridge, UK), pp. 1–335.

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," Anim. Behav. **62**, 313–329.

Childerhouse, S., Jackson, J., Baker, C. S., Gales, N., Clapham, P. J., and Brownell Jr, R. L. (**2008**). "*Megaptera novaeangliae* (Oceania subpopulation)," IUCN 2012, IUCN Red List of Threatened Species, Version 2012.2. Available from www.iucnredlist.org (accessed April 2016).

Cholewiak, D. M., Sousa-Lima, R. S., and Cerchio, S. (**2012**). "Humpback whale song hierarchical structure: historical context and discussion of current classification issues," Marine Mammal Sci. **29**, E312–E332.

Connor, D. A. (**1982**). "Dialects versus geographic variation in mammalian vocalizations," Anim. Behav. **30**, 297-298.

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*)," J. Acoust. Soc. Am. **122**, 2893–2905.

Eriksen, N., Miller, L. A., Tougaard, J., and Helweg, D. A. (**2005**). "Cultural change in the songs of humpback whales (*Megaptera novaeangliae*) from Tonga," Behaviour **42**, 305–328.

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., Gibbs, R. A., Hanson, M. B., Korneliussen, T. S., Martin, M. D., Robertson, K. M., Sousa, V. C., Vieira, F. G., Vinař, T., Wade, P., Worley, K. C., Excoffier, L., Morin, P. A., Gilbert, M. T. P., and Wolf, J. B. W. (**2016**). "Genome-culture coevolution promotes rapid divergence of killer whale ecotypes," Nat. Commun. **7**, doi:10.1038/ncomms11693.

Garland, E. C., Goldizen, A. W., Rekdahl, M. L., Constantine, R., Garrigue, C., Daeschler Hauser, N., Poole, M. M., Robbins, J., and Noad, M. J. (**2011**). "Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale," Curr. Biol. **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., Noad, M. J., Goldizen, A. W., Lilley, M. S., Rekdahl, M. L., Constantine, R., Garrigue, C., Daeschler Hauser, N., Poole, M. M., and Robbins, J. (**2013**). "Quantifying humpback whale song sequences to understand the dynamics of song exchange at the ocean basin scale," J. Acoust. Soc. Am. **133**, 560–569.

Garland, E. C., Goldizen, A. W., Lilley, M. S., Rekdahl, M. L., Constantine, R., Garrigue, C., Daeschler Hauser, N., Poole, M. M., Robbins, J., and Noad, M. J. (**2015**). "Population structure of humpback whales in the western and central South Pacific Ocean as determined by vocal exchange among populations," Conserv. Biol. **29**, 1198-1207.

Green, S. R., Mercado III, E., Pack, A. A., and Herman. L. M. (**2011**). "Recurring patterns in the songs of humpback whales (*Megaptera novaeangliae*)," Behav. Process. **86**, 284-294.

Helweg, D. A., Herman, L. M., Yamamoto, S., and Forestell, P. H. (**1990**). "Comparison of songs of humpback whales (*Megaptera novaeangliae*) recorded in Japan, Hawaii, and Mexico during the winter of 1989," Sci. Rep. Cetacean. Res. **1**, 1-20.

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 **135**, 1–27.

Herman, L. M., and Tavolga, W. N. (**1980**) "The communication systems of cetaceans," in *Cetacean Behavior: Mechanisms and Functions*, edited by L. M. Herman, John Wiley (New York, USA), pp. 149–209.

Janik, V. M. (**2014**). "Cetacean vocal learning and communication," Curr. Opin. Neurobiol. **28**, 60–65.

Kershenbaum, A., and Garland, E. C. (**2015**). "Quantifying similarity in animal vocal sequences: which metric performs best?," M. Ecol. Evol. **6**, 1452–1461.

Kershenbaum, A., Ilany, A., Blaustein, L., and Geffen, E. (**2012**). "Syntactic structure and geographical dialects in the songs of male rock hyraxes," Proc. R. Soc. B **279**, 2974–2981.

Kohonen, T. (**1985**). "Median strings," Pattern Recogn. Lett. **3**, 309–313.

Levenshtein, V. I. (**1966**). "Binary codes capable of correcting deletions, insertions and reversals," Dokl. Phys. **10**, 707–710.

Margoliash, D., Staicer, C. A. and Inoue, S. A. (**1991**). "Stereotyped and plastic song in adult indigo buntings, *Passerina cyanea*," Anim. Behav. **42**, 367-388.

Meirmans, P. G. (**2012**). "AMOVA-Based Clustering of Population Genetic Data," J. Hered. **103**, 744-750.

Miksis-Olds, J. L., Buck, J. R., Noad, M. J., Cato, D. H., and Stokes, M. D. (**2008**). "Information theory analysis of Australian humpback whale song," J. Acoust. Soc. Am. **124**, 2385–2393.

Noad, M. J., Cato, D. H., Bryden, M. M., Jenner, M.-N., and Jenner, K. C. S. (**2000**). "Cultural revolution in whale songs," Nature (London) **408**, 537.

Payne, K., and Payne, R. (**1985**). "Large-scale changes over 19 years in songs of humpback whales in Bermuda," Z. Tierpsychol. **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," in in *Communication and Behavior of Whales*, edited by R. Payne, AAAS Selected Symposia Series (Westview, Boulder, CO), pp. 9–57.

Payne, R., and Guinee, L. N. (**1983**). "Humpback whale (*Megaptera novaeangliae*) songs as an indicator of 'stocks," in *Communication and Behavior of Whales*, edited by R. Payne, AAAS Selected Symposia Series (Westview, Boulder, CO), pp. 333–358.

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

Placer, J., Slobodchikoff, C. N., Burns, J., Placer, J., and Middleton, R. (**2006**). "Using self-organizing maps to recognize acoustic units associated with information content in animal vocalizations," J. Acoust. Soc. Am. **119**, 3140–3146.

R Development Core Team. (**2015**). "R: a language and environment for statistical Computing," R Foundation for Statistical Computing, Vienna.

Ranjard, L., and Ross, H. A. (**2007**). "A Method for Bird Song Segmentation and Pairwise Distance Measure of Syllables and Songs," Proceedings of the Fourth International Conference on Bio-Acoustics **29**, 185–192.

Ranjard, L., and Ross, H. A. (**2008**). "Unsupervised bird song syllable classification using evolving neural networks," J. Acoust. Soc. Am. **123**, 4358-4368.

Rekdahl, M. R., Dunlop, R. A., Noad, M. J., and Goldizen, A. W. (**2013**). "Temporal stability and change in the social call repertoire of migrating humpback whales," J. Acoust. Soc. Am. **133**, 1785–1795.

Rendell, L., and Whitehead, H. (**2001**). "Culture in whales and dolphins," Behav. Brain Sci. **24**, 309–382, discussion 324–382.

Riesch, R., Barrett-Lennard, L. G., Ellis, G. M., Ford, J. K. B., and Deecke, V. B. (**2012**). "Cultural traditions and the evolution of reproductive isolation: ecological speciation in killer whales?," Biol. J. Linn. Soc. **106**, 1–17.

Smith, J. N. (**2009**). "Song function in humpback whales (*Megaptera novaeangliae*): the use of song in the social interactions of singers on migration," unpublished Ph.D. thesis, The University of Queensland, pp.1-131.

Smith, J. N., Goldizen, A. W., Dunlop, R. A., and Noad, M. J. (**2008**). "Songs of male humpback whales, *Megaptera novaeangliae*, are involved in intersexual interaction," Anim. Behav. **76**, 467-477.

Sokal, R. R., and Rohlf, F. J. (**1962**). "The comparison of dendrograms by objective methods," Taxon **11**, 33-40.

Stimpert, A. K., Au, W. W. L., Parks, S. E., Hurst, T., and Wiley, D. N. (**2011**). "Common humpback whale (*Megaptera novaeangliae*) sound types for passive acoustic monitoring," J. Acoust. Soc. Am. **129**, 476–482.

Suzuki, R., and Shimodaira, H. (**2004**). "An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: how accurate are these clusters?," Poster presented at the 15th Annual International Conference of Genome Informatics, Posters and Software Demonstrations. Yokohama, Japan. (http://www.is.titech.ac.jp/~shimo/pub/GIW2004/suzukiGIW2004.pdf)

Tougaard, J., and Eriksen, E. (**2006**). "Analysing differences among animal songs quantitatively by means of the Levenshtein distance measure," Behaviour **143**, 239–252.

Wieling, M., and Nerbonne, J. (**2015**). "Advances in Dialectometry," Annu. Rev. Linguist. **1**, 243-264.

Winn, H. E., and Winn, L. K. (**1978**). "The song of the humpback whale *Megaptera novaeangliae* in the West Indies," Mar. Biol. **47**, 97–114.

TABLES

TABLE I. Variables measured for each unit.

Measurement	Description
Duration (s)	Vocalization length
Minimum frequency (Hz)	Minimum frequency
Maximum frequency (Hz)	Maximum frequency
Start frequency (Hz)	Start frequency
End frequency (Hz)	End frequency
Frequency range (as ratio)	Max freq/min freq
Frequency trend (as ratio)	Start freq/end freq
Bandwidth (Hz)	Max-min freq
Inflections	Number of reversals in slope
Peak frequency (Hz)	Frequency of the spectral peak
Pulse rate (/s)	for pulsative sounds

TABLE II. Set medians from the Blue song type with and without weighting. N is the number of strings for each theme present in the data. Weight is un-w = un-weighted, β = 1 is the default weight of exponential coefficient, β = 0.5 is weighted to relax the exponential coefficient to 0.5, and β = 0.25 is weighted to relax the exponential coefficient to 0.25 (see Fig. 2). Sum similarity is the highest summed similarity score of a string within the set. This string became the set median string. Note the set median can change in arrangement between each of the four analyses (un-weighted, β = 1, β = 0.5 and β = 0.25). % Theme similarity is the average LSI similarity of all strings to all other strings within the theme. Differences between the weighted and un-weighted set median sequences are underlined. Each letter or combination of letters represents a unit type*. A comma separates units.

Theme	N	Weight	Sum similarity	% Theme similarity	Set median unit string/sequence
23	1	un-w	1.00	100	w, dws, w, nws, w, dws, w, dws, w, modws, be
		β = 1	1.00	100	w, dws, w, nws, w, dws, w, dws, w, modws, be
		β = 0.5	1.00	100	w, dws, w, nws, w, dws, w, dws, w, modws, be
		β = 0.25	1.00	100	w, dws, w, nws, w, dws, w, dws, w, modws, be
24	19	un-w	13.96	62.2	as/aws, as/aws, as/aws, e
		β = 1	14.13	64.8	as/aws, as/aws, as/aws, e
		β = 0.5	14.24	67.1	as/aws, as/aws, as/aws, e
		β = 0.25	14.363712	69.5	as/aws, as/aws, as/aws, e
25a	20	un-w	16.15	73.4	am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t
		β = 1	16.83	78.9	am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t
		β = 0.5	17.05	80.8	am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t
		β = 0.25	17.21	82.0	am(s), gt, gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t
25b	2	un-w	1.83	91.7	am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t, sq, t, mods

		$\beta = 1$	1.83	91.7	am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t, sq, t, mods
		$\beta = 0.5$	1.83	91.7	am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t, sq, t, mods
		$\beta = 0.25$	1.83	91.7	am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t, sq, t, mods
26b	28	un-w	14.86	37.1	s, am, um, modws, um, modws, um, modws
		$\beta = 1$	15.77	43.1	s, am, um, modws, um, modws, um, modws
		$\beta = 0.5$	17.74	52.2	s, am, um, modws, um, modws, am, modws, um, modws
		$\beta = 0.25$	20.13	63.0	s, am, um, modws, um, modws, am, modws, um, modws
27	79	un-w	44.87	41.8	lb, ba, ti(a), sq-ds, ti(a), sq-ds, ti(a), sq-ds, ti(a), sq-ds
		$\beta = 1$	57.60	60.6	lb, ba, ti(a), sq-ds, ti(n), sq-ds, ti(a), sq-ds, ti(n), sq-ds
		$\beta = 0.5$	63.81	71.2	lb, ba, ti(a), sq-ds, ti(n), sq-ds, ti(a), sq-ds, ti(n), sq-ds
		$\beta = 0.25$	68.92	80.3	lb, ba, ti(a), sq-ds, ti(n), sq-ds, ti(a), sq-ds, ti(n), sq-ds
28a	19	un-w	13.89	60.2	lb, ba, am(pul), v, v, v, am(pul), v, v, v, am(pul), v, v, v
		$\beta = 1$	15.19	70.3	lb, ba, am(pul), v, v, v, am(pul), v, v, v, am(pul), v, v, v
		$\beta = 0.5$	15.81	75.5	lb, ba, am(pul), v, v, v, am(pul), v, v, v, am(pul), v, v, v
		$\beta = 0.25$	16.27	79.5	lb, ba, am(pul), v, v, v, am(pul), v, v, v, am(pul), v, v, v
29	11	un-w	7.85	61.4	be, c, c, c
		$\beta = 1$	7.85	62.3	be, c, c, c
		$\beta = 0.5$	7.88	63.3	be, c, c, c
		$\beta = 0.25$	8.13	66.5	be, c, c, c
30b	33	un-w	16.52	44.0	gr/gw, p(ch), c(w), c
		$\beta = 1$	18.52	53.0	gr/gw, p(ch), c(w), c
		$\beta = 0.5$	20.10	58.5	gr, p, gr, p, c, c
		$\beta = 0.25$	22.21	63.9	gr, p, gr, p, c, c

*Unit names: am=ascending moan, am(pul)=pulsative ascending moan, am(s)=short ascending moan, as/aws=ascending shriek/ascending whistle, ba=bark, be=bellows, c=croak, c(w)=croak-whoop, dws=descending whistle, e=e-sound, gr=groan, gr/gw=groan/growl, gt=grunt, lb=long bark, mods=modulated shriek, modws=modulated whistle, nws=n-shaped whistle, p=purr, p(ch)=chainsaw purr, s=siren, sq=squeak, sq-ds=squeak-descending shriek, t=trumpet, ti(a)=ascending trill, ti(n)=n-shaped trill, um=u-shaped moan, v=violin, w=whoop.

TABLE III. A sample of the unit strings/sequences (*i.e.,* phrases) assigned to Themes 25a, 28a and 30b. The un-weighted set median unit string/sequence from Table II is shown below each theme. Each letter or combination of letters represents a unit type*. A comma separates units. Note the variety of unit types and lengths of sequences/strings.

Theme	Unit string/sequence
25a	am(s), gt, gt, gt, gt, am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t, sq, t, sq
	am(s), ba, ba, gt, gt, gt, gt, am(s), t, sq, t, sq, t
	am(s), gt, gt, gt, gt, am(s), t, t, t, sq, t
	w, w/ba, w/ba, ba, ba, am(s), t, sq, t, sq, t
	am(s), gt, gt, gt, gt, am(s), gt, gt, gt, gt, am(s), t, sq, t, t
Set median	am(s), gt, gt, gt, gt, gt, am(s), t, sq, t, sq, t
28a	lb, ba, nm(pul), v, v, v, mm(pul), sq, sq, v, v, mm(pul), v, v, v
	ba, ba, am(pul), sq, sq, sq, sq, am, sq, sq, sq, sq, sq, sq, am, v, v, sq, sq, v
	lb, ba, am(pul), v, v, v, am(pul), v, v
	ba, ba, am(pul), v, v, v, am(pul), v, v, v, am(pul), v, v, v
	lb, ba/am, ti(a), sq, v, v, am(pul), v, v, v, am(pul), v, v, v, am(pul), v, v, v
Set median	lb, ba, am(pul), v, v, v, am(pul), v, v, v, am(pul), v, v, v
30b	gr/gw, p(ch), c(w), c(w)
	gr, p, gr, p, c, c
	gr/gw, p(ch), gr/gw, p(ch), c, c, c
	gr/gw, p, c(w), c(w)
	gr, p, gr, p, c, c, c(w)
Set median	gr/gw, p(ch), c(w), c

*Unit names: am=ascending moan, am(pul)=pulsative ascending moan, am(s)=short ascending moan, ba=bark, ba/am= bark/ascending moan, c=croak, c(w)=croak-whoop, gr=groan, gr/gw=groan/growl, gt=grunt, lb=long bark, mm(pul)=pulsative modulated moan, nm(pul)=pulsative n-shaped moan, p=purr, p(ch)=chainsaw purr, sq=squeak, t=trumpet, ti(a)=ascending trill, v=violin, w=whoop, w/ba=whoop/bark.