Machine Source Localization of *Tursiops truncatus* Whistle-like Sounds in a Reverberant Aquatic Environment

SF Woodward[1]*, D Reiss[2], MO Magnasco[1]

1 Laboratory of Integrative Neuroscience, Center for Studies in Physics and Biology, The Rockefeller University, New York, NY, United States
2 Department of Psychology, Hunter College, City University of New York, New York, NY, United States

* sf.woodward@gmail.com

Abstract

Most research into bottlenose dolphins' (*Tursiops truncatus'*) capacity for communication has centered on tonal calls termed *whistles*, in particular individually distinctive contact calls referred to as *signature whistles*. While "non-signature" whistles exist, and may be important components of bottlenose dolphins' communicative repertoire, they have not been studied extensively. This is in part due to the difficulty of attributing whistles to specific individuals, a challenge that has limited the study of not only non-signature whistles but the study of general acoustic exchanges among socializing dolphins. In this paper, we propose the first machine-learning-based approach to identifying the source locations of semi-stationary, tonal, whistle-like sounds in a highly reverberant space, specifically a half-cylindrical dolphin pool. We deliver time-of-flight and normalized cross-correlation measurements to a random forest model for high-feature-volume classification and feature selection, and subsequently deliver the selected features into linear discriminant analysis, linear and quadratic Support Vector Machine (SVM), and Gaussian process models. In our 14-point setup, we achieve perfect classification accuracy and high (MAD of 0.6557 m, IQR = 0.3395 - 1.5694) regression accuracy with fewer than 10,000 features. The regression models yielded better accuracy than the established Steered-Response (SRP) method when all training data were used, and comparable accuracy - even when interpolating at several meters - in the lateral directions when deprived of training data at testing sites; our methods additionally boast improved computation time and the potential for superior accuracy in all domains with more training data.

Introduction

Dolphin communication research is in an active period of growth. Many researchers expect to find significant communicative capacity in dolphins given their complex social structure [1–3], advanced cognition including the capacity for mirror self-recognition [4], culturally transmitted tool-use and other behaviors [5], varied and adaptive foraging strategies [6], and their capacity for metacognition [7]. Moreover, given dolphins' well-studied acoustic sensitivity and echolocation ability [8–10], some researchers have speculated that dolphin vocal communication might share properties with human languages [11–13]. However, there is an insufficiency of work in this area to make substantive comparisons.

Among most dolphin species, a particular tonal class of call, termed the *whistle*, has been identified as socially important. In particular, for the common bottlenose dolphin, *Tursiops truncatus* – arguably the focal species of most dolphin cognitive and communication research – research has focused on *signature whistles*, individually distinctive whistles [14–16] that may convey an individual's identity to conspecifics [15, 17] and that can be mimicked, potentially to gain conspecifics' attention [18].

Signature whistle studies aside, most studies of bottlenose dolphin calls concern group-wide repertoires of whistles and other, pulse-form call types [19–23]; there is a paucity of studies that seek to examine individual repertoires of non-signature whistles or the phenomenon of non-signature acoustic exchanges among dolphins. Regarding the latter, difficulties with sound attribution at best allow for sparse sampling of exchanges [17, 24]. Nevertheless, such studies constitute a logical prerequisite to an understanding of the communicative potential of whistles.

The scarcity of such studies can be explained in part by a methodological limitation in the way in which dolphin sounds are recorded. In particular, no established method exists for recording the whistles of an entire social group of dolphins so as to reliably attribute the signals to specific dolphins. The general problem of sound attribution, which is encountered in almost every area of communication research, is typically approached in one of two ways: (1) by attaching transducers to all potential sound sources, in which case the source identities of sounds can usually be obtained by discarding all but the highest-amplitude sounds in each source-distinctive recorder, or (2), by using a fixed array (or arrays) of transducers, a physics-based algorithm for identifying the physical origin of each sound, and cameras that monitor the physical locations of all potential sources for matching.

While notable progress has been made implementing attached transducers (or tags) to identify the sources of dolphin whistles [25–27], shortfalls include the need to manually tag every member of the group under consideration, the tendency of tags to fall off, and the tags' inherent lack of convenient means for visualizing caller behavior. Most significant to research with captive dolphins, the use of tags can conflict with best husbandry practices (e.g., due to risk of skin irritation, of ingestion) and be forbidden, as is the case at the National Aquarium. At such locations, less invasive means of sound attribution are necessary. Unfortunately, a reliable implementation of the array/camera approach to dolphin whistles has not been achieved, though it has been achieved for more tractable dolphin clicks [28]. In the context of whistles in reverberant environments, authors have noted the complications introduced by multipath effects – resulting from the combination of sounds received from both the sound source and acoustically reflective boundaries – to standard signal processing techniques. These complications generally arise from the overlap of original and reflected sounds that confound standard, whole-signal methods of obtaining time-of-flight differences. Standard techniques have at best obtained modest results in relatively irregular, low-reverberation environments where they have been evaluated [29–32]. In unpublished work, we have achieved similar results. One method of improving a standard signal processing tool for reverberant conditions, the cross-correlation, has been proposed without rigorous demonstration and has not be reproduced [33]. Among all previous methods we have identified those falling under the umbrella of Steered-Response Power (SRP) most effective. In short, these methods rely on maximizing the sum of cross-correlations between all pairs of hydrophone signals with respect to sets of time shifts between signals/hydrophones, each set corresponding to a hypothetical source location in the pool [34]; hypothetical source locations may correspond to equally-spaced grid points in the suspect zone, the naive approach, or be iteratively chosen in a more efficient fashion [35]. While the details of the particular SRP method employed are left

vague, Rebecca E. Thomas et al. [36] have rigorously demonstrated such a method with reasonable success (used for about 40% recall of caller identity) [36].

We propose the first machine-learning-based solution to the problem of localizing whistle-like sounds in a highly reverberant environment, a half-cylindrical concrete dolphin pool, located at the National Aquarium in Baltimore, Maryland. We apply it to a broad variety of artificial tonal whistle-like sounds that vary over a range of values within a universally recognized parameter space for classifying dolphin sounds, for a limited number of sampling points. We begin with a random forest classification model and later find that a linear classification model achieves similar results, as well as a regression model that achieves dolphin-length accuracy depending on training circumstances (discussed below). The latter two models rely on small (or parsimonious) feature sets containing fewer than 10,000 features to locate single whistles. We implement an SRP method to compare our results.

Materials and methods

Sample Set

All data were obtained from equipment deployed at the Dolphin Discovery exhibit of the National Aquarium in Baltimore, Maryland. The exhibit's 110-ft-diameter cylindrical pool is subdivided into one approximate half cylinder, termed the *exhibit pool* (EP), as well three smaller holding pools, by thick concrete walls and 6 ft x 4.25 ft (1.83 m x 1.30 m) perforated wooden gates; all pools are acoustically linked. The data were obtained from the EP, when the seven resident dolphins were in the holding pools; their natural sounds were present in recordings.

To ensure that the sound samples used for classification were not previously distorted by multipath phenomena (i.e., were not pre-recorded), were obtained in sufficient quantity at several precise, known locations inside the EP, and were representative of the approximate "whistle space" for *Tursiops truncatus*, we chose to use computer-generated whistle-like sounds that would be played over an underwater Lubbell LL916H speaker.

We generated 128 unique sounds (with analysis done on 127) to fill the available time. To be acoustically similar to actual *T. truncatus* whistles, these sounds were to be "tonal" – describable as smooth functions in time-frequency space, excluding harmonics – and to be defined by parameters and parameter ranges, given in Table 1, representative of those used and observed by field researchers to characterize dolphin whistles [37,38]. To construct a waveform to be played, we began with an instantaneous frequency, $f(t)$, that described a goal time-frequency (or spectrographic) trace, for instance the trace shown in Fig 1. For simplicity, and consistent with the parameters typically used to describe dolphin whistles, we approximated dolphin whistles as sinusoidal traces in spectrographic space – thus $f(t)$ was always a sinusoid. Based on the standard definition of the instantaneous frequency as $f(t) = \frac{1}{2\pi}\frac{d\Phi(t)}{dt}$, we obtained the phase $\Phi(t)$ by integration of $f(t)$ with respect to time. The phase could be straightforwardly transformed into a playable waveform $y(t)$ as $y(t) = A(t)sin(\Phi(t))$, where $A(t)$ represented a piecewise function that modulated the intensity of the signal at different times (the beginning and end of a signal were gradually increased to full intensity and decreased to zero, respectively, as functions of the "Power Onset/Decay Rate," and the absolute beginning occurred at a peak or trough of the sinusoid $f(t)$ according to "Phase Start"). Alternatively, the phase derived for $f(t)$ could be transformed with a heuristic into a waveform corresponding to a slightly modified version of $f(t)$, specifically a quasi-sinusoid with "sharpened" peaks and approximate whistle-harmonic-like traces higher in frequency than the fundamental. This heuristic is

Table 1. Parameters of training set sinusoids.

Parameter	Value Set
Duration (sec)	[0.3, 1]
Number of Cycles	[1, 2]
Center Frequency (Hz)	[6000, 10500]
Cycle Amplitude (Hz)	[2000, 5000]
Phase Start (rad)	$[-\frac{\pi}{2}, \frac{\pi}{2}]$
Power Onset/Decay Rate *	[0.1, 0.25]

* Values indicate fraction of signal length over which a sin^2 rise/falls occurs.

$y(t) = A(t)\frac{arcsin(m \cdot sin(\Phi(t)))}{arcsin(\Phi(t))}$, where m is a parameter that simultaneously affects the "sharpness" of the peaks and the number of harmonics; we used a value of 0.8.

Waveforms were played in Matlab through a MOTU 8M audio interface at calibrated volumes and a sampling rate of 192 kHz. An example of pre-speaker output is given in Fig 1.

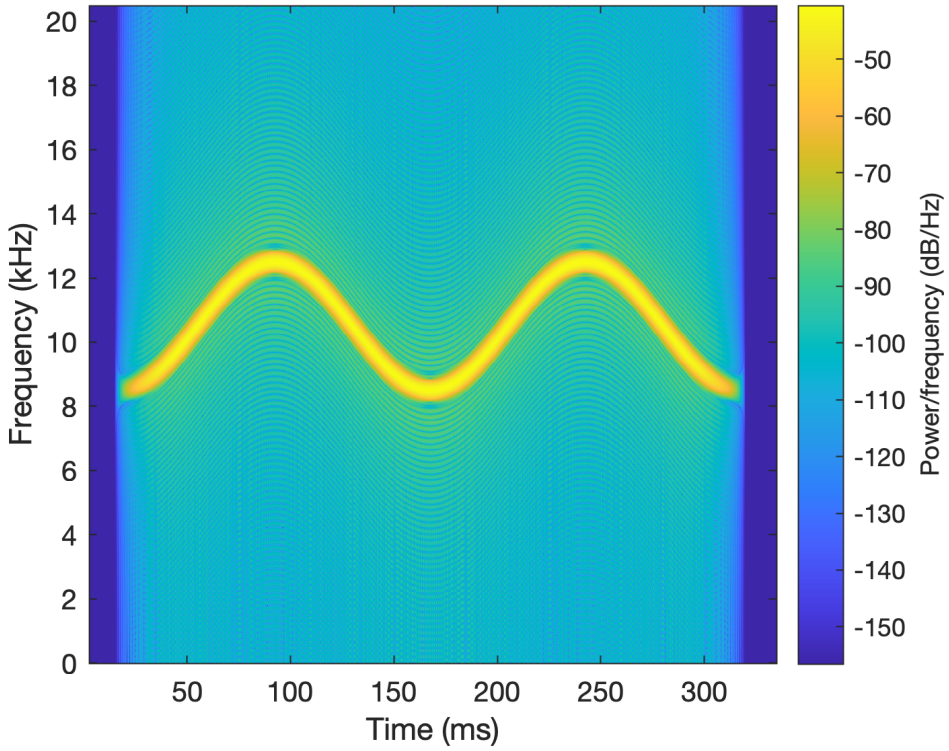

Fig 1. Spectrogram of an artificial whistle. Displayed is a standard, 1024-bin, Hamming-window spectrogram of one of the 128 whistle-like sounds that were generated (and here sampled) at 192 kHz; frequency resolution of the plot is 187.5 Hz (smoothing added). Note that the spectrogram was constructed from the unplayed source signal. In this case, Duration = 1 second, Number of Cycles = 2, Center Frequency = 10500 Hertz, Cycle Amplitude = 2000 Hertz, Phase Start = -π/2.

The 128 sounds were played at each of 14 locations within the EP; they corresponded to 7 unique positions on the water surface on a 3 x 5 cross, at 6 ft (1.83

m) and 18 ft (5.49 m) deep. Approximate surface positions are shown in Fig 2; the difference between adjacent horizontal and vertical positions was 10-15 ft (3.05-4.57 m). The LL916H speaker was suspended by rope from a custom flotation device and moved across the pool surface by four additional ropes extending from the device to research assistants standing on ladders poolside. Importantly, the speaker was permitted to sway from its center point by approximately 1/3 m (as much as 1 m) in arbitrary direction during calibration. These assistants also used handheld Bosch 225 ft (68.58 m) Laser Measure devices to determine the device's distance from their reference points (several measurements were taken for each location), and through a least-squares trilateration procedure [39] the device location could always be placed on a Cartesian coordinate system common with the hydrophones. Each sound in a 128-sound run was played after a 2-second delay as well as a 0.25-second, 2-kHz tone, that allowed for the creation of a second set of time-stamps in order to compensate for clock drift during the automated signal extraction.

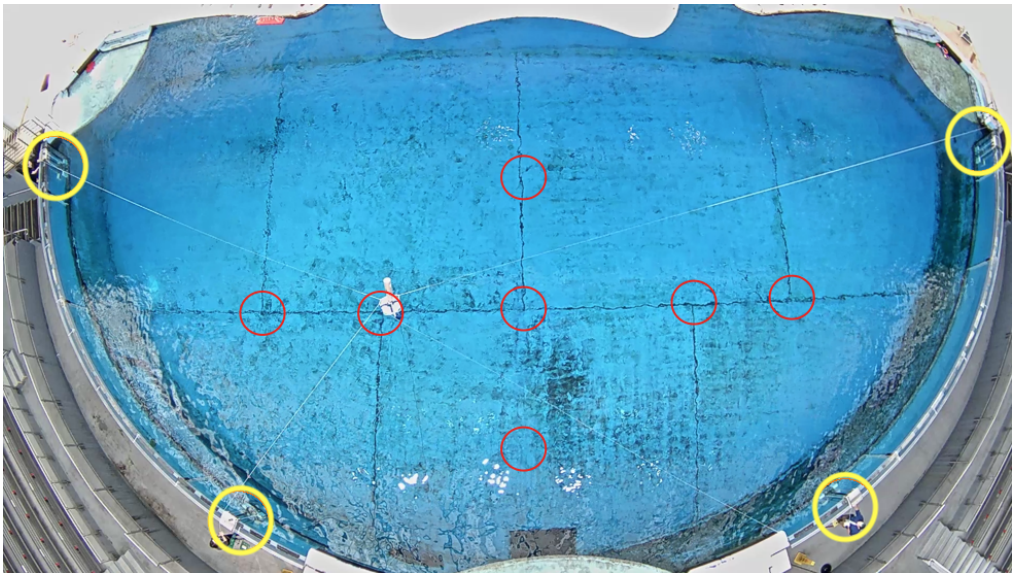

Fig 2. Pool grid point and hydrophone array layout. The National Aquarium Exhibit Pool (EP) is shown, as visualized by the central overhead AXIS P1435-LE camera. Circled in red are the approximate surface projections of the fourteen points (each of the seven points represents a pair) at which sounds were played. Circled in yellow are the four hydrophone arrays, each containing four hydrophones; they are numbered 1-4 from left to right.

Recording System

Acoustic and visual data were obtained from a custom audiovisual system consisting of 16 hydrophones (SQ-26-08's from Cetacean Research Technology, with approximately flat frequency responses between 200 and 25,000 Hz) split among 4 semi-permanent, tamper-resistant arrays and 5 overhead cameras – for the purpose of this study, only one central AXIS P1435-LE camera, managed by Xeoma surveillance software, was used. The four arrays were spaced approximately equally around the half-circle boundary of the EP (a "splay" configuration). The physical coordinates of all individual hydrophones were obtained by making underwater tape-measure measurements as well as above-water laser-rangefinder measurements; various calibrations were performed that are outside the scope of the present paper. Audio recordings were collected at 192

kHz by two networked MOTU 8M audio interfaces into the Audacity AUP sound format, to avoid the size limitations of standard audio formats. The same two audio interfaces were involved in playing the sounds. Standard passive system operation was managed by Matlab scripts recording to contiguous WAV files; for consistency Matlab was also used for most data management and handling. Data are available at `https://doi.org/10.6084/m9.figshare.7956212`.

Classification and Regression

1,605 recorded tones were successfully extracted to individual 2-second-long, 16-channel WAV's that were approximately but not precisely aligned in time. Each tone was labeled with a number designating at which of 14 pool locations it was played. A random 10% of sounds were set aside for final testing, with sinusoids and quasi-sinusoids with the same parameters grouped together given their close similarity.

Each sound was initially digested into 1,333,877 continuous, numerical features (or variables), together composing the so-called feature set.

The first 120 features were *time-difference-of-arrivals* (TDOA's) obtained using the Generalized Cross-Correlation Phase Transform (GCC-PHAT) method [40], which in currently unpublished work we found to be most successful among correlation-based methods for obtaining whistle TDOA's. Briefly, a TDOA is the time difference between the appearance of a signal (such as a whistle) in one sensor (such as a hydrophone) versus another. Possession of a signal's TDOA's for all pairs of 4 sensors (with known geometry) in 3-dimensional space is theoretically, but often not practically, sufficient to calculate the exact source position of that signal from geometry [41]. While we establish elsewhere that the 120 TDOA's (i.e., one TDOA for each distinct pair of our 16 hydrophones, excluding self-pairs) obtained from GCC-PHAT are practically not sufficient to calculate the exact source position of our signals using a standard geometric technique that accommodates more than four sensors, Spherical Interpolation [42, 43], we suspected that these TDOA's might still contain information helpful to a machine learning model.

The next 6601 x 136 features consisted of elements from standard, normalized circular cross-correlations [44]: for each unique pair of the 16 hydrophones (including self-pairs), 136 in total, we computed the standard circular cross-correlation of a whistle's two audio snippets. While each correlation series was initially 384,000 elements long (192,000 samples/second x 2 seconds), we only kept the central 6601 elements from each, corresponding to a time-shift range of approximately ± 17 milliseconds; based on geometry, the first incidence of any sound originating from inside the pool must have arrived within 17 milliseconds between any two sensors (conservatively), meaning cross-correlation elements corresponding to greater delays could be expected to be less helpful to in-pool source location prediction.

Lastly, we included 27,126 x 16 discrete Fourier transform elements (one set for each of 16 hydrophones, for frequencies from 0 Hz to 27,126 Hz). However, preliminary analysis found these features to be unhelpful to classification. Thus, they were discarded from the feature set, leaving 897,871 features.

Possessing the above feature set for each whistle, each "labeled" by the whistle source location and coordinates, we constructed predictive models on the 90% of whistles made available for model construction (or training). We began with multiclass classification [45], training models to predict which of the 14 possible locations a novel whistle originated from. Given our limited computational resources, our feature set remained too large to accommodate most classifiers. A notable exception was the Breiman random forest [46, 47], which was suitable not only for classification – being a powerful nonlinear multiclass classifier with built-in resistance to overfitting – but for feature reduction (i.e., the process of shrinking feature set size while minimizing loss of

classification accuracy), via the permuted variable delta error metric. The permuted variable delta error roughly describes the increase in classification error when a particular feature is effectively randomized, providing a measure of that feature's importance in classification. We grew a Breiman random forest composed of CART decision trees [48] on the training data; each tree was sequentially trained on a random subset of ~75% of the training samples using a random $\sim\sqrt{897,871}$ feature subset (as per standard practice). Out-of-bag (OOB) error, referring to the classification error on samples not randomly chosen for the training subset, was used for validation. While an introduction to machine learning is beyond the scope of this paper, these and subsequent techniques are standard with reader-friendly documentation available at such sources as MathWorks (`https://www.mathworks.com`) and sciikit-learn (`https://scikit-learn.org/stable/`), to which we direct interested readers.

We subsequently used permuted variable delta error as a measure of feature importance, both to examine the selected features for physical significance – recall that cross-correlation element features correspond to pairs of sensors – and to obtain a reduced feature set appropriate for training additional models. The reduced feature set included the 6,788 features with nonzero permuted variable delta error. On the reduced feature set, we considered a basic CART decision tree [45, 48], a linear and quadratic Support Vector Machine (SVM) [45, 49], and linear discriminant analysis [50].

We also considered Gaussian process regression (also termed *kriging*) [51, 52] – a nontraditional, nonparametric method of regression that could accommodate our under-constrained data. Whereas the purpose of our classification models was to predict from which of 14 possible points a sound originated, the purpose of our regression models was to predict the three Cartesian coordinates from which the sound originated. Localization by regression was performed two ways. In the first way, all training sounds were used to generate three models (one per dimension) for predicting the coordinates of all test sounds. In the second way, training sounds from all but one grid point were used to generate models for predicting the coordinates of test sounds from the excluded point; the process was repeated for all grid points, the results aggregated. While we were doubtful of our models' ability to precisely interpolate at distances of several meters from 14 points, this test was envisioned to show that reduced-feature-set models are capable of a degree of spatial interpolation.

For comparison, we employed a Steered-Response Power approach [34] to localizing the sounds in the test set. As the details to the particular method used by Thomas et al. [36] are unpublished, we followed a standard procedure: for a hypothetical sound originating from each of every point of a 6" (15.25 cm) virtual gridding of the pool, we calculated the theoretical differences in the time of arrival for our 16 hydrophones. Shifting the 16 signals of a received whistle in our test set by each of every set of differences, then cross-correlating the shifted signals and summing the results along both axes, the predicted source location corresponded to the grid point that produced the largest value thus computed. We calculated the speed of sound from the Del Grosso equation [53]; the pool salinity was 31.5 ppt and temperature 26.04 °C. As an aside, we attempted to replace the standard cross-correlation in the prior calculation with the Generalized Cross Correlation with Phase Transform [40], which constitutes the SRP-PHAT technique for sound localization discussed in [34], but quickly found the results to be inferior.

Lastly, we obtained a minimal, nearly sufficient feature set by training a single, sparse-feature (or parsimonious) decision tree classifier on all features of all training data. We then investigated these minimal features for physical significance, by mapping features' importance (again, using a random forest's permuted variable delta error) back to the sensor and array pairs from which they were derived.

Results

The random forest classification model trained on the full feature set, as described above, reached 100.0% OOB accuracy at a size of approximately 180 trees. We continued training to 300 trees, and evaluated the resulting model on the test set: 100.0% accuracy was achieved, with 6,788 features possessing permuted variable delta error greater than 0 (based on OOB evaluations). Note that, given the stochastic construction of the random forest, these features did not represent a unique set or superset of sufficient features for obtaining 100.0% test accuracy. When we considered which array pairs the 6,778 TDOA and cross-correlation features represented, we found that all pairs of the four hydrophone arrays were represented with no significant preference.

We trained several more models on the reduced, 6,788-item feature set, including a basic decision tree, a linear and quadratic SVM, and linear discriminant analysis, using 10-fold cross-validation. The quadratic SVM as well as linear discriminant analysis achieved 100.0% cross-validation and 100.0% test accuracy, the basic decision tree achieved 96.90% cross-validation and 97.75% test accuracy (95% CI [97.06 - 98.44]), and the linear SVM achieved 100.0% cross-validation accuracy and 99.44% test accuracy (95% CI [98.34 - 100.0]). Confidence intervals reflect Wilson scores.

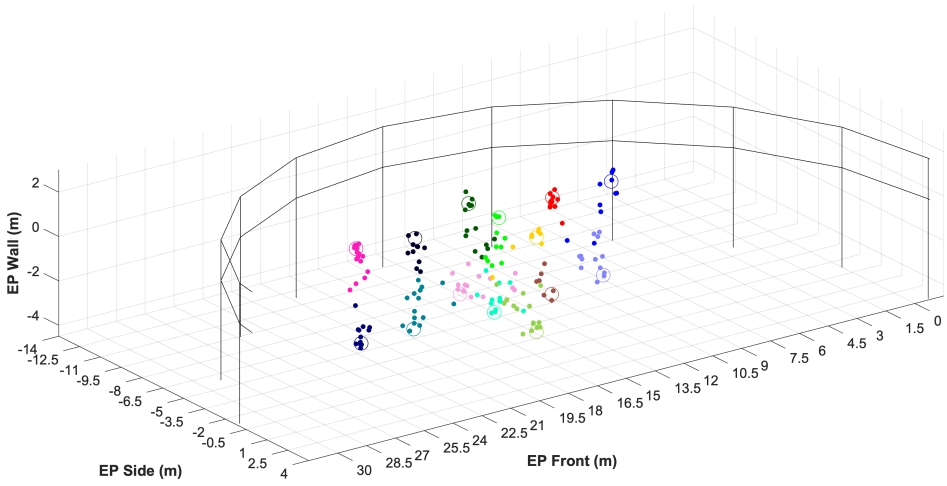

Fig 3. Predictions of test sound coordinates by Gaussian process regressors. The half-cylindrical National Aquarium EP is depicted. Large unfilled circles indicate the true coordinates of the test sounds; each has a unique color. Small filled circles indicate the test sound coordinates predicted by Gaussian process regression, colors matching their respective true coordinates.

Again using the reduced feature set, we performed Gaussian process regression (kriging) to predict test sound coordinates, generating one model for each Cartesian axis. A random subset of predicted coordinates are plotted in Fig 3. The calculated Euclidean (straight-line) error or median absolute deviation (MAD) was 0.6557 m (IQR = 0.3395 - 1.5694); along the "EP Front" axis MAD was 0.1909 m (IQR = 0.0702 - 0.3891), along the "EP Side" axis MAD was 0.1301 m (IQR = 0.0481 - 0.3367), and along the "EP Wall" axis MAD was 0.5191 m (IQR = 0.1644 - 1.1771).

When the Gaussian process regression models were generated to predict the coordinates of test sounds on grid points from which they received no training data (three models were generated for each grid point, generated from training data from all other grid points), Euclidean MAD was 3.3691 m (IQR = 2.8497 - 3.7480); along the

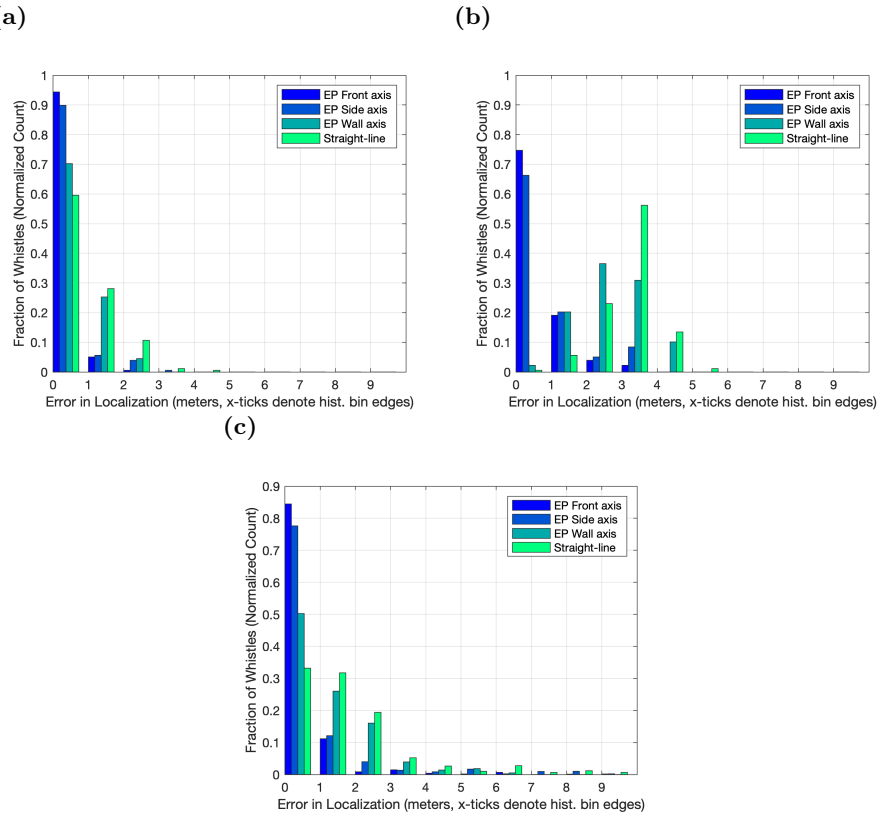

Fig 4. Multi bar graphs displaying the histogrammed localization error (absolute deviation) for three whistle localization methods. Both straight-line (Euclidean) error and error along each of three component Cartesian axes are displayed. (A) Localization error for Gaussian process regression, training on all training whistles and predicting coordinates of all test whistles. (B) Localization error for Gaussian process regression, using models deprived of training data from grid points of evaluated test sounds. (C) Localization error for SRP on test whistles.

"EP Front" axis MAD was 0.5617 m (IQR = 0.2583 - 1.0282), along the "EP Side" axis MAD was 0.5024 m (IQR = 0.1658 - 1.6241), and along the "EP Wall" axis MAD was 2.7324 m (IQR = 2.1410 - 3.3854).

For the SRP method applied to the test sounds, Euclidean MAD was 1.5572 m (IQR = 0.7277 - 2.4839); MAD along the "EP Front" axis was 0.4267 m (IQR = 0.1829 - 0.7315), MAD along the "EP Side" axis was 0.5182 m (IQR = 0.2134 - 0.9754), and MAD along the "EP Wall" axis was 0.9144 m (IQR = 0.1524 - 1.9812).

The histogrammed localization errors (absolute deviations) for all the above predictions are displayed in Fig 4. Anderson-Darling tests rejected the null hypotheses that the sets of errors were drawn from normal distributions (5% significance level), and we therefore proceeded with non-parametric statistics. A Kruskal-Wallis test on the error sets' mean ranks determined the sets did not originate from the same distribution (5% significance level); results for Dunn's post-hoc comparisons of individual pairs are given in Fig 5.

Lastly, we trained a single, sparse-feature decision tree (a so-called parsimonious model) on the full training set. The severe feature reduction left 22 features. While the decision tree achieved only 96.63% accuracy (95% CI [93.98 - 99.28]) on the test set, a random forest trained on the same features achieved 98.88% test accuracy (95% CI

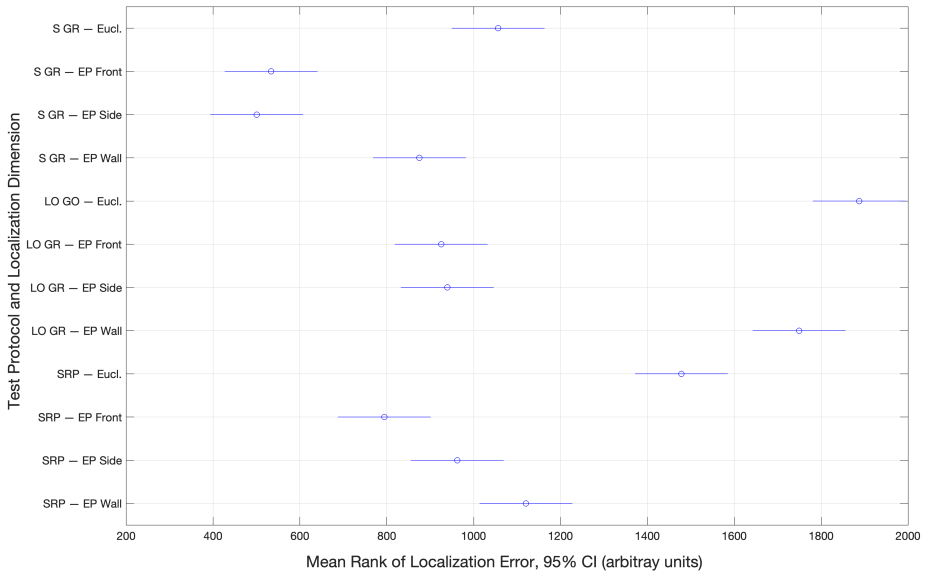

Fig 5. Mean ranks of test sound localization error. The mean rank for each test protocol and axis of localization, computed ahead of a Kruskal-Wallis test (discussed in the text), are shown with 95% CI's. Comparing every pair of groups using post-hoc Dunn's tests with an overall significance threshold of 0.05, we confirmed that interval overlaps visualized here reflect an inability to reject the null hypothesis that the underlying samples are drawn from the same distribution; similarly, interval non-overlaps visualized here reflect rejection of the null hypothesis that the underlying samples are drawn from the same distribution. Note, "S GR" (Standard Gaussian regression) refers to Gaussian regression performed with training data from grid locations of test sounds made available for model building; "LO GR" (Leave-out Gaussian regression) refers to Gaussian regression performed with training data for grid locations of test sounds excluded from model building.

[97.73 - 100.0]). Thus, we considered this feature set both sufficient and sparse enough to meaningfully ask whether classification is making use of features derived from a spatially mixed set of hydrophone and hydrophone array pairs, consistent with a geometric approach to sound source localization. The permuted variable delta error was summed across hydrophone and hydrophone array pairs, which is visualized in Fig 6. Overall, we note that, directly or indirectly, features representing all pairs of hydrophone arrays are utilized for classification.

Discussion

We provided a proof of concept that sound source localization of semi-stationary bottlenose whistle playbacks can be achieved implicitly as a classification task and explicitly as a regression task in a standard, highly reverberant, half-cylindrical captive dolphin enclosure. Moreover, for the same conditions we showed that, for the localization of whistles originating near training-set sounds (within ˜1/3 m, the speaker sway range), Gaussian regression outperforms a standard Steered-Response Power (SRP) approach to whistle localization. Localizing in the lateral directions, which is often sufficient for distinguishing among potential sound sources based on overhead imaging,

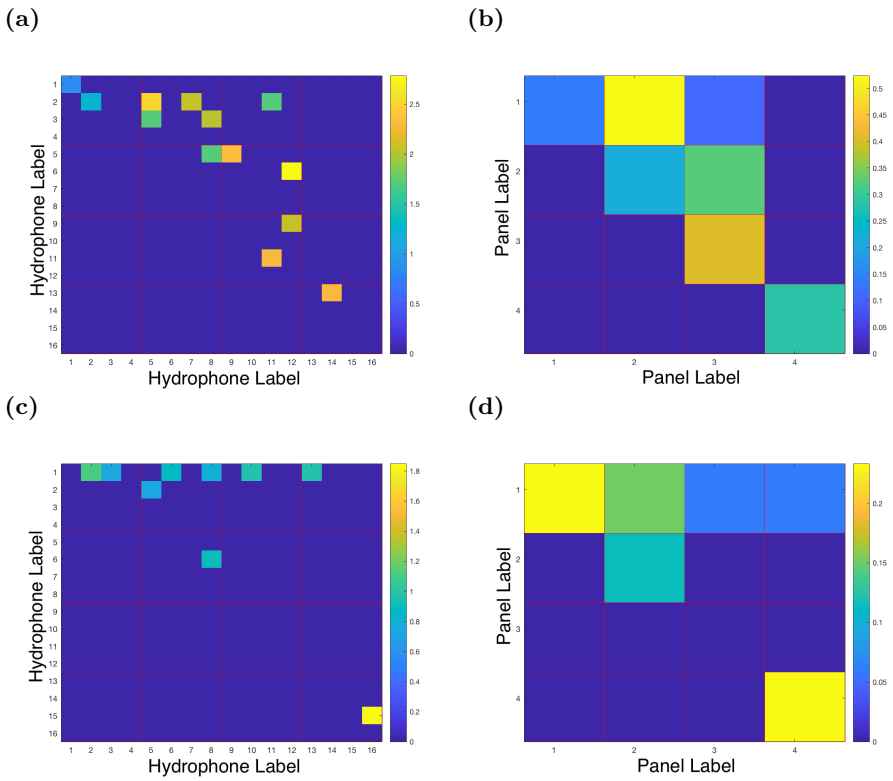

Fig 6. Cross-hydrophone and cross-hydrophone-array feature importances for the parsimonious random forest. Feature importance values for a parsimonious, 22-feature random forest model, summed across corresponding hydrophone pairs and averaged across corresponding hydrophone array pairs. (A) Cross-hydrophone importances for cross-correlation features. Hydrophones belonging to common panels (1-4, 5-8, 9-12, 13-16) are grouped by red boxes. (B) Cross-hydrophone-array importances for cross-correlation features. Panels are numbered 1-4 from left-to-right in Fig 2. (C) Cross-hydrophone importances for TDOA features. (D) Cross-hydrophone-array importances for TDOA features.

Gaussian process models prompted to predict test sound coordinates at novel grid points (potentially interpolating over several meters) did so with similar accuracy to SRP.

The data consisted of 16 independent recordings (from four four-hydrophone arrays) of 127 unique bottlenose-dolphin-whistle-like sounds played at 14 positions in the dolphin exhibit pool (EP) at the National Aquarium, semi-randomly divided into "training" and "test" sets for model building and evaluation, respectively. First, we showed that a random forest classifier with fewer than 200 trees can achieve 100% testing accuracy at the task of predicting from which of the 14 locations a sound originated, using 6,788 of 897,871 available features, including TDOA's obtained from GCC-PHAT as well as normalized cross-correlations from all pairs of sensors. We then showed that linear discriminant analysis and a quadratic SVM can achieve the same classification accuracy on the reduced, 6,788-feature set. If the linear model in particular were to remain accurate when trained on a finer grid of training/testing points (finer by about two fold, which would reduce the distance between grid points to approximately the length of a mature bottlenose dolphin), it would constitute a simple and computationally efficient method of locating the origin of tonal sounds in a reverberant environment.

Although it remains unclear to what extent sounds originating off-grid are classified to the most logical (i.e., nearest) grid points, we note that our classifiers' success was achieved despite the ˜1/3 m drift of the speaker during play-time; this together with the success of regression may indicate a degree of smoothness in the classifiers' decision-making. Also, that a linear classifier, which by definition cannot support nonlinear decision making, suffices for this task on features that are generally expected to vary continuously in value across space (TDOA's, cross-correlations) is reassuring. Nevertheless, this question does warrant further investigation, perhaps using faster-moving sound sources – an investigation of this nature should also be performed to better evaluate the method's ability to localize sounds produced by dolphins.

We more suitably addressed the question of off-grid prediction using Gaussian process regression to predict the coordinates of the test sounds. This method was also quite successful when trained on the full training data set, achieving test MAD of 0.6557 m (IQR = 0.3395 - 1.5694) – less than the expected length of an adult common bottlenose dolphin [54]. In order to better assess the regression models' capacity for interpolation, we evaluated the regression models' performance on test sounds for which no training sounds from the same grid points were used for model generation. While the regressors' overall performance on novel points was not satisfactory, admitting error larger than average dolphin length (MAD of 3.3691 m, IQR = 2.8497 - 3.7480), when we decomposed the error along three Cartesian axes ("EP Front" MAD of 0.5617 m with IQR = 0.2583 - 1.0282, "EP Side" MAD of 0.5024 m with IQR = 0.1658 - 1.6241, and "EP Wall" MAD of 2.7324 m with IQR = 2.1410 - 3.3854), we found that the overall prediction error was significantly dominated (referring to Fig 5) by localization error in the direction of pool depth. This is significant because sounds from only two distinct pool depths were obtained, which is intuitively unsuitable for interpolation. We think it is reasonable to suggest interpolation in this direction would improve with finer sampling. Moreover, we note that MAD for interpolation in the other two directions was less than average adult dolphin body length.

For comparison, we localized the test sounds using a standard SRP approach that has met success elsewhere, particularly in Thomas et al. [36]. Referring to Fig 5, Gaussian regression significantly outperformed SRP when all training data was used. When the regression models were deprived of training data from grid points of evaluated test sounds, SRP performed better overall and along the "EP Wall" (i.e., depth) axis; there was no significant performance difference along the other two component axes. While this suggests that with the current training data Gaussian regression does not outperform SRP in three dimensions when prompted to interpolate at longer distances, it also suggests the interpolation models are capable of comparable accuracy in the lateral directions; the same might be expected in the vertical direction were more than two depth points available for interpolation. Moreover, the results suggest that Gaussian regression can perform just as well as SRP for localizing sounds across the pool surface (i.e., disregarding depth), which is often sufficient for distinguishing among potential sound sources based on overhead imaging (as in Thomas et al.), even at the disadvantage of needing to interpolate over distances of several meters.

Nevertheless, it remains unclear to what extent naturally produced dolphin whistles can be localized. Such whistles are produced by sources that are faster-moving and possessive of different anisotropic properties than our speaker. Evaluation would require a large set of dolphin whistles generated at known locations in the pool, which we do not possess at present, and cannot obtain due to removal of our equipment. However, even were an evaluation of real dolphin whistles to fail due to the models' inability to generalize in "whistle space," we note that in general captive dolphins' whistle repertoires tend to be limited – groups seem to possess less than 100 unique types [21] – and that it would be realistic to train classification/regression models with whistles

closely resembling group members' sounds, avoiding the need for the model to generalize over all whistle space.

Lastly, we showed that an extremely sparse, 22-item feature set that lends itself to relatively strong classification accuracy includes time-of-flight comparisons from all four pairs of arrays. As sound amplitude information was removed in the process of feature creation, this suggests that the classification and regression methods discussed here implicitly use time-of-arrival information for classification from four maximally spaced sensors, consistent with a naive analytic-geometric approach to sound source localization. However, the inner logic of the models ultimately remains unknown.

Overall, we feel this study offers a strong argument that machine learning methods are suitable to solving the problem of bottlenose whistle localization in highly reverberant aquaria, where tag-based solutions to whistle attribution are not feasible. We offer evidence to suggest that these methods might be capable of greater accuracy than SRP methods given adequate training data, coming at smaller computational expense – requiring evaluation of approximately 6,788 features per sound versus performing multiple signal cross-correlations per sound. While we caution that these methods still must be evaluated for real dolphin whistles (representing sources that are faster-moving and possessing different anisotropy than our speaker), we opine that our results are encouraging and warrant further research.

Acknowledgments

We thank the National Aquarium for participating in this study, as well the National Science Foundation (Awards 1530544, 1607280), the Eric and Wendy Schmidt Fund for Strategic Innovation, and the Rockefeller University for funding. While regrettably we cannot name everyone, we also thank the approximately two dozen people at the National Aquarium, the Rockefeller University, and Hunter College for assisting with various aspects of the project.

References

1. Shirihai H, Jarrett B. Whales, Dolphins, and Other Marine Mammals of the World. Princeton and Oxford: Princeton University Press; 2006.
2. Connor RC, Heithaus MR, Barre LM. Complex social structure, alliance stability and mating access in a bottlenose dolphin 'super-alliance'. Proceedings of the Royal Society B: Biological Sciences. 2001;268(1464):263–267.
3. Krutzen M, Sherwin WB, Connor RC, Barre LM, Van de Casteele T, Mann J, et al. Contrasting relatedness patterns in bottlenose dolphins (Tursiops sp.) with different alliance strategies. Proceedings of the Royal Society B: Biological Sciences. 2003;270(1514):497–502.
4. Reiss D, Marino L. Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. Proceedings of the National Academy of Sciences. 2001;98(10):5937–5942.
5. Krutzen M, Mann J, Heithaus MR, Connor RC, Bejder L, Sherwin WB. Cultural transmission of tool use in bottlenose dolphins. Proceedings of the National Academy of Sciences. 2005;102(25):8939–8943.
6. Sargeant BL, Mann J, Berggren P, Krutzen M. Specialization and development of beach hunting, a rare foraging behavior, by wild bottlenose dolphins (Tursiop sp.). Canadian Journal of Zoology. 2005;83(11):1400–1410.

7. Smith JD. Inaugurating the Study of Animal Metacognition. International Journal of Comparative Psychology. 2010;23(3):401–413.

8. Au WWL, Moore PWB, Pawloski D. Echolocation transmitting beam of the Atlantic bottlenose dolphin. The Journal of the Acoustical Society of America. 1986;80:688–691.

9. Johnson CS. Discussion. In: Busnel RG, editor. Animal Sonar Systems Biology and Bionics. Jouy-en-Josas, France; 1967. p. 384–398.

10. Au WWL. Echolocation in Dolphins. In: Hearing by Whales and Dolphins. New York City, New York: Springer; 2000.

11. Lilly JC, Miller AM. Vocal Exchanges between Dolphins. Science. 1961;134(3493):1873–1876.

12. Dreher JJ. Linguistic Considerations of Porpoise Sounds. The Journal of the Acoustical Society of America. 1961;33(12):1799–1800.

13. McCowan B, Hanser SF, Doyle LR. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. Animal Behaviour. 1998;57(2):409–419.

14. Caldwell MC, Caldwell DK. Individualized Whistle Contours in Bottlenosed Dolphins (*Tursiops truncatus*). Nature. 1965;207(1):434–435.

15. Caldwell MC, Caldwell DK, Tyack PL. Review of the signature-whistle-hypothesis for the Atlantic bottlenose dolphin. In: Leatherwood S, Reeves RR, editors. The Bottlenose Dolphin. San Diego; 1990. p. 199–234.

16. Sayigh LS, Esch HC, Wells RS, Janik VM. Facts about signature whistles of bottlenose dolphins, *Tursiops truncatus*. Animal Behaviour. 2007;74(6):1631–1642.

17. Janik VM, Sayigh LS. Communication in bottlenose dolphins: 50 years of signature whistle research. Journal of Comparative Physiology A. 2013;199(6):479–489.

18. Tyack PL. Whistle Repertoires of Two Bottlenosed Dolphins, *Tursiops truncatus*: Mimicry of Signature Whistles? Behavioral Ecology and Sociobiology. 1986;18(4):251–257.

19. McCowan B. A New Quantitative Technique for Categorizing Whistles Using Simulated Signals and Whistles from Captive Bottlenose Dolphins (Delphinidae, *Tursiops truncatus*). Ethology. 1995;100:177–193.

20. McCowan B, Reiss D. Whistle Contour Development in Captive-Born Infant Dolphins (*Tursiops truncatus*): Role of Learning. Journal of Comparative Psychology. 1995;109(3):242–260.

21. McCowan B, Reiss D. Quantitative Comparison of Whistle Repertoires from Captive Adult Bottlenose Dolphins (Delphinidae, *Tursiops truncatus*): a Re-evaluation of the Signature Whistle Hypothesis. Ethology. 1995;100:194–209.

22. McCowan B, Reiss D, Gubbins C. Social familiarity influences whistle acoustic structure in adult female bottlenose dolphins (*Tursiops truncatus*). Aquatic Mammals. 1998;24(1):27–40.

23. McCowan B, Reiss D. The fallacy of 'signature whistles' in bottlenose dolphins: a comparative perspective of 'signature information' in animal vocalizations. Animal Behaviour. 2001;62(6):1151–1162.

24. Janik VM, King SL, Sayigh LS, Wells RS. Identifying signature whistles from recordings of groups of unrestrained bottlenose dolphins (*Tursiops truncatus*). Marine Mammal Science. 2013;29(1):109–122.

25. Tyack PL. An optical telemetry device to identify which dolphin produces a sound. The Journal of the Acoustical Society of America. 1985;78(5):1892–1895.

26. Watwood SL, Owen ECG, Tyack PL, Wells RS. Signature whistle use by temporarily restrained and free-swimming bottlenose dolphins, Tursiops truncatus. Animal Behaviour. 2005;69(6):1373–1386.

27. Akamatsu T, Wang D, Wang K, Naito Y. A method for individual identification of echolocation signals in free-ranging finless porpoises carrying data loggers. The Journal of the Acoustical Society of America. 2000;108(3):1353–5.

28. Watkins WA, Schevill WE. Listening to Hawaiian Spinner Porpoises, *Stenella* Cf. *Longirostris*, with a Three-Dimensional Hydrophone Array. Journal of Mammalogy. 1974;55(2):319–328.

29. Bell BM, Ewart TE. Separating Multipaths by Global Optimization of a Multidimensional Matched Filter. IEEE Transactions on Acoustic, Speech, and Signal Processing. 1986;ASSP-34(5):1029–1036.

30. Freitag LE, Tyack PL. Passive acoustic localization of the Atlantic bottlenose dolphin using whistles and echolocation clicks. The Journal of the Acoustical Society of America. 1993;93(4):2197–2205.

31. Janik VM, Thompson M. A Two-Dimensional Acoustic Localization System for Marine Mammals. Marine Mammal Science. 2000;16(2):437–447.

32. López-Rivas RM, Bazúa-Durán C. Who is whistling? Localizing and identifying phonating dolphins in captivity. Applied Acoustics. 2010;71(11):1057–1062.

33. Spiesberger JL. Linking auto- and cross-correlation functions with correlation equations: Application to estimating the relative travel times and amplitudes of multipath. The Journal of the Acoustical Society of America. 1998;104(1):300–312.

34. DiBiase JH. A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays; 2000.

35. Do H, Silverman HF, Yu Y. A Real-Time SRP-PHAT Source Location Implementation Using Stochastic Region Contract (SRC) on a Large-Aperture Array. Proc ICASSP-2007. 2007; p. I–121 – I–124.

36. Thomas RE, Fristrup KM, Tyack PL. Linking the sounds of dolphins to their locations and behavior using video and multichannel acoustic recordings. The Journal of the Acoustical Society of America. 2002;112(4):1692–1701.

37. Steiner WW. Species-Specific Differences in Pure Tonal Whistle Vocalizations of Five Western North Atlantic Dolphin Species. Behavioral Ecology and Sociobiology. 1981;9(4):241–246.

38. Rendell LE, Matthews JN, Gill A, Gordon JCD, MacDonald DW. Quantitative analysis of tonal calls from five odontocete species, examining interspecific and intraspecific variation. Journal of Zoology, London. 1999;249:403–410.

39. Navidi W, Jr WSM, undefined, Hereman W. Statistical methods in surveying by trilateration. Computational Statistics & Data Analysis. 1998;27:209–227.

40. Knapp CH, Carter C. The Generalized Correlation Method for Estimation of Time Delay. IEEE Transactions on Acoustic, Speech, and Signal Processing. 1976;24(4):320–327.

41. Li X, Deng ZD, Rauchenstein LT, Carlson TJ. Source-localization algorithms and applications using time of arrival and time difference of arrival measurements. Review of Scientific Instruments. 2016;87(4):041502–13.

42. Smith JO, Abel JS. The Spherical Interpolation Method of Source Localization. IEEE Journal of Oceanic Engineering. 1987;OE-12(1):246–252.

43. Zimmer WMX. Passive Acoustic Monitoring of Cetaceans. Cambridge: Cambridge University Press; 2011.

44. Oppenheim AV, Schafer RW, Buck JR. Discrete-Time Signal Processing. Upper Saddler River, NJ; 1999.

45. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review. 2006;26:159–190. doi:10.1007/s10462-007-9052-3.

46. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.

47. Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing. 2016;114:24–31. doi:10.1016/j.isprsjprs.2016.01.011.

48. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.

49. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273–297. doi:10.1007/bf00994018.

50. Fisher RA. The use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 1936;7:179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

51. Wahba G. Spline Models for Observational Data; 1990.

52. Williams CKI. Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond. In: Jordan MI, editor. Learning in Graphical Models. NATO ASI Series (Series D: Behavioural and Social Sciences). vol. 89. Springer, Dordrecht; 1998.

53. Del Grosso VA. New equation for the speed of sound in natural waters (with comparisons to other equations). The Journal of the Acoustical Society of America. 1974;56(4):1084–1091.

54. Fernandez S, Hohn AA. Age, growth, and calving season of bottlenose dolphins, Tursiops truncatus, off coastal Texas. Fishery Bulletin - National Oceanic and Atmospheric Administration. 1998;(96):357–365.