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 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 SVM, and Gaussian process models. In our 14-point setup, we achieve perfect classification accuracy and high (3.22 ± 2.63 feet) regression accuracy with less than 10,000 features, suggesting an upgrade in accuracy and computational efficiency to the whole-pool-sampling SRP-PHAT method that is the only competitive alternative at present, apart from tag-based methods.

Author summary

The common bottlenose dolphin (*Tursiops truncatus*) has attracted attention as a distinctly nonhuman, yet intelligent and social species that may be capable of complex communication. Despite the great interest in probing the "vocal" interactions of socializing dolphins to evaluate this possibility, a prerequisite to any rigorous attempt is the matching of vocalizations with their corresponding vocalizers. At present, no reliable method exists for consistently performing this matching, particularly over long periods of time and with complete information about the physical condition of all vocalizers. In this study, we propose the first machine learning-based method for accomplishing sound localization – the primary step of sound attribution – of dolphin-like vocalizations in a dolphin pool. On our sample data, this method proves extremely accurate within the body length of a dolphin, and is suggestive of greater practical reliability than other available methods.

Introduction

Dolphin communication research is an active period of growth. Many researchers expect to find significant communicative capacity in dolphins given their complex social structure [?, ?, ?], advanced cognition including the capacity for mirror self-recognition [?], culturally transmitted tool-use and other behaviors [?], varied and adaptive foraging strategies [?], and their capacity for metacognition [?]. Moreover, given dolphins' well-studied acoustic sensitivity and echolocation ability [?, ?, ?], some researchers have speculated that dolphin vocal communication might share properties with human languages [?, ?, ?]. However, there is an insufficiency of work in this area to make significant 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 [?, ?, ?] that may convey an individual's identity to conspecifics [?, ?] and that can be mimicked, potentially to gain conspecifics' attention [?].

Signature whistle studies aside, most studies of bottlenose dolphin calls concern group-wide repertoires of whistles and other, pulse-form call types [?, ?, ?, ?, ?]; there is only 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 [?, ?]. 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 [?, ?, ?], 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. On the other hand, a consistently reliable implementation of the array/camera approach to dolphin whistles has not been achieved, even if it has been achieved for dolphin clicks [?]. 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 [?, ?, ?, ?]. 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 [?]. Among all the methods attempted, one, termed Steer Response Power

Phase Transform (SRP-PHAT), has achieved more success than others (about 40% recall of caller identity), however it relies on a computationally exhausting sampling of all space and has not yet been tested in highly reverberant space [**?**].

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 eventually find that a linear classification model that achieves similar results, as well as a regression model that achieves dolphin-length accuracy. The latter two models rely on tight feature sets containing less than 10,000 features to locate a single whistle, and even with preprocessing avoid the computational burden of the full-space cross-correlation sampling required by SRP-PHAT.

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'-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'x4.25' 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.

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 by field researchers to classify dolphin whistles [**?**, **?**]. In time-frequency space, the sounds were functionally described as either sinusoids or pseudo-sinusoids of the form $\frac{arcsin(m \cdot sin(2\pi f))}{arcsin(2\pi f)}$, the latter class possessing harmonic-like stacking, similar to real whistles. Waveforms were obtained for the desired time-frequency traces by a standard process of integrating instantaneous frequency with respect to time, modified by window functions to ensure realistic onset and decay rates and generally to ensure good behavior, and played in Matlab through a MOTU 8M audio interface at calibrated volumes and a rate of 192 kHz.

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 feet and 18 feet deep. Approximate surface positions are shown in Fig **??**; the difference between adjacent horizontal and vertical positions was 10-15 feet. 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 in by as much as a few feet in arbitrary direction during calibration. These assistants were also used handheld Bosch 225 ft. Laser Measure devices to determine the device's

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 (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.

distance from their reference points (several measurements were taken for each location), and through a triangulation process 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 after 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.

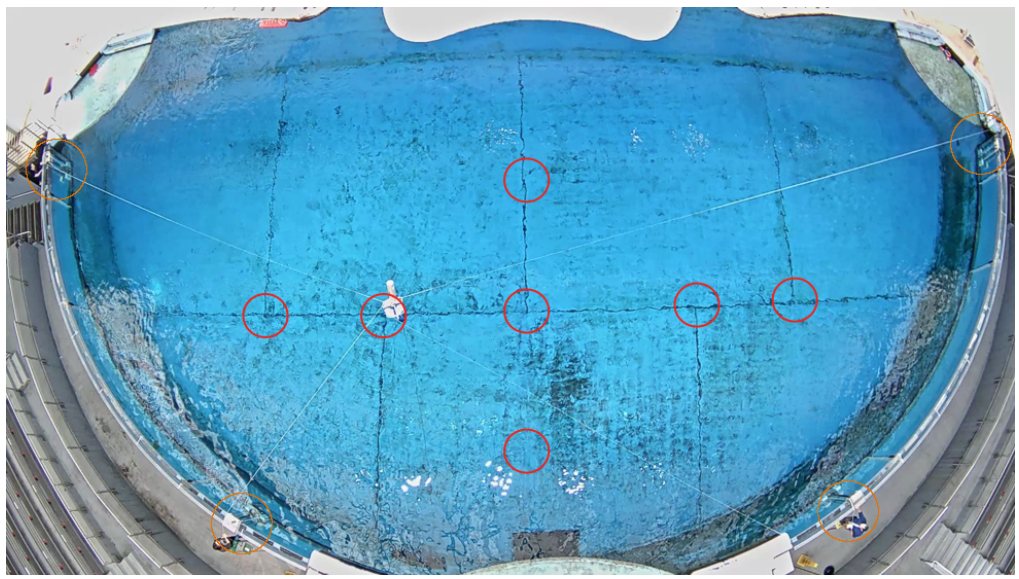

Fig 1. Pool test point and hydrophone array layout. The National Aquarium Exhibit Pool (EP) is shown. Circled in red are the approximate surface projections of the fourteen points at which sounds were played. Circled in orange are the four hydrophone arrays, each containing four hydrophones.

Recording System

Acoustic and visual data was 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 a central AXIS P1435-LE camera, managed by Xeoma surveillance software was used. The four arrays were spaced approximately equally around 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. During this test, sounds were collected at 192 kHz by two networked MOTU 8M into the Audacity AUP sound format, to avoid the size limitations of standard audio formats – this system was also used for playing the sounds. Standard passive system operation was managed by Matlab scripts recording to successive WAV files; for consistency Matlab was also used for most data management and handling. Data is available at 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 the region in which it was played. A random 10% of sounds were set aside for final testing, sinusoids and pseudo-sinusoids with the same parameters grouped together.

Each sound was initially digested into 1,333,877 continuous, numerical features: 120 *time-difference-of-arrivals* (TDOA's) obtained using the Generalized Cross-Correlation Phase Transform (GCC-PHAT) method [**?**], which in currently unpublished work we found to be most successful among correlation-based methods for obtaining whistle TDOA's (if still too imprecise for achieving reliable sound localization with Spherical Interpolation), 6601 x 136 standard normalized cross-correlations (truncated in time to discard physically impossible delay peaks), and 27,126 x 12 truncated Fourier transforms. Preliminary analysis found the Fourier transform features to be completely disregarded during classification, so they were discarded, leaving 897,871 features.

Given our computational resources, this feature set remained too large to accomodate most classifiers. A notable exception was random forests, which were suitable not only for classification – being powerful nonlinear classifiers with built-in resistance to overfitting – but for feature reduction, via the delta error metric that can serve as a measure of feature importance. We grew a Breiman random forest composed of CART decision trees on the training data; each tree was trained on a random subset of ˜75% of the training samples using a random ˜$\sqrt{897,871}$ feature subset. Out-of-bag (OOB) error was used for validation.

We subsequently used delta error as a measure of feature importance, both to examine the selected features for physical significance – recall that cross-correlation features correspond to pairs of sensors – and to obtain a reduced feature set appropriate for training additional models. On the reduced feature set, we considered a basic decision tree, a linear and quadratic SVM, and linear discriminant analysis. We also considered Gaussian process regression (also termed *kriging*) – a nontraditional, nonparametric method of regression that could accommodate our under-constrained data.

We then assessed our ability to locate sounds not originating on the training/testing grid using Gaussian process models. We trained models on training data exclusive of a single grid point, and then evaluated the regression's performance predicting the coordinates of test sounds from that point. We repeated this process for all grid points and generated statistics.

Lastly, we obtained a minimal, nearly sufficient feature set by training a single sparse 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 delta error) back to the sensor and array pairs that they represented.

Results

The random forest trained on the full feature set, as specified 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 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 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 96.9% cross-validation and 97.75% test accuracy, and the linear SVM achieved 100.0% cross-validation accuracy and 99.44% test accuracy.

On the reduced feature set we also performed Gaussian process regression (kriging), generating one model for each spatial dimension. The predicted locations of testing samples for a subset of test locations are plotted in Fig **??**. The calculated error was 3.22 ± 2.63 feet.

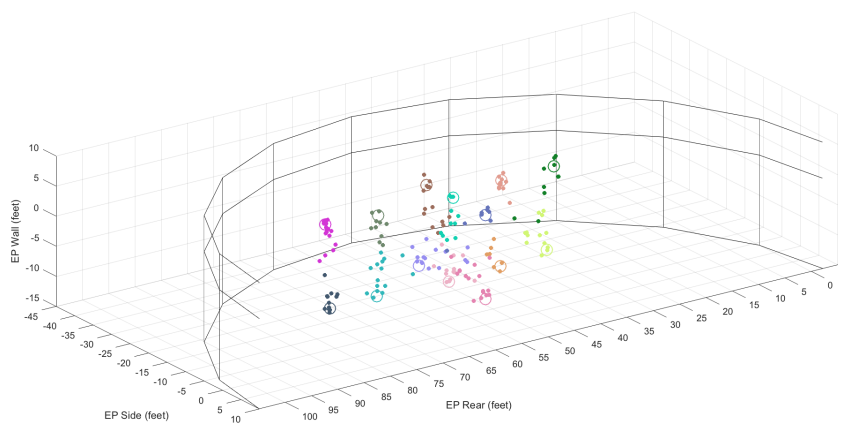

Fig 2. Predictions of test sound locations by Gaussian process regressors. The half-cylindrical National Aquarium EP is depicted. Large unfilled circles indicate the true coordinates of the test sounds; each has a unique color. Small filled circles indicate the test sound coordinates predicted by Gaussian process regression, colors matching their respective true coordinates.

When the Gaussian process regression models were prompted to predict the coordinates of test sounds from single grid points from which they received no training data, the error was significantly greater, at 10.73 ± 2.45 feet. We also calculated the error separately in each dimension: referring to Fig **??**, in the horizontal direction the error was 2.45 ± 2.26, in the vertical direction the error was 3.11 ± 3.32, and into the figure the error was 8.95 ± 3.07; based on one-sided, unpaired t-testing, the differences between all three are statistically significant (<0.05).

Next, we trained a single, sparse decision tree on the full training set. The severe

feature reduction left 22 features. While the decision tree achieved only 96.63% accuracy on the test set, a random forest trained on the same features achieved 98.88% test accuracy. Thus, we considered this feature set both sufficient and sparse enough to meaningfully ask these features tell us about sensor-sensor pairs a classifier might emphasize. The delta error was summed across hydrophone and array pairs, visualized in Fig **??**. Overall, we note that, directly or indirectly, features representing all pairs of hydrophone arrays are utilized.

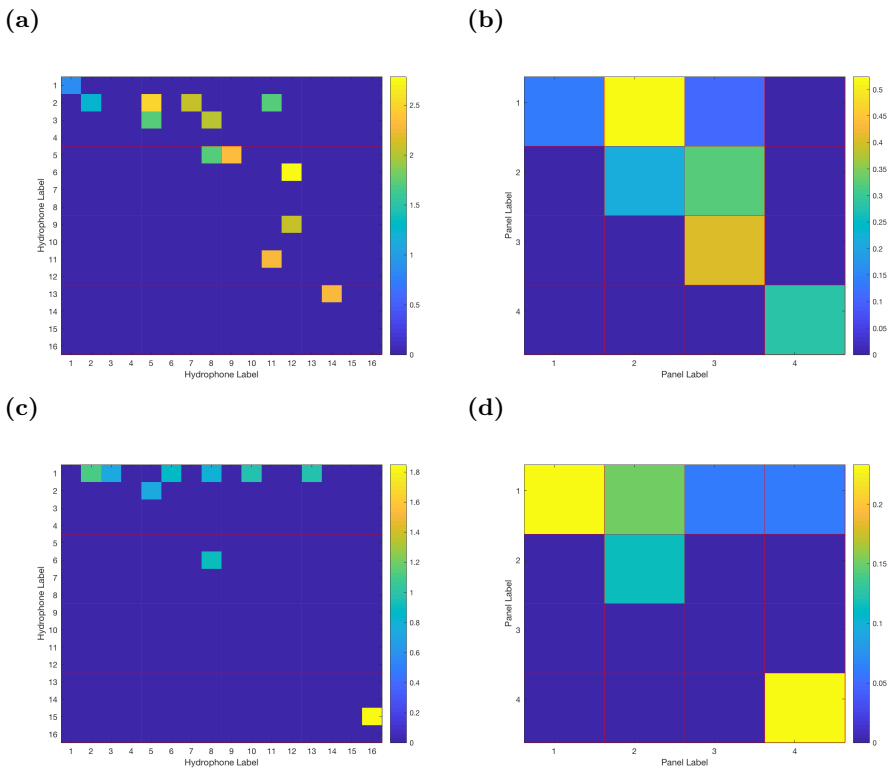

Fig 3. Minimal Cross-Hydrophone and Cross-Array Feature Importances. Aggregate feature importance values for a minimal set of 22 features across hydrophone and hydrophone array pairs are formed by summing the importances (delta error) of features across hydrophone pairs, and subsequently averaging hydrophone pair sums across 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-array importances for cross-correlation features. Facing the pool from the panels increase in number from left to right. C: Cross-hydrophone importances for TDOA features. D: Cross-array importances for TDOA features.

Discussion

We provide a proof of concept that sound source localization of bottlenose whistles can be achieved implicitly as a classification task and explicitly as a regression task in a highly reverberant, half-cylindrical aquatic environment. We began with 127 unique tonal sounds played at 14 positions in the primary dolphin pool at the National Aquarium, recorded with four four-hydrophone arrays. First, we showed that a random forest classifier with less than 200 trees can achieve 100% testing accuracy using 6,788 of 897,871 features, including TDOA's obtained from GCC-PHAT and normalized

cross-correlations between all pairs of sensors. We then showed that linear discriminant analysis and a quadratic SVM can achieve the same results on the reduced feature set. If the linear model in particular were to remain valid 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, a concern even for a classifier trained on a finer grid, we note that our classifiers' success was achieved despite the few-foot drift of the speaker during play-time; this 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 (TDOA's, cross-correlations) that are generally expected to vary continuously in value across space is reassuring. Nevertheless, this question does warrant further investigation, perhaps with deliberately, faster moving sources.

We more suitably addressed the question of off-grid prediction for Gaussian process regression, which was also quite successful when trained on the full training data set, achieving test error of 3.22 ± 2.63 feet – less than the expected length of an adult common bottlenose dolphin [**?**]. Not only is regression inherently suited to interpolation, but it was straightforward to assess regression's performance on test data from grid points excluded during training. While the regression's overall performance on novel points was not satisfactory, admitting error larger than average dolphin length at 10.73 +/- 2.45 feet, when we decomposed the error into three dimensions (2.45 +/- 2.26 feet in X, 3.11 +/- 3.32 feet in Y, 8.95 +/- 3.07 feet in Z, all statistically different under one-sided t-testing), we saw that much of the error originated in the direction of pool depth. This is unsurprising given that all data are evenly distributed between only two depths; we would expect a significant boost to performance by introducing additional depth(s) to the data.

It also remains unclear to what extent sounds outside the training set, specifically real dolphin whistles, are properly assigned. A test of this would require a large set of dolphin whistles played at known locations in the pool, which would do not possess at present. However, even were an evaluation of real dolphin whistles to fail, we note that in general captive dolphins' "vocabularies" tend to be limited – groups seem to possess less than 100 unique types [**?**] – 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 in whistle type space.

We also 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 much sound amplitude information was removed in the process of feature creation, this suggests that the decision tree and random forest 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 decision making of the models ultimately remains unclear.

Overall, we feel this study offers a strong demonstration that machine-learning methods are suitable to solving the problem of sound-localization for tonal whistles in highly reverberant aquaria. Were the performance of the methods presented here extend to a finer grid – which was not and will not be feasible in our own work at the National Aquarium – they would constitute the most accurate methods to sound source localization of dolphin-whistle-like sounds in a highly reverberant environment yet proposed that avoid the need for tagging; currently, successful localization of whistles in similar environments is no greater than 70% [**?**, **?**, **?**, **?**]. Moreover, as these methods do

not require the computation of cross-correlations across the whole sample space, we expect these methods to be less computationally expensive than SRP-PHAT, the primary alternative. With a set of four permanent hydrophone arrays surrounding a subject enclosure, automated overhead tracking, and suitable training set, this method may allow for the creation of a high-fidelity record of dolphin exchanges suitable to statistical analysis in many settings.

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.