Linking the sounds of dolphins to their locations and behavior using video and multichannel acoustic recordings

Rebecca E. Thomas a)  
Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543

Kurt M. Fristrup  
Bioacoustics Research Program, Cornell Laboratory of Ornithology, Ithaca, New York 14850

Peter L. Tyack  
Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543

(Received 1 May 2002; revised 3 May 2002; accepted 22 May 2002)

It is difficult to attribute underwater animal sounds to the individuals producing them. This paper presents a system developed to solve this problem for dolphins by linking acoustic locations of the sounds of captive bottlenose dolphins with an overhead video image. A time-delay beamforming algorithm localized dolphin sounds obtained from an array of hydrophones dispersed around a lagoon. The localized positions of vocalizing dolphins were projected onto video images. The performance of the system was measured for artificial calibration signals as well as for dolphin sounds. The performance of the system for calibration signals was analyzed in terms of acoustic localization error, video projection error, and combined acoustic localization and video error. The 95% confidence bounds for these were 1.5, 2.1, and 2.1 m, respectively. Performance of the system was analyzed for three types of dolphin sounds: echolocation clicks, whistles, and burst-pulsed sounds. The mean errors for these were 0.8, 1.3, and 1.3 m, respectively. The 95% confidence bound for all vocalizations was 2.8 m, roughly the length of an adult bottlenose dolphin. This system represents a significant advance for studying the function of vocalizations of marine animals in relation to their context, as the sounds can be identified to the vocalizing dolphin and linked to its concurrent behavior. © 2002 Acoustical Society of America. [DOI: 10.1121/1.1494805]

PACS numbers: 43.80.Ka, 43.30.Sf, 43.80.Jz [WA]

I. INTRODUCTION

Studies of animal communication and social behavior ideally use methods where signals and actions can be associated with individuals. Only under these conditions can one fully study social interactions involving signal and response. Studies of marine mammal acoustic repertoires have been hampered by the difficulty of identifying which animal produces each sound. Marine mammals generally do not open their mouths when they vocalize underwater, nor do they regularly release bubbles. Humans cannot routinely use their auditory capabilities to localize underwater sounds directly. However, these technical difficulties need not inhibit research on marine mammal acoustic behavior. Acoustic communication is especially important for these species, because they are often out of sight of each other, but can remain in acoustic contact at long ranges.

Several solutions have been proposed and used to solve the problem of identifying the vocalizing animal, including isolating animals (Lilly and Miller, 1961; Caldwell et al., 1990; Sayigh et al., 1990), attaching tags to animals (Evans and Sutherland, 1963; Tyack, 1991; Tyack and Recchia 1991; Nowacek et al., 1998), and identifying the vocalizing animal using bubblestreams (Dahlheim and Awbrey, 1982; McCowan, 1995; McCowan and Reiss, 1995; Herzing, 1996). These methods have produced valuable results, but in the case of isolation or tags can yield biased data by altering the behavior of the animal. Bubblestream emissions occur relatively rarely (Caldwell et al., 1990; Fripp, 1999), and themselves may function as a behavioral display. Also, sounds produced with bubblestreams may not be a random subsample of all the sounds emitted (Fripp, 1999).

Methods to locate the source of a sound and to link this acoustic location to visual images promise the ability to obtain unbiased results without modifying the behavior of the study animals. However, the first part of this method, acoustic localization, has proven difficult to achieve for marine mammals, and few implementations have been linked to visual data to allow identification of the vocalizer. Linear arrays for sound beamforming and dispersed arrays for localization are widely used techniques for determining the location or direction of a sound source. Beamforming using towed linear arrays is often suitable for field work at sea with free-ranging animals, because the array can be easily towed while maintaining fixed distances between hydrophones. When the length of the array is small relative to the distance to the sound source, only the bearing to the source is obtained. This bearing indicates angle relative to the line of the array, so the locus of possible locations is a cone. With careful positioning of the array relative to the animals, focal follows of single animals can be performed. For example, Miller and Tyack (1998) followed free-ranging killer whales while beamforming their calls using a small towed array.

Dispersed array localization has been used more often...
than linear beamforming to study cetacean sounds. Watkins and Schevill (1972) used a drifting three-dimensional array to determine \( x - y \) positions of finback whales, right whales, and white-beaked dolphins. Although the technique used passive acoustic localization, they needed to produce intermittent pings to determine hydrophone locations. The calibration pings can interfere with the behavior being observed; they had the effect of temporarily halting sperm whale sound production (Watkins and Schevill, 1975).

Clark et al. (1986) used a fixed array to study bowhead whales during their spring migration off Pt. Barrow, Alaska. Placing the hydrophones at fixed locations in the ice eliminated the need for intermittent pings. Spiesberger and Fris-trup (1990) developed a method combining passive localization of vocalizing animals with acoustic tomography. Their technique allowed for localization of vocalizing animals in addition to construction of maps of sound speed and wind (or current) fields. However, their technique has yet to be implemented in its entirety for marine mammals. Freitag and Ty-ack (1993) demonstrated the feasibility of acoustic localization of bottlenose dolphin sounds in a captive environment. Although successful, their study demonstrated that reverberation presents a problem for localization in captive environments, and they did not attempt to link acoustic locations to visual. Janik (2000) demonstrated the feasibility of a two-dimensional acoustic localization system for determining positions of vocalizing bottlenose dolphins in the wild. Reverberation did not appear to be a significant problem in this study, most likely due to the fact that the fixed array was located in a large channel rather than in a reverberant captive pool.

A few methods have endeavored to combine acoustic location or bearing data with concurrent video recordings to obtain the identity of a vocalizing animal. Dudzinski et al. (1995) developed a video/acoustic system for underwater recording of dolphin interactions. Analysis of the data used a human observer to aurally determine if the sound came from the right, left, or center of the system, but did not allow for pinpointing the sound source on the video image. Brensing et al. (2001) designed a system of two pairs of closely spaced hydrophones to localize dolphins within a pool. They had a video component to the system, but did not show video overlay.

Few, if any, studies have been able to link the locations of sounds to a detailed record of behavior. The technique presented here projects acoustic localization results onto a video recording of dolphin behavior. An observer watching the video can see behaviors happening, hear the concurrent vocalizations, and also see on the screen which individual produced the vocalization. Having a record of which animal produced which sound, and what it was doing when it produced the sound, is important for understanding the function of communicative sounds. The identity of the calling animal can be determined, and the behavioral context of the sound can be revisited. A number of research directions that require knowledge of the identity of the vocalizing animal can be explored, as well as data archived for future reference and reanalysis.

The focus of this study was not a demonstration of the merits of a particular localization algorithm, but a demonstration that acoustic localization data can be combined with video analysis methods to successfully associate dolphin sounds with the individuals that produce them. Video recordings have long been used to supplement visual observations, and to provide an archival record of behavior. The capacity to augment the video imagery and soundtrack with a visual indication of the animal producing each sound magnifies the scientific and archival value of these recordings. Acoustic localization software provided by the Cornell Bioacoustics Research Program was used here because it was available for this study and easily adapted to its needs. Any other localization package that yielded equivalent or better localization accuracy could be substituted in its place, without substantially affecting the conclusions of this work.

II. METHODS

A. Study site

This research was performed at Dolphin Quest Bermuda’s interim facility at the Maritime Museum in the Naval Dockyards on Ireland Island, Bermuda in the fall of 1999. At this time, the dolphins’ social group was composed of two mother–calf pairs, a juvenile male and a juvenile female, for a total of six animals. The lagoon facility measured roughly 30 meters by 45 meters. The sides of the lagoon were composed of irregular limestone bricks, covered with algae and other organisms. The ramparts of the Naval Dockyards directly abutted the lagoon on one side. Eight hydrophones were placed around three sides of the lagoon, and the video camera was placed on the rampart next to the lagoon, approximately 9 m high (Fig. 1). The lagoon was connected to the ocean via a short channel with a gate on the lagoon side, allowing the water level in the lagoon to change with the local tides. The temperature of the seawater at a depth of approximately 0.5 m was recorded each day of observation from a mercury thermometer. The speed of sound was calculated from temperature and salinity from standard equations (Urick, 1982). Because salinity and depth have less significant effects than temperature, salinity was assumed to be constant at 36.6 ppt, and the pressure effect on the speed of sound was assumed to be nil, the same as assuming a depth of 0 m for the path from the animal in the shallow lagoon to the hydrophone.

B. Acoustic localization

Eight hydrophones (High Tech Inc. HTI-94-SSQ) were placed around three of the four sides of the lagoon. The fourth side of the lagoon directly abutted the ramparts of the Dockyard fort, with no convenient spot to anchor a hydrophone. An eight-channel TASCAM DA-88 multitrack recorder digitally sampled signals simultaneously from all the hydrophones and then saved the digitized signals to tape. The TASCAM DA-88 sampled at 48 kHz and had a flat frequency response (±0.5 dB) from 20 Hz to 20 kHz. The frequency response of the hydrophones was 2 Hz to 30 kHz, so the recorded signals were frequency limited by the...
TASCAM recorders to the range of 20 Hz to 20 kHz. Eight channels of 1-s data (Microsoft wave format) occupied 0.73 Mb of disk space.

1. Data acquisition

The digital data on tape were transferred directly to a computer running Microsoft Windows using the “Translator Plus” digital audio format converter (Spectral Inc., Woodinville, WA) and STUDIO TRACKS XP software.

2. Signal selection process

A suite of MATLAB (MathWorks, Natick, MA) programs was used for performing the various aspects of signal selection. An automated energy detector was used to detect, extract, and save sounds above a preset energy threshold (modified from Fripp et al., 1997). This enabled more extensive screening of the multichannel data than would be feasible using people to screen sounds, while taking advantage of human judgment to edit the relatively compact set of detections. The operator viewed eight-channel spectrograms of the extracted sound cuts. The operator would discard the cut if it contained only noise or if the dolphin vocalization was excessively contaminated by transient signals (e.g., snapping shrimp clicks). Otherwise, the operator would highlight the sounds in time and frequency. These time and frequency parameters were saved for the localization step.

3. Localization algorithm

A brief summary of the algorithm used in this study is provided here, but any acoustic localization algorithm that provided comparable results could be substituted. The algorithm chose points in space and calculated the time delays between phones that would result if the sound came from that point in space. The value of each cross-correlation function at the corresponding time delay was determined. The sum of the cross-correlation values is proportional to the beamformed energy corresponding to the designated point, and the search algorithm looked for the point that maximized the summed correlation values. [See Johnson and Dudgeon (1993) for a detailed description of beamforming.] This contrasts to alternative schemes in which the time delays associated with the peaks of each cross-correlation function are taken, and a least-squares fit is performed to obtain a location estimate from these time delays (Spiesberger and Fristrup, 1990). The beamforming algorithm produces accurate locations in acoustic scenarios (overlapping sound transients, multipath) that can present difficulties for algorithms based on peak picking. Details of the algorithm used in this study can be found in Fristrup and Dunsmore (unpublished).

As the lagoon was not very deep, the localization was performed only in the $x$ and $y$ dimensions. A typical localization of a dolphin vocalization took about 22 s on a Pentium III 700 MHz computer. The algorithm gave two outputs: the $x$–$y$ localization coordinates and a term that quantified the quality of the localization. A heuristic threshold for localization quality was implemented to remove outliers caused by poor localization quality.

4. Calibration of acoustic localization system: Locations of hydrophones

Precise hydrophone locations are needed for accurate localization. Approximate interhydrophone distances were measured using a tape measure to obtain approximate hydrophone positions. To obtain more precise locations, a calibration signal set was played at each hydrophone (ostensibly yielding a time delay of 0 for that hydrophone). The calibration signal set consisted of a set of five pseudorandom sequences and a set of four upsweeps (4–7 kHz). The pseudorandom sequences, also known as Barker codes (Barker, 1953), are binary sequences with good autocorrelation properties; autocorrelated Barker codes have high main lobes and low sidelobes. The sequence used for this study was an 11-bit Barker code ($1,-1,1,1,-1,1,1,1,-1,1,-1$), set so each
bit lasted a little over 0.01 s, with the total 11-bit sequence lasting about 0.125 s, or 1/8 s. The upsweeps each lasted about 1 s.

Received calibration signals were saved as multichannel sound cuts. Localization was performed on each sound cut to obtain a series of estimates of hydrophone locations. Interphone time delays were calculated from these estimated locations and also saved. Outlier localizations were removed, yielding a set of localizations and corresponding sets of calculated time delays for each of the eight hydrophones. The time delays were converted into distances using the speed of sound, yielding eight sets of interhydrophone distances. These sets were condensed into the matrix $D_{\text{exp}}$. The matrix $D_{\text{exp}}$ is an $8 \times 8$ matrix of interhydrophone distances. To determine the best-fit $x$–$y$ positions that would yield these distances, a multidimensional unconstrained nonlinear minimization [Nelder-Mead (Press et al., 1992)] was performed in MATLAB. This type of minimization requires no assumptions about the function to be minimized. The function to be minimized was $\|D_{\text{pos}} - D_{\text{exp}}\|$, minimizing over $D_{\text{pos}}$. $D_{\text{pos}}$ is the set of interposition distances between a set of eight $x$–$y$ positions. The $D_{\text{pos}}$ which minimizes $\|D_{\text{pos}} - D_{\text{exp}}\|$ should be the distances between the correct hydrophone positions. In the minimization, $D_{\text{pos}}$ was initialized as the set of distances between the tape-measured hydrophone positions. For reference in the real world, one hydrophone was arbitrarily set to (0,0), and the $x$ coordinate of an adjacent hydrophone set to 0. The rest of the hydrophone coordinates were calculated from these two hydrophones and the best-fit $D_{\text{pos}}$ resulting from the minimization.

5. Calibration of the acoustic localization system: Accuracy and precision

In order to determine the accuracy of the acoustic localization system, calibration transects were performed along a dock that separated the lagoon into two parts as well as along a side wall of the lagoon (Fig. 1). The same calibration signal as used for calibrating the hydrophone locations was played at set locations along three transect lines, transect 1, transect 2, and transect 3. The lines were measured using a tape measure. The error was calculated as the distance from mean localized position to actual position (calculated using tape measured lines).

C. Video system

The video camera was placed at a height of approximately 9 m, on the side of the lagoon abutting the ramparts. The camera was placed in approximately the same position each day. A wide-angle lens (KVC-05 0.5 x) and polarizing filter were attached to the 3 ccd digital video camcorder (SONY TRV-900) used for recording video. At the beginning of each recording session, the timestamps on the video and audio recorders were synchronized to a digital chronometer. Once the camera was in place for the day, it was not moved. Video recordings were later imported into a PC computer using a DVRaptor card and ADOBE PREMIERE software.

There are two parts to calibrating a camera image. The first is the intrinsic calibration, in which the internal geometric and optical workings of the camera are calibrated. As long as the focal length of the lens is kept constant, the intrinsic parameters will not vary with recording of different images. The second part of calibrating a camera image is the extrinsic calibration. In this calibration, parameters such as the distance to objects in the world frame coordinate system, rotation, etc. are calibrated. For our system we always used the same configuration of lenses, filters, and zoom; therefore, the intrinsic calibration only needed to be performed once. The camera was installed each morning and then not moved for the day, so the extrinsic calibration needed to be performed once for each day of observation. A camera calibration toolbox for MATLAB obtained from Cal-Tech (Bouguet, 2000) was used for both the intrinsic and extrinsic calibrations. The projection feature of the calibration toolbox transformed the real-world coordinates of localizations into video space.

1. Intrinsic calibration

The camera calibration toolbox included a checkerboard that we attached to stiff matting board. The checkerboard was videotaped and 20 images representing a diversity of video angles and distances were used for the intrinsic calibration. The toolbox, with some initial user input, automatically finds the corners of the checkerboard boxes and performs the intrinsic calibration. Parameters from this calibration were saved in a file for later use.

2. Extrinsic calibration

After solving for the intrinsic parameters, the extrinsic parameters were calculated by relating known points in the world coordinate system with their pixel analogs in the video image. Video frame pixel coordinates were obtained by plotting the image in MATLAB and using the mouse to click on locations in the image. We chose five easily recognizable locations to be used for daily calibration. The daily calibration yielded the extrinsic calibration parameters for that day.

One way of measuring the error of the video imaging system is in terms of the projection error, which is the difference between the actual position and the projected position. Using a tape measure we determined the world coordinates at 100 set locations around the lagoon (videotaped on 1 November 1999). The real world positions of these coordinates were termed $X_{\text{R}}$ (in meters). These positions were projected (from the extrinsic parameters calculated from the five specified points) into video frame coordinates. We called these projected coordinates $X_{\text{V}}$ (in pixels). From the video image we determined the pixel location of the $X_{\text{R}}$ coordinates by plotting the image in MATLAB and using the mouse to click on each location in the video frame. We termed these video frame coordinates $X_{\text{mi}}$. We backprojected $X_{\text{mi}}$ to real-world coordinates, $X_{\text{mi}}$ (in meters) for comparison with $X_{\text{R}}$. The projection error in terms of real-world coordinates is calculated as $E_{i} = X_{\text{R}} - X_{\text{mi}}$. The error in terms of pixels is $e_{i} = x_{\text{ti}} - x_{\text{mi}}$. 


Thomas et al.: Dolphin localization and video 1695
D. Method for fusion of acoustic localization and video imaging

Projected sound-source positions derived from acoustic localization were plotted in MATLAB against a blue background and exported as numbered picture files. The localizations were projected onto the water's surface, taking into account daily tidal fluctuations. The tidal height was assumed to not change significantly over each 2-h observation period. The picture files were imported into ADOBE PREMIERE as an animated video clip. The video sequence from the same time sequence as the localizations was also imported. The video sequence and animated video clip were overlaid, setting the blue background in the animated localization clip to transparent. This results in the localized positions appearing as bull's-eyes on the video clip [see Fig. 7(a) in Sec. III].

A set of sound playbacks across the lagoon (transect 1) was videotaped to compare positions obtained from both acoustic localization and video imaging. Pixel positions of the acoustic source were measured directly from the video images by clicking the mouse on the position of the source where the cable exited the water. The difference in localized position and video frame position was calculated for each playback location to obtain the error of the overall localization/video system for calibration sounds.

Dolphin localization: A set of 222 sounds from dolphins swimming freely in the lagoon was localized. Of the subset of these sounds that could be localized to a dolphin or a group of dolphins, ten sounds of each of three sound types were randomly chosen. The three sound types were echolocation clicks, burst-pulsed sounds, and whistles. For each vocalization, the error was calculated as the distance between the localized position and the nearest dolphin. When possible, the blowhole was used as the reference point for the dolphin. If more than one dolphin was in the immediate area, the error was calculated as the distance between the localized position and the mean position of the two nearest dolphins. Instances when more than two dolphins were in the immediate area were not used.

III. RESULTS

This paper presents the results of a combined acoustic localization and video imaging system. The results are broken down into three categories: acoustic localization results, video imaging results, and combined acoustic localization and video imaging results. The results of the combined system are broken down into two categories: results from the artificial sound source, and results from dolphins.

A. Error of acoustic localization system

The error and standard deviation of the acoustic localization system were determined by comparing known source playback locations along the transect lines to the correspond-

![Figure 2](image2.png)

**FIG. 2.** Histograms of error and standard deviation of the acoustic localization system.

![Figure 3](image3.png)

**FIG. 3.** (a) Error in localization plotted versus distance from source position to centroid of array. (b) Error in localization plotted versus distance from nearest wall.

D. Thomas et al.: Dolphin localization and video
ing positions calculated from acoustic localization. The histo-
grams of error and standard deviation are shown in Fig. 2.

The localization error was less than 1.5 m for 95% of the
measurements (mean error=0.54 meters). Thus, the accuracy
of the localization system can be thought of as being better
than 1.5 m. The mean standard deviation was 0.64 m. Thus,
the precision of localization can be thought of as being two
standard deviations, or 1.28 m.

Figure 3 presents the error as a function of distance from
the centroid of the array as well as of distance from the
nearest wall. The error appeared to increase both with in-
creasing distance from the centroid of the array as well as
with decreasing distance to the nearest wall.

B. Error of video imaging system

The histogram of the video projection error in terms of
pixels is shown in Fig. 4(a). The size of each video image in
pixels was 480×720. Of the error in the x dimension, 95%
was less than 11.6 pixels. This was less than 1.6% percent of
the image size in the x dimension. Of the error in the y
dimension, 95% was less than 6.4 pixels, which was less
than 1.3% of the image size in the y dimension. The mean
error was 6.0 pixels in the x dimension and 2.4 pixels in the
y dimension.

The projection error was also calculated in terms of real-
world coordinates. The histogram of this error is shown in
Fig. 4(b). For ease of visualization, the projection error in
terms of real-world coordinates is plotted at the location of
each coordinate in Fig. 5. Of the projection error, 95% was
less than 0.9 m in the x axis, and 2.0 m in the y axis. The
mean projection error was 0.5 m in the x axis and 0.7 m in
the y axis. The overall 2D projection error was less than 2.1
m for 95% of the measurements. The mean 2D projection
error was 0.9 m.
C. Synthesis of acoustic localization and video

1. Artificial sound source

Video recordings were performed during acoustic calibration transect 1 across the center of the lagoon (shown in Fig. 1). This allowed direct comparison of the projected positions obtained by acoustic localization to positions obtained directly from video images. Figure 6 presents histograms of the error between localized positions and video frame positions in terms of both pixel and real-world coordinates. Of the error measurements in the $x$ and $y$ axes, 95% were less than 10.0 and 9.8 pixels, respectively [Fig. 6(a)]. In real-world coordinates this corresponds to errors less than 0.9 and 2.0 m, respectively [Fig. 6(b)]. The mean errors in the $x$ and $y$ axes were 4.7 and 7.5 pixels, respectively. In real-world coordinates this corresponds to mean errors of 0.4 and 1.7 m in the $x$ and $y$ axes, respectively. The 2D error between the localized and video frame positions was less than 2.1 m for 95% of the measurements [Fig. 6(c)]. The mean 2D error was 1.8 m.

2. Dolphin sounds

The final set of results reports the error of the system with the dolphins as the sound source, matching localized sounds with the video images of the vocalizing dolphins. An example of matching a localized vocalization to a dolphin is shown in Fig. 7(a), with the spectrogram of the localized echolocation click shown in Fig. 7(b). The $x$ axis, $y$ axis, and 2D results from the comparisons of localized positions to video frame positions are shown for each of the three sound types in Table I. The means were calculated for each sound type as well as over all three sound types: 95% of the errors were less than 0.8 m in the $x$ dimension, 2.5 m in the $y$ dimension, and 2.9 m in the combined dimensions (2D). The
mean error of echolocation clicks appears to be lower than that of the other two sound types. However, the difference in error between the sound types cannot be considered to be statistically significant (ANOVA, $p = 0.068$).

In addition to determining the errors between localized positions and video frame positions of vocalizing dolphins, the percent of localized sounds that could be attributed to a dolphin or a group of dolphins was also calculated (Table II). When only one dolphin was in the immediate vicinity of the localization (no other dolphins were within 1.5 m), the vocalization was determined to be from that individual. In some cases, more than one dolphin could be seen within 1.5 m of the localized position. In these instances the vocalization could only be identified as coming from that group of dolphins. In other cases it could not be determined that a dolphin was in the area. However, since in some areas the video image did not extend to the bottom of the lagoon, there were occasions when a dolphin was believed to be in the area, but the presence of the dolphin could not be confirmed. A dolphin might be believed to be in the area if the path between the dolphin’s previous and following surfacings crossed the area around the localized position.

### IV. DISCUSSION

The localization error of 95% of the measurements from the calibration signals was less than 1.5 m. This is less than the average length of an adult bottlenose dolphin (1.9–3.9 m, Read et al., 1993). The precision (2 standard deviations) of the localization of the calibration signals was 1.28 m. In this sense, we should be able to distinguish sound sources that are greater than 1.28 m apart. The error was not constant at all points within the array, but was larger near the wall of the lagoon (Fig. 3). This trend was observed primarily in transect 3. The other transects showed no trend of increasing error with either decreasing distance from the wall or increasing distance from the centroid of the array. These transect points along the wall were also the points furthest from the centroid. Therefore, it was not possible to rigorously separate the effects of decreasing distance from the wall and increasing distance from the centroid. However, the tight linkage between error greater than about 0.75 m and proximity to the wall suggests that the wall may have been the main factor. Reflections of sound off the wall can confuse the localization algorithm. This problem with multipath is common among localization algorithms (e.g., Freitag and Tyack, 1993; discussion in Spiesberger and Fristrup, 1990; Spiesberger, 1999). Confusion is likely to be greatest when the sound originates near the wall and the reflected sound cannot be isolated from the direct path. There are several practical ways to reduce this problem. Localized positions from reflected sounds will appear to come from outside the lagoon, which we know is not possible. Setting more detailed search boundaries for the algorithm might alleviate this problem. Another solution would be to keep the dolphins away from the walls, which in this setting was not practical. Spiesberger (1998, 1999, 2000) presents an algorithm that can deal with some aspects of multipath.

The error of the video system was measured in terms of pixels [Fig. 4(a)] and then projected into meters [Fig. 4(b)]. The absolute pixel error appeared to be the greatest in the $x$ axis. However, the errors in terms of percentage of image size were similar for both the $x$ and $y$ axes. When the error of the video system was projected into meters, the error was greater in the $y$ dimension. Error in the $y$ dimension appeared to increase with increasing $y$ position (Fig. 5), which was also increasing distance from the camera. The lagoon was longer in the $y$ dimension than in the $x$ dimension, so fitting more of the $y$ axis into the image resulted in lower resolution and greater error at the limits. This skewed distribution of error could be improved by placing the video camera higher and at less of an angle to the lagoon (e.g., suspend the camera from an overhead position, as in Nowacek et al., 2000). Another feature of Fig. 4(a) is the bimodality of the $x$ error. Due to practical considerations of accessibility to different parts of the lagoon, the $x$ axis was sampled more heavily at the extremes than at the middle. If one side were better calibrated at the expense of the other side, it would result in the observed bimodality of the $x$ error.

The error of the overall system (Fig. 6) must take into account the errors of both the acoustic localization and video projection components. The 95% error bound calculated from the combined acoustic localization and video transect was 2.1 m. A comparison of this error bound to the 2.1-m error bound for the video calibrations and the 1.5-m error bound from the acoustic transects suggests that the video component of the system dominates the error.

Although the transects using calibration sounds from a mechanical source demonstrate that the system ostensibly had an error less than 2.1 m, the true test of the system is with sounds from the dolphins. The 95% error bound for the dolphin vocalizations was 2.9 m, roughly the length of an adult dolphin’s body. The errors in video position and localized position of dolphin sounds were worse than those from calibration sounds from a mechanical source. This was probably due to a combination of several factors. The first factor

**TABLE I.** Mean error between localized and video frame positions for each of three vocalization types, the mean error over all three vocalization types, and the 95% error limit over all three vocalization types.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>X Dimension (m)</th>
<th>Y Dimension (m)</th>
<th>2D (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burst pulsed sounds</td>
<td>0.33</td>
<td>1.29</td>
<td>1.34</td>
</tr>
<tr>
<td>Echolocation clicks</td>
<td>0.52</td>
<td>0.65</td>
<td>0.83</td>
</tr>
<tr>
<td>Whistles</td>
<td>0.34</td>
<td>1.24</td>
<td>1.28</td>
</tr>
<tr>
<td>Overall mean</td>
<td>0.40</td>
<td>1.06</td>
<td>1.13</td>
</tr>
<tr>
<td>95% error limit</td>
<td>0.75</td>
<td>2.52</td>
<td>2.91</td>
</tr>
</tbody>
</table>

**TABLE II.** Number and percentage of localized vocalizations that could be attributed to a dolphin or group of dolphins.

<table>
<thead>
<tr>
<th></th>
<th>Number (Total=222)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cannot determine if dolphin in area:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No reason to believe there is a dolphin:</td>
<td>8</td>
<td>3.6%</td>
</tr>
<tr>
<td>Reason to believe there is a dolphin:</td>
<td>27</td>
<td>12.2%</td>
</tr>
<tr>
<td>Can identify to group of dolphins:</td>
<td>101</td>
<td>45.5%</td>
</tr>
<tr>
<td>Can identify to individual dolphin:</td>
<td>86</td>
<td>38.7%</td>
</tr>
</tbody>
</table>

has to do with the calibration signal itself. The Barker codes and frequency upsweeps used for calibration were specifically designed to be easily locatable, while dolphin sounds may or may not be. In addition, many of the dolphin sounds may have lower signal-to-noise ratios than the calibration sounds, making localizations more susceptible to contamination from ambient noise, such as snapping shrimp clicks. Another factor was the movement of the dolphin while vocalizing, which may result in smearing of the localization. Also, if the synchronization between the acoustic localization and the video imaging is not perfect, rapid movements of the dolphin will result in an increased discrepancy between the localized position and the position on the video image. The third factor concerned the limited visibility of the water column. For the dolphin sounds, most of the matches between the localizations and the dolphin image on the video were to whatever portion of the dolphin’s anatomy was visible. Since dolphins can be up 4 m long, using a portion of the anatomy far from the head could cause significant error. In addition, although the operator attempted to keep track of all the dolphins in the pool, it is possible that limited visibility may have resulted in an ID mismatch.

In addition, although the operator attempted to keep track of all the dolphins in the pool, it is possible that limited visibility may have resulted in an ID mismatch. This system of acoustic localization using a fixed array combined with elevated video imaging has several benefits. The elevated video enables increased visibility into the water column, along with possibilities of more detailed behavior analyses. The localization component enables matching of the vocalization to concurrent behavior. Another advantage of this system is that it is not necessary that the human observers be near the animals under observation. This may be important as the animals were fed by human trainers, and the presence of any humans may interrupt the dolphins’ normal social routine. The system could be used in a wide variety of captive situations, as long as the appropriate localization and video information could be obtained. Use of the system in reverberant captive environments may be limited by the localization algorithm used. The primary disadvantage of this system is that fixed arrays and elevated video cameras can be difficult to implement in ocean situations with free-ranging dolphins. Possibilities for implementation include placing fixed hydrophone arrays and video equipment in a bay or channel which is frequented by wild animals (Janik, 2000). A smaller system could possibly be implemented off of a boat using a towed array and video system on a tethered blimp [combining the systems described by Miller and Tyack (1998) and Nowacek et al. (2000)]. Obtaining error results for these situations would be facilitated by further theoretical work on the algorithm.

There are several avenues for improvement of the system. Increased resolution would enable identification of the vocalizing dolphin when dolphins are closer together. The video projection error appears to be the dominant source of error. However, since the video calibration error was sometimes over 10 pixels, the solution is probably not to increase image resolution. In one sense, this level of video error is not inherent to the system, and the results could probably be easily improved by superior camera placement and improved camera calibration. Placing the camera over the lagoon would reduce the error in at least one dimension. Using more than 5 points to perform daily calibrations, as well as using points spaced more evenly in the image, would probably increase accuracy of the calibration as well. Reducing this video error would be a relatively easy way to increase resolution of the entire system.

Increasing resolution of the acoustic localization system would also be helpful. Use of even slightly incorrect hydrophone positions can cause significant errors in localization. Thus, any method of improving calibration of hydrophone positions would likely decrease localization error. Placing the source a known distance from each hydrophone instead of directly adjacent to each hydrophone might avoid possible hydrophone overloading, near-field effects, as well as strong reverberation effects from the wall next to each hydrophone. Also, the localization algorithm could be improved to reduce its sensitivity to reverberation. Use of a sampling rate higher than 48 kHz could improve the resolution of the system. Although data storage might become more of an issue, more of the higher-frequency energy in echolocation clicks and burst pulsed sounds would be captured.

V. CONCLUDING REMARKS

The combination of acoustic localization and video sampling techniques allows us to link dolphin sounds with the identity of the vocalizing dolphin. If the behavior of the vocalizing dolphin is known, either from the video record or from more detailed real-time behavioral sampling, sounds can be linked to the contexts under which they were made. This is very important if we desire to ascertain the functions of vocalizations. Possible uses of this system include studying numerous aspects of the social contexts and behavioral function of vocal behavior as well as purely acoustic aspects of sounds under normal social conditions (e.g., directionality). There exist several avenues for improvement of the system. However, many new and interesting questions about marine mammal behavior can be asked and answered with the current resolution. For instance, echolocation behavior of foraging animals could be studied, capturing the behavior of the echolocating dolphin, and possibly its prey as well, on video. The system might help resolve the current debate on the existence of signature whistles in normally socializing bottlenose dolphins (McCowan and Reiss, 1995, 2001) by providing unbiased data sets of whistles identified to individual. The system might also help in testing the hypothesis that animals use cues in signal directionality for pod cohesion and communication (Lammers, 2001; Miller, 2002).

ACKNOWLEDGMENTS

The authors would like to thank DolphinQuest Bermuda and its kind and helpful staff for use of the facility, for their efforts on behalf of the project, and for their flexibility and incredible support. The Bermuda Maritime Museum was also very helpful and understanding of research needs and schedules. Stephanie Watwood was indispensable for data collection in the field, and deserves extra gratitude for her tireless efforts. Thanks to Kathy Dunsmore for assistance in development and writing of the localization algorithm. Thanks to Jay Wheeler, Amy Aprill, and Eli Woolery for assistance.
with data collection. The manuscript was improved by revisions made by Michael Casso, John Buck, John Leonard, Amy Samuels, and three anonymous reviewers. Funding was provided by the Waikoloa Marine Life fund, Grant No. IBN-9975523 from the National Science Foundation, a graduate student fellowship for R. Thomas from the National Science Foundation, and an Ocean Ventures Fund Grant for R. Thomas from the Woods Hole Oceanographic Institution. This is Woods Hole Oceanographic Institution Contribution Number 10444.

1We follow the usage of Tyack and Miller (2002; section 6.2.2) calling "vocalizations" any sounds produced by specializations of the respiratory tract. We do not mean to imply that all sounds called vocalizations are made by the larynx.


