Animal Behavior and Cognition
ABC 2016, 3(4): 288-298
DOI: 10.12966/abc.07.11.2016

Whose ~~Line~~ *Sound* is it Anyway? Identifying the Vocalizer on Underwater Video by Localizing with a Hydrophone Array

Matthias Hoffmann-Kuhnt[1*], Denise Herzing[2], Abel Ho[1], & Mandar A. Chitre[1]

[1]Acoustic Research Laboratory, Tropical Marine Science Institute, National University of Singapore
[2]Wild Dolphin Project and Florida Atlantic University

*Corresponding author (Email: tmsmh@nus.edu.sg)

Citation – Hoffmann-Kuhnt, M., Herzing, D. L., Ho, A., & Chitre, M. A. (2016). Whose ~~line~~ sound is it anyway? Identifying the vocalizer on underwater video by localizing with a hydrophone array. *Animal Behavior and Cognition, 3*(4), 288–298. doi: 10.12966/abc.07.11.2016

Abstract - A new device that combined high-resolution (1080p) wide-angle video and three channels of high-frequency acoustic recordings (at 500 kHz per channel) in a portable underwater housing was designed and tested with wild bottlenose and spotted dolphins in the Bahamas. It consisted of three hydrophones, a GoPro camera, a small Fit PC, a set of custom preamplifiers and a high-frequency data acquisition board. Recordings were obtained to identify individual vocalizing animals through time-delay-of-arrival localizing in post-processing. The calculated source positions were then overlaid onto the video – providing the ability to identify the vocalizing animal on the recorded video. The new tool allowed for much clearer analysis of the acoustic behavior of cetaceans than was possible before.

Keywords – Spotted dolphin, Bottlenose dolphin, Echolocation, Vocalization, Whistles, Localization, *Stenella frontalis*, *Tursiops truncatus*

Localizing a vocalizing dolphin underwater has been a challenge for decades. Although traditional methods of hydrophone arrays or towed hydrophone arrays from the surface have been used (Freitag & Tyack, 1993; Lammers & Au, 2003), the challenge of localizing a dolphin with underwater video has yet to be achieved. In using such an array, researchers were able to tell that vocalizing animals were present, and to calculate the approximate position of the animals in reference to the array – but any correlation with underwater video was not possible.

In the Bahamas a long-term study of dolphins, using underwater video and hydrophone recordings has correlated general types of vocalizations with behavior (Herzing, 1996, 2000). However, dolphin-to-dolphin acoustic signal use has yet to be applied to underwater video as it has been to localizing dolphin vocalizations using towed hydrophone arrays (Lammers & Au, 2003). The problems are vast. Detection of a sound source underwater versus in-air is influenced by changes in the medium properties (speed of sound, absorption, etc.). The speed of sound in water is 4.5 times higher than in air, and because, for the human hearing system, localizing a source depends largely on arrival time differences of the same signal between the two ears, our in-air ability to determine the source of a sound fails underwater. A second issue affecting localization is that, in air, sound travels to the inner ear via the outer ear canal, making this

the only access route for external sounds that are not physically connected to our head (bone conduction). This means that reception depends on our outer ears and our brain is adapted for this physiology. When we listen to sound sources underwater, the physics change because now sounds can travel straight from water through tissue and bones to our inner ears. As a result, sounds seem to be coming from everywhere. These physical changes could be compensated, artificially, by placing sensors at 4.5 times the distance between our ears, to adjust to the higher sound speed.

For video taken underwater this can be achieved by having two hydrophones separated by about 70 cm and feeding these signals into the stereo channels of the camera. This would solve the azimuthal problem of locating sound sources underwater, but not the elevation problem. In the latter, two synchronized sensors result in a circle of uncertainty created by the intersection of two spheres around the two sensors, on which the sound source could potentially lie. To determine the exact location of a sound source, a minimum of four synchronized sensors are necessary to localize the position. Three spheres would intersect at two points (reducing the uncertainty, but not eliminating it) but four spheres would intersect at a single point in space, as long as the fourth sensor does not lie in the plane created by the three other sensors. Such an arrangement would be of great value, because it would enable researchers to identify the vocalizer while observing its behavior.

Method

Materials

Underwater Dolphin Data Acquisition System (UDDAS). In 1999, Marc Lammers (Au, Lammers, & Aubauer, 1999) built a single channel recording device that could be synchronized with a video camera, in an underwater housing, with a sampling frequency of 300 kHz (see Figure 1). It had two gain settings and the acquisition software was based on Labview. The system was first used at the Wild Dolphin Project in 2002 and data were collected with this system throughout several seasons. Because that system had only one hydrophone, identifying a particular vocalizing animal in a group of dolphins through acoustic means was not possible. Furthermore, to balance recording time, data load, etc., – the system was set to sample at 220 kHz – thus not quite covering the complete frequency range of the dolphins being observed.

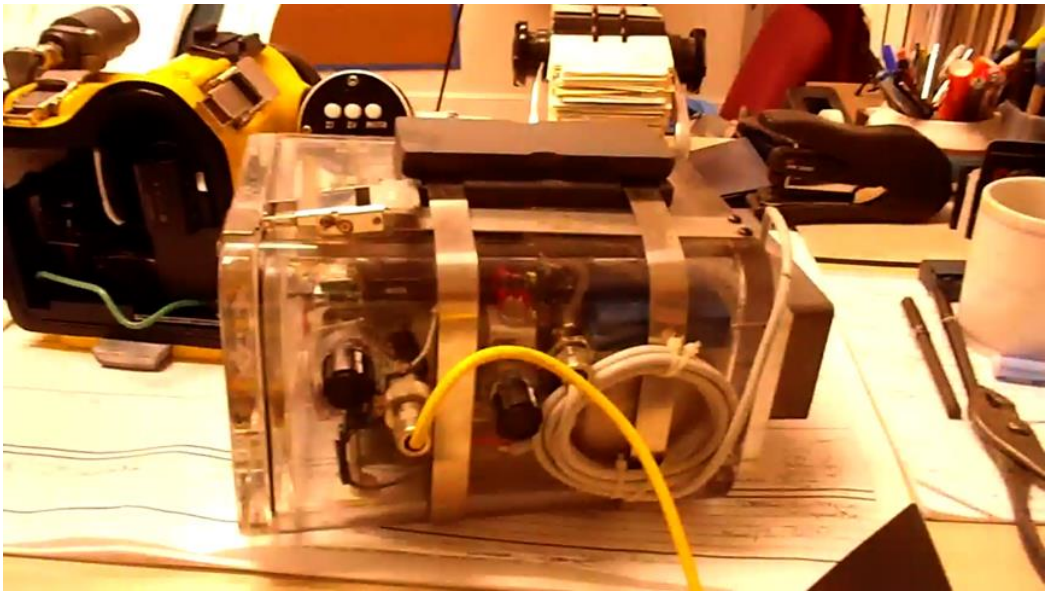

Figure 1. "UDDAS" the single channel recording device with the hydrophone (right), the synchronization cable (front) and the connection mount to the camera (top).

We thus decided to create a system that would be able to combine video and three channels of synchronized audio and that would be able to cover the whole entire range of dolphin vocalizations by sampling at a frequency of 500 kHz. In the time since the first single-hydrophone system was designed, technology and processing power have increased substantially, making it possible to combine all of the requirements into one package that was is still maneuverable by a swimmer.

A first "pre-runner" version of the new system came about in 2007 and is shown in Figure 2. This 3-hydrophone system was used to record humpback whales in the wintering grounds of Hawaii in the hopes of determining the location of the sound production system in the humpback whale (Potter, Pack, Hoffmann-Kuhnt et al., 2007; Potter, Pack, Reidenberg, et al., 2007).

Figure 2. Photo of the "pre-runner" system – consisting of a Sony underwater housing with camera, three hydrophones and the computer section with a PC104 system.

The system was limited to a frequency range of up to 20 kHz. This was sufficient for humpback whale vocalizations, but would only cover a small part of the range of frequencies in dolphin vocalizations.

New System. For the new system, we decided to maximize the potential by sampling at 500 kHz on each of the three hydrophones, thus covering the entire range of the dolphins' vocalizations. The system consisted of three Reson TC 4013 hydrophones, a National Instruments ® USB 6356 data acquisition card, a FIT PC Intense, custom-build built preamplifiers for each of the hydrophones, a GoPro Hero 3+ camera, an Arduino board to control the camera, a stripped down 7-in field LCD screen, and four 5200 mAh Turnigy lithium polymer batteries to power the system (see Figures 3 and 4). The housing was made from a large schedule-80 PVC pipe with an inner diameter of 206 mm, a length of 275 mm, and a wall thickness of 6mm. The front and back covers where made from clear acrylic: in the front, to provide a visual port for the GoPro camera, and, in the back, a viewport for the LCD screen. The complete electronic system was mounted in a system of stainless steel rods and rings that fit the inside of the PVC pipe and was attached to the front Plexiglas. Two external magnetic switches on the back Plexiglas allowed the user to power the system on and off and start and stop recordings.

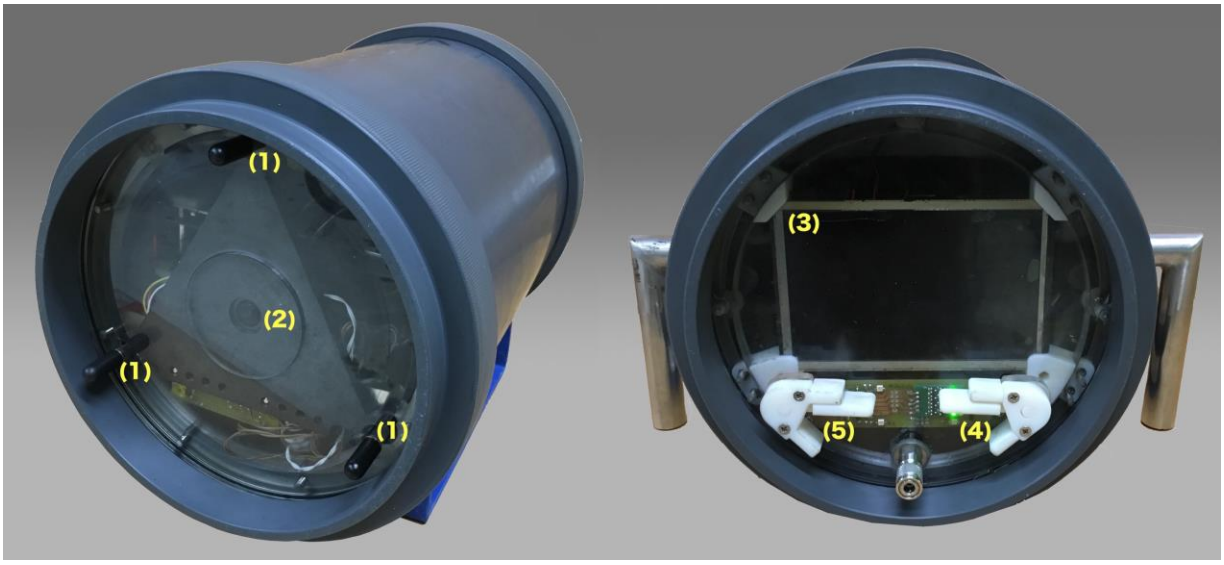

Figure 3. Photos of the system with front view on the left and back view on the right. (1) hydrophones, (2) GoPro video camera, (3) HDMI video screen, (4) power switch, and (5) record on/off switch.

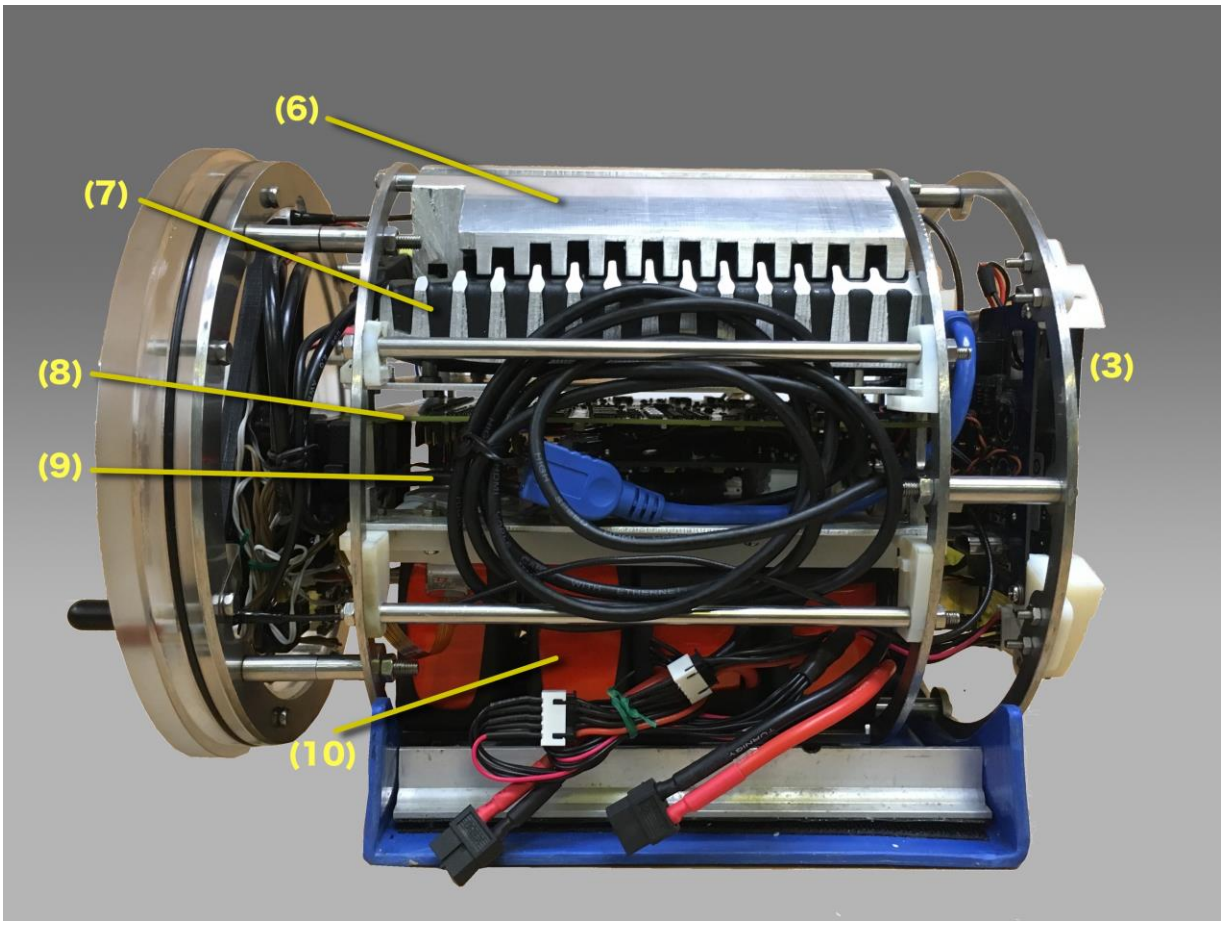

Figure 4. Side view of the system without the housing. (6) heat sink, (7) Fit PC Intense, (8) Data acquisition board, (9) amplifiers, (10) batteries, and (3) HDMI video screen.

Data Acquisition

Data acquisition was performed in Matlab with a custom script that streamed data from the three channels to a disk. A fourth channel contained a short signal generated by the Ardunio board that was also fed into the audio channel of the GoPro camera to enable post-recording synchronization. The position of the camera was centered relative to the three hydrophones that were separated from each other by 17 cm. With clicks consisting of a broadband frequency content and a large Signal-to-Noise Ratio (SNR) the timing resolution was very good and, even at a small aperture of 17 cm, yielded good angular resolution of about six degrees. Because the position of the hydrophones and the camera never changed, and the lens had a fixed focal length, it was then possible to localize the location of the sound sources and calculate if they originated with the visual field of the camera.

To obtain a recording, the system was powered on and handed down to a snorkeler in the water. The swimmer would then approach the dolphin(s) and start recording. Although, theoretically, recordings could last as long as the battery would allow (more than 3 hours), the recordings were kept to shorter sections of 3-5 min to allow for easier and faster processing later, since because the file size for the acoustic data would grow very large, very quickly. Once a session with the dolphins was finished and the snorkeler had returned to the boat, both the video and the acoustic files were downloaded to a laptop computer for processing.

Data Processing

The processing of the corresponding audio and video files followed a series of steps:

Read files:

The complete video and acoustic files were read into memory. The first frame of a video sequence to be analyzed would be loaded and the segments of the three synchronized audio channels that corresponded to that first frame were used for analysis.

Find clicks:

On each audio channel, a Hilbert transform was performed to find the points of highest energy in that segment. Within a small window (200 data points), the corresponding peaks were then used to perform Time-Delay-of-Arrival (TDoA) localizing and the resulting 3d- coordinates converted to azimuth (∅̂) and elevation (θ̂) angles as shown in the following equation:

$$(\hat{\emptyset}, \hat{\theta}) = \arg\min_{(\emptyset,\theta)} \left| \frac{1}{c} \begin{pmatrix} \sin\emptyset\cos\theta \\ \sin\theta \\ \cos\emptyset\cos\theta \end{pmatrix}^{T} \mathbf{P} - \tau \right|^{2}$$

where c is the sound speed, P is a matrix containing the position column vectors of each hydrophone, and τ is a vector containing the time difference of arrival for each hydrophone with respect to the reference hydrophone (at the origin of the coordinate system). The minimization problem was solved numerically using gradient descent starting from (0, 0). Because clicks are impulsive and very short, the window size in which to localize was kept small to reduce spurious alias that could occur when the algorithm combined energy from separate clicks.

Find whistles:

A similar procedure was used to find whistles in the segment. Again, a Hilbert transform was used to determine the location of highest energy in the channels – then a larger window (400 data points or more) was used to match and cross-correlate across the channels. In contrast to clicks, whistles can last over longer periods than the length of a video frame; thus, the window used for the cross correlation had to be larger. Then, TDoA localizing again resulted in azimuthal and elevation information.

Plot localization data:

The resulting azimuthal and elevation data for both clicks and whistles were then plotted on the video frame, if the angles were within the visual field of the camera (see Figure 5). Clicks were indicated with a red square and whistles with a yellow star. If the calculated angles lay outside the visual field, a red square for clicks and a yellow square for whistles were plotted in the bottom right corner of the video frame to indicate that the processing had detected a signal but that it was coming from a location outside the field of view of the frame (Figure 6). With only three hydrophones, there was a front-back ambiguity in reference to the plane of the hydrophones, but that was not relevant for clicks, since as the housing blocked signals coming from the back. For whistles this could be potentially more of a problem but could be resolved by adding at 4th hydrophone outside the plane.

Plotting of the time series:

The time series of one of the audio channels was plotted below the video frame with a padding of ± 4 seconds to display where in the overall time the analyzed section was (see Figure 5). The padding was added so that the waveform display would not pass too fast for the human eye if only the audio data for each particular frame would have been displayed if the resulting movie was would be played at the original speed.

Plotting of spectrogram:

A vertical (waterfall) spectrogram was plotted on the right side of the video frame to display the frequency content of the analyzed signal (see Figure 5, right). The same padding of ±4 s was also applied to the spectrogram.

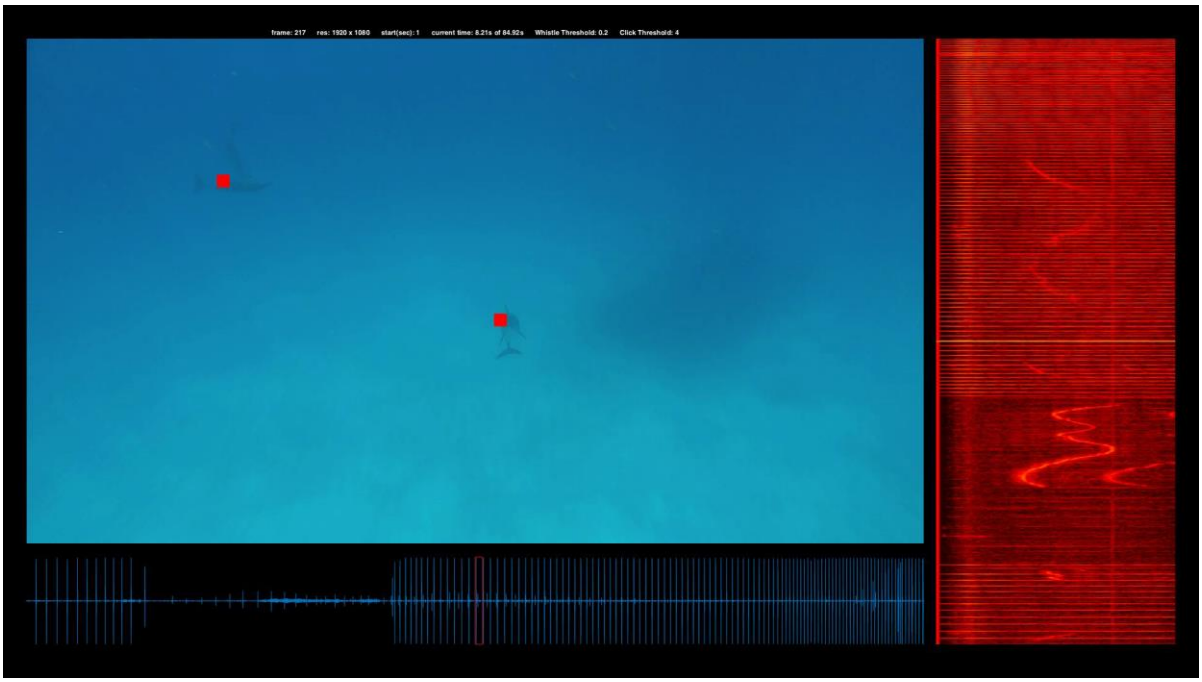

Figure 5. Screen capture of one of the frames during processing. The red squares identify the sound sources (in this case the two dolphins). The time series is plotted below the video frame and the spectrogram to the right. A red box in the center of the time series indicates the relevant data for each particular frame. A yellow line in the center of the spectrogram indicates also the current position.

The above steps were then repeated for successive video frames, creating a new movie that was then itself written to written to disk.

A down-sampled version of the two bottom hydrophone channels (resampled at 48 kHz to provide standard audio) of the entire section of interest were written as a separate .wav file, and later added to the

new movie as the audio track. This roughly corresponded to a “left” and “right” channel on a regular video camera and provided some form of stereo audio when viewed offline later.

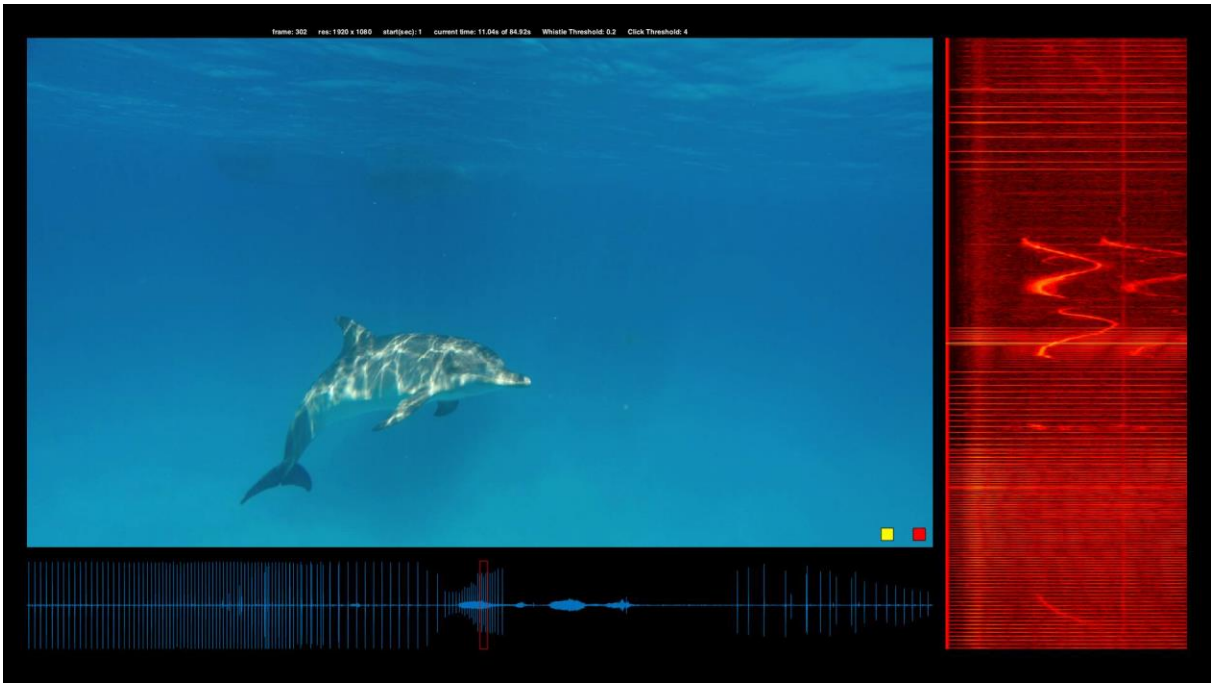

Figure 6. A frame of the processed movie that shows both a whistle and a click (yellow and red square in the bottom left of the video frame) detected by the software but neither was emitted by the dolphin visible in the video frame.

Ground truth for the system and the data processing was ascertained at a large saltwater pool at Ocean Park Hong Kong where a small hand-held underwater clicker was positioned at various locations in the visual field of the system while emitting clicks and the recorded data were used to confirm the accuracy of the procedure.

Results

The results of the recordings and analysis that followed clearly demonstrated that the method of identifying individual moving sound sources and tracking them for the duration of the vocalizations worked and was robust for most of the time (on some occasions when the SNR was too low, the system could not accurately identify the vocalizing animal). Even multiple sources were correctly identified – i.e., when a group of dolphins were vocalizing simultaneously (e.g., Figure 7).

Detection of clicks was, in general, more robust than the detection of whistles. This stemmed, in part, from the fact that clicks are very short and have a clear definition of peak energy, whereas the energy in whistles is more spread out, making them harder to localize. The duration of any clicks in the recordings was on a much smaller time scale (40–70 microseconds) than a frame of the video (1/30 of a second – a difference of three orders of magnitude) – so clicks did not last across several frames. This meant that, for each frame analyzed, the peak of the click was exactly defined and could be used to calculate the horizontal and vertical angle where it originated. Whistles on the other hand had durations of 1-2 s and thus could cover many (up to around 60) frames. The detection of either the onset of the whistle, or a distinct peak, could be calculated in the first frame where it occurred. However, for the following frames, the software was not able to calculate a time-of-arrival difference because the energy from the whistle appeared as a continuous waveform without an onset. In addition, the detection of both clicks and

whistles was based on the time series in any of the channels, and a threshold was set, below which any energy was ignored by the processing code. A lot of the whistles were rather quiet and thus the signal-to-noise ratio was very low. This meant that any whistle that did not exceed the threshold that was set above the background noise levels would not be detected. Furthermore, unlike clicks, which are impulsive, whistles might not have a clear onset in any of the three different channels (e.g., a very gradual and slow increase in the voltage recorded). This could lead to a "blur" in the position displayed – thus the localization of whistles is less accurate when the whistle is quiet.

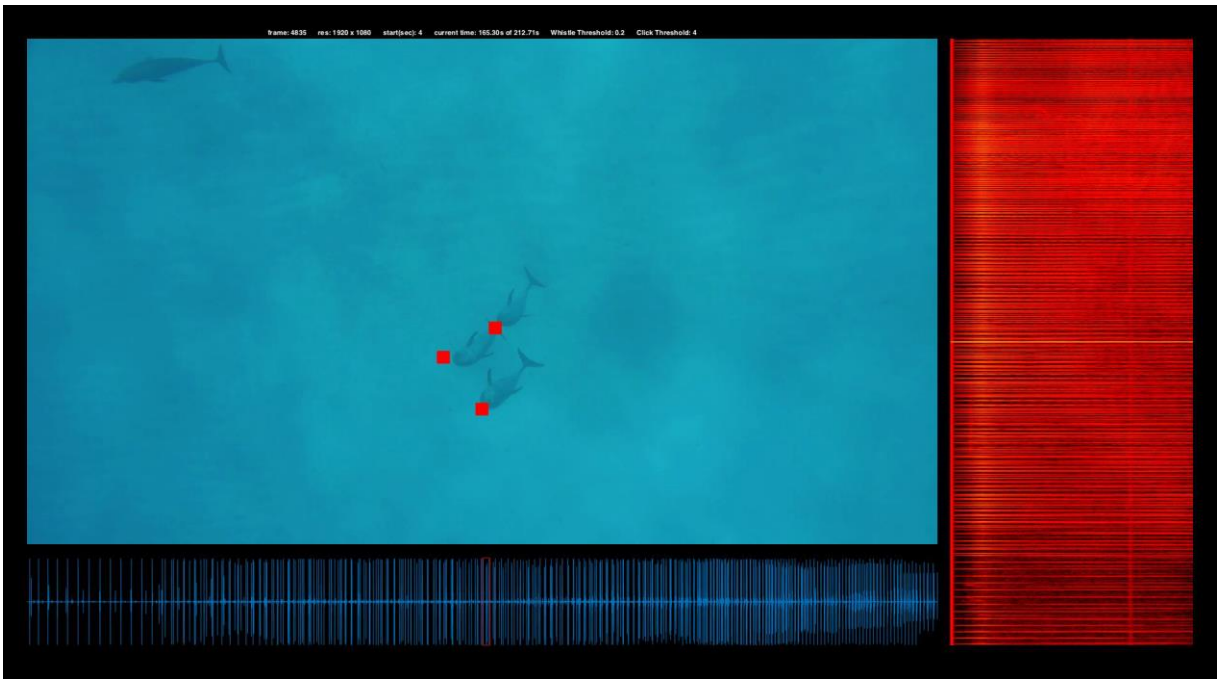

Figure 7. Screen capture of three animals vocalizing simultaneously and the software correctly identifying the three animals.

Surface Reflections

One interesting result of the analysis was that sometimes the reflections of the sounds from the surface were louder than the original sound. This was caused by multipath reflections that interfered positively – which meant that the software detected the highest energy and thus placed the location of the sound in the surface reflection of the dolphin as shown for a whistle in Figure 8. Setting the threshold correctly was also not always easy since because the noise level, as well as the level of the dolphin vocalizations, could change over the time and would then need to be adjusted for optimal detection. Further work is necessary to fine-tune and enhance the software of this system.

Figure 8. Detected whistles in a sequence where the surface reflection was stronger than the original direct path.

Discussion

The results of these recordings show that a much more detailed analysis of the behavior of animals is now possible by combining acoustic localization and video in one dataset. Where previously there was little or no evidence on which members of a group of dolphins was emitting sounds, it is now possible to often accurately determine the source of those sounds and correlate that with the behavior seen on the video. Sound is probably the dolphins' most important sense underwater, allowing the animals to communicate over long distances and, thus, playing a very important part in the social repertoire of these marine mammals. Because observable behavior that also includes vocalizations can now also be investigated, this opens up new vistas for research on social interaction. Possible future application could occur in the following scenarios:

1. Mother-calf interactions

Young calves learn how to search for food and hunt prey from their mothers, and although this teaching/learning behavior has been observed and documented previously (Bender, Herzing, & Bjorklund, 2009), it was not clear what role acoustics would play in this. For example, to what extent might a mother demonstrate how to echolocate, and how much of the vocalizations recorded in this context come from the calf? With the new technique described above, it would be possible to determine which of the two animals is vocalizing even when both animals swim right next to each other, as it is often the case with mother-calf pairs.

2. Interspecies vocalizations

In the wild, home ranges of different species of dolphins often overlap and interactions and collaboration between these species happen on a daily basis. For example, male spotted dolphins have been observed to socialize with their bottlenose counterparts (Elliser & Herzing, 2016; Herzing & Elliser, 2013) and here it would be very interesting to investigate what the underlying acoustics of such interactions may be. For instance, it is not clear which of the two different groups

initiates the interaction, and understanding the vocalizations better could help to determine that.

3. Localizing echolocators in group foraging by bottlenose dolphins

Bottlenose dolphins can be observed foraging in groups by echolocating for prey that is buried below the sand ("crater feeding" – Herzing 2004; Rossbach & Herzing, 1997) but very little is known about whether all animals echolocate simultaneously or whether just a few members in the group echolocate at any given time. Because this task is dependent on echolocation clicks, a much more detailed investigation and analysis of that behavior could be done with the new device.

4. Vocalizations in agonistic male coalitions

Male dolphins of some species often spend a large amount of time in male-male coalitions (Connor, Heithaus, & Barre, 1999; Connor, Smolker, & Richards, 1992) and the accompanying acoustic interactions both within and across coalitions are not well understood. A clear determination of which of the partners in the coalition are communicating would be of great value to understanding the dynamics of coalition behavior.

5. Humpback whale competitive groups

In Hawaiian waters male humpback whales form groups that compete for access to females (Baker & Herman, 1984; Tyack & Whitehead, 1983). Their complex social behavior within this competitive group is accompanied by vocalizations (different from the "songs" that are sung by solitary males), but it remains unclear what the functions of these vocalizations are. Identifying which males are vocalizing at what time would enhance the understanding of the complex social behavior of the whales and maybe lead to the identification of the male that would eventually reproduce with the female. Modifications to the system (increasing the aperture, etc.) might be necessary to adjust to the lower frequencies of the vocalizations in this case.

6. Localization of sound production areas in baleen whales

How many of the baleen whales produce sound is barely understood, and current knowledge and theories are based almost solely on anatomical evidence. If video of a vocalizing whale can be obtained with simultaneous acoustic recordings from the array, the results could pinpoint where the sounds originate on the whale's body. In combination with the anatomical findings, this could lead to a greatly improved model of baleen whale sound production. The findings of Potter, Pack, Hoffmann-Kuhnt et al. (2007) were based on a very small sampling of recordings and new recordings would strengthen the evidence and could extend the findings to other species of baleen whales. Here, similar modifications to the system as mentioned before might be necessary to compensate to the substantially lower frequencies of whale vocalizations.

The above mentioned scenarios are just a small subset of the questions that might be answered with the new methodology. Addressing these questions, and many others, would greatly enhance our understanding of the function of acoustics in marine mammal behavior.

References

Au, W. W. L., Lammers, M. O., & Aubauer, R. (1999). A portable broadband data acquisition system for field studies in bioacoustics. *Marine Mammal Science, 15*, 526–531. doi: 10.1111/j.1748-7692.1999.tb00817

Baker, C. S., & Herman, L. M. (1984). Aggressive behavior between humpback whales (*Megaptera novaeangliae*) wintering in Hawaiian waters. *Canadian Journal of Zoology, 62*, 1922–1937.

Bender, C. E., Herzing, D. L., & Bjorklund, D. F. (2009). Evidence of teaching in Atlantic spotted dolphins (*Stenella frontalis*) by mother dolphins foraging in the presence of their calves. *Animal Cognition, 12*, 43–53.

Connor R. C., Heithaus, M. R., & Barre, L. M. (1999). Superalliance of bottlenose dolphins. *Nature, 397*, 571–572. doi:10.1038/17501

Connor R. C., Smolker R. A., & Richards A. F. (1992). Two levels of alliance formation among male bottlenose dolphins (*Tursiops* sp.). *Proceedings National Academy of Sciences USA, 89*, 987–990.

Elliser C. R., & Herzing, D. L. (2016). Long-term interspecies association patterns of Atlantic bottlenose dolphins, *Tursiops truncatus*, and Atlantic spotted dolphins, *Stenella frontalis*, in the Bahamas. *Marine Mammal Science, 32*, 38–56, doi: 10.111/mms.12242

Freitag L., & Tyack, P. L. (1993). Passive acoustic localization of the Atlantic bottle-nosed-dolphin using whistles and echolocation clicks. *Journal of the Acoustical Society of America, 93*, 2197–2205

Herzing, D. L. (1996). Vocalizations and associated underwater behavior of free-ranging Atlantic spotted dolphins, *Stenella frontalis* and bottlenose dolphins, *Tursiops truncatus*. *Aquatic Mammals, 22*, 61–79.

Herzing, D. L. (2000). Acoustics and social behavior of wild dolphins: Implications for a sound society. In W.W. L. Au, A. N. Popper, & R. R. Fay (Eds.), *Hearing in whales, Springer-Verlag handbook of auditory research* (pp. 225–272). New York, NY: Springer-Verlag.

Herzing, D. L. (2004). Social and nonsocial uses of echolocation in free-ranging *Stenella frontalis* and *Tursiops truncatus*. In: J. A. Thomas, C. F. Moss, & M. Vater (Eds.), *Echolocation in bats and dolphins* (pp. 404 – 410). Chicago, IL: The University of Chicago Press.

Herzing, D. L., & Elliser, C. R. (2013). Directionality of sexual activities during mixed species encounters between Atlantic spotted dolphin (*Stenella frontalis*) and bottlenose dolphin (*Tursiops truncatus*). *International Journal of Comparative Psychology, 26*, 124–134.

Herzing, D. L., & Johnson, C. M. (1997). Interspecific interactions between Atlantic spotted dolphins (*Stenella frontalis*) and bottlenose dolphins (*Tursiops truncatus*) in the Bahamas, 1985–1995. *Aquatic Mammals, 23*, 85–99.

Lammers M. O., & Au W. W. L. (2003). "Directionality in the whistles of Hawaiian spinner dolphins (*Stenella longirostris*): A signal feature to cue direction of movement?" *Marine Mammal Science, 19*, 249–264.

Potter, J. R., Pack, A. A., Hoffmann-Kuhnt, M., Koay, T. B., Seekings, P. J., & Chitre, M. A. (2007). A synchronized acoustic array, rangefinder & video system with examples from 'singing' humpback whales (*Megaptera novaeangliae*). *Proceedings of the 21st conference of the European Cetacean Society*, 23–25 Apr 2007, Donsotia, San Sebastian, Spain.

Potter, J. R., Pack, A. A., Reidenberg, J., Hoffmann‐Kuhnt, M., Seekings, P. J., Chitre, M. A., …Herman L. M. (2007, Nov). *Humpback whale song source location in the head, source levels and directionality from in-situ rebreather diver recordings*. Presented at the 17th Biennial conference on the Biology of Marine Mammals, 29 Nov–3 Dec 2007, Cape Town, South Africa.

Rossbach, K., & Herzing, D. L. (1997). Underwater observations of benthic-feeding bottlenose dolphins (*Tursiops truncatus*) near Grand Bahamas Island, Bahamas. *Marine Mammal Science 13*, 498–504.

Tyack, P., & Whitehead, H. (1983). Male competition in large groups of wintering humpback whales, *Behaviour, 83*, 132–154. doi: http://dx.doi.org/10.1163/156853982X00067