High frequency short range Physeter monitoring and model of the organ of the sonar

Maxence Ferrari
U. Toulon, LIS, CNRS, AMU
CNRS LAMFA, U. Picardie J. Verne
Amiens, France
maxence.ferrari@lis-lab.fr

Hervé Glotin
U. Toulon, AMU, CNRS
LIS, DYNI, Marseille, France
SMIoT Toulon, France
glotin@univ-tln.fr (corr. auth.)

Ricard Marxer
U. Toulon, AMU, CNRS
LIS, DYNI, Marseille, France
ricard.marxer@lis-lab.fr

Valentin Barchasz
U. Toulon, AMU, CNRS
SMIoT Toulon, France
valentin.barchasz@gmail.com

Véronique Sarano
Longitude 181, France
veronique.sarano@gmail.com

Valentin Giés
U. Toulon, AMU, CNRS
SMIoT Toulon, France
valentin.gies@univ-tln.fr

Mark Asch
CNRS LAMFA, U. Picardie J. Verne
name of organization (of Aff.)
Amiens, France
email address

François Sarano
Longitude 181, France
francois.sarano@gmail.com

***Abstract*—Passive acoustics allow to study large animals and obtain information that could not be gathered through other methods. In this paper we study a set of near-field audiovisual recordings of a sperm whale pod acquired with a high-frequency and small aperture antenna. We show how we analyze those recordings in order to increase our knowledge regarding the characterization of their vocalization.**

***Index Terms*—Passive Acoustic Monitoring, Cetacean Survey, Abyss Monitoring, 3D Tracking, Long Term Survey, Transient Analysis, Weak Signal Detection, Autoencoder, FDTD.**

I. Introduction

Due to their large size and long dives, sperm whales are impossible to study in controlled conditions. The production of their vocalizations remains less understood than that of other smaller cetaceans such as dolphins. While anatomic descriptions have been performed via dissections, functional aspects and mechanisms involved are still unclear. We study their acoustic production through data-driven techniques on multi-channel near-field audio-visual recordings. Under the authority of Marine Megafauna Conservation Organisation directed by H. Vitry and, as part of the global program Maubydick, a team led by F. Sarano has been conducting a longitudinal study on the same group of 27 sperm whales since 2013. The main goal is to understand the relationship between individuals inside the family group and the dynamic of the Mauritian population. The main originality is that, since 2017, the data protocol is reinforced under the supervision of H. Glotin by a high sampling rate hydrophone array that can record their most acoustic intimate behaviour without disturbing them. We show in this paper the first results of these unique recording of this endangered species, and the challenges opened for the analyses, clustering and classification at the group versus the individual levels of this complex transients including the most advanced deep learning representation.

Two main studies are in progress: i) characterisation of the vocalisation localization by multi-modal analysis; and ii) an exploration of meaningful information contained in clicks including individual signature. i) The Direction of Arrival (DoA) is characterised using Generalized Cross-Correlation (GCC) beamforming with adaptive time-frequency weighting and pooling [1]. DoAs are crossed with the animal positions obtained from the video by a simple tracking algorithm. In the second study clicks are extracted and their DoA estimated. Deep learning is employed to analyse fundamental aspects of the clicks. Thus we propose new model, stereo autoencoders (SAE) to analyse these complex transients.

II. Material

During the years, Franois Sarano and his team have been periodically returning to Mauritius island in order to record local Sperm Whales (*Physeter macrocephalus*, *Pm*). Each year, we have been able to improve the recording protocol. Since 2017, on the initiative of H. Glotin, V. and F. Sarano has been using a GoPro Hero 4 mounted on an a stereophonic acoustic antenna of our design, based on our JASON SMIoT Toulon ultra high velocity DAQ designed for this extreme recordings. Our protocol evolved each year, with the access to additional high quality hydrophones: 2 hydrophones in 2017, 3 in 2018, and 4 in 2019. The hydrophones are from Cetacean Research. The DAQ is the Qualilife sound card [2], which was used in this study at 16 bits@600 kHz. It is able to record at a

We thank Vitry and Preud'Homme of Marine Megafauna Conservation Organisation, Mauritius, and R. Heuzey of 'Un ocan de vie' ONG, France.*******DGA Smiles Sabiod eadm FUI GIAS, Rgion haut de France

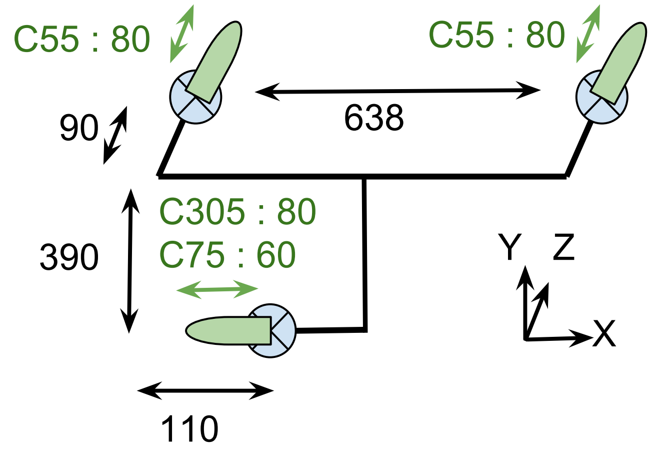

Fig. 1. Blueprint of the 2018 antenna

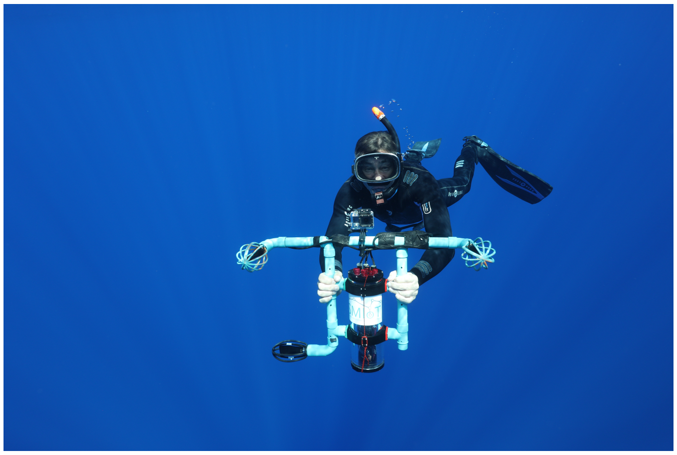

Fig. 2. Franois Sarano holding the 2018 antenna (Image: F. Guerin).

sampling rate up to 2 MHz per channel, up to 5 channels. In this paper, we focus on the results of 2018. The antenna is composed of two C55 hydrophones. The third one is a C305 which has been changed by a C75 because the C305 was too directive. The audio recording was on most of the time (during all dives and part of boat transfers between dives), while the video recordings were only done during dives. The audio files are 1min 12sec long (350 MB) and continuous, while the video recording was turn on manually.

III. Click Detection

Before doing any of the experiment we use a simple click detector on all the sound files. We cross-correlate the signal with one period of a 12.5 kHz sine which act as a band-pass filter (bandwidth of echolocation clicks is 10–15 kHz [3]), followed by a Teager-Kaiser filter [4], [5] and the extraction of local maxima in 20 ms windows (twice the largest Inter-Pulse Interval (IPI) of 10 ms [6]). We then convert the maxima's values into dB. The maxima usually form two distributions : one that emanate from the click and one that emanate from the maxima that are between click, which are maxima created by white noise. We thus filter out the noise by fitting two gaussians on the distribution formed, and only keeping values that are above the time the standard deviation of the gaussian with the smallest mean [7].

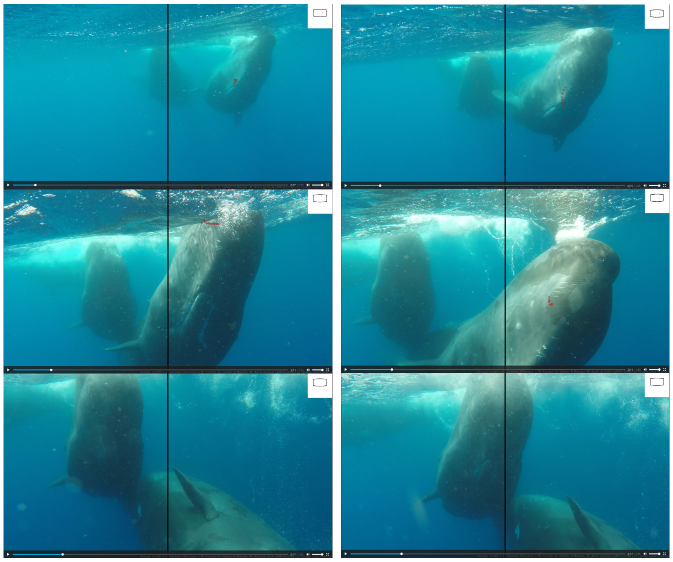

Fig. 3. Six frames from a video where the click have been localized. The top right corner show for the frame the clicks' azimuth / elevation, with the black border being the GoPro screen border. Other videos are available on http://sabiod.univ-tln.fr/workspace/Sarano_2018

IV. Linking clicks to their emitter

Since the 2018 antenna had 3 hydrophones, we were able to compute the elevation and azimuth of the clicks origin. With the click detected in II, we computed the two independent TDoA (Time Difference of Arrival) using the method describe in [8]. In order to obtain the angles from the TDoA, we had to suppose that the sperm whales where far from the antenna. With the elevation and azimuth of each angle, each click origin can be plotted on the video, as 3 shows. To do so, we converted the elevation and azimuth to xy pixel coordinate while taking into account the distortion added by the fish eye lens of the GoPro. Unfortunately, the GoPro elevation was lost. Most clicks seem to be shifted down in the video, which could be explained by a wrong estimation of the GoPro elevation. The GoPro video where re-synchronized with the audio recording using cross-correlation. Each point (DOA of a click) stays for 7 frames (starting from the frames the corresponding click is eared) on the video to make them easier to see. However the antenna does not have means to measure its rotation in space, which mean that every oscillation (which are strong due to waves) will shift the scene. Seven frames is already long enough for a point to give the impression that it is located where it should not be, when it was in fact in the right place in the first frames it was display. Since F. Sarano is able to identify the sperm whales in the video, this localization allowed us to link clicks to their emitter. This will allow us to analyze the link between a click sequence and the sperm whales behaviour that resulted.

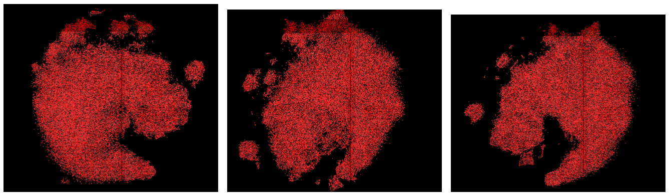

Fig. 4. Various angle of view of the 3D TSNE

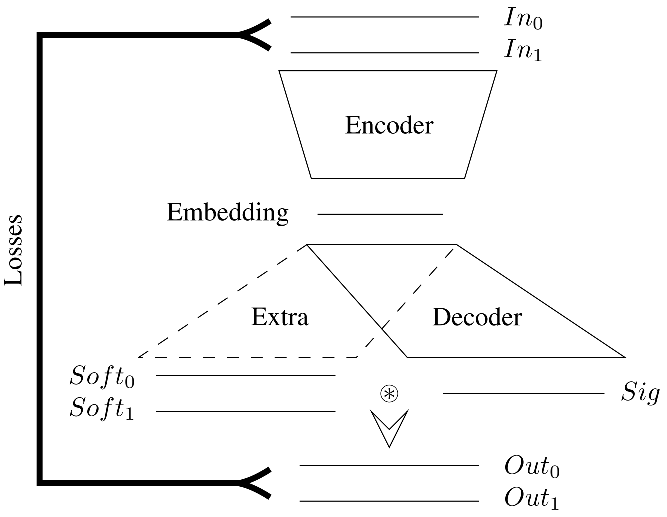

Fig. 5. Autoencoder architecture

Layer type, Activation	Input shape	Kernel, stride	Filters
Encoder			
Convolution layer, tanh	2*12000*1	1*11, 1*4	64
Convolution layer, leaky relu 5%	2*3000*64*1	1*1*64, 1*1*1	1
Convolution layer, tanh	2*3000*64	1*11, 1*4	128
Convolution layer, leaky relu 5%	2*750*128*1	1*1*128, 1*1*1	1
Convolution layer, tanh	2*750*128	1*11, 1*4	128
Convolution layer, leaky relu 5%	2*188*128*1	1*1*128, 1*1*1	1
Convolution layer	2*188*128	1*11, 1*4	128
Convolution layer, leaky relu 5%	2*47*128*1	1*1*128, 1*1*1	1
Convolution layer	2*47*128	1*11, 1*4	128
Convolution layer, leaky relu 5%	2*12*128*1	1*1*128, 1*1*1	1
Convolution layer	2*12*128	1*11, 1*4	128
Convolution layer	2*12*128	2*1, 1*1	256
Dense layer leaky relu 5%	3072		2048
Dense layer leaky relu 5%	2048		512
Dense layer	512		128
Decoder			
Dense layer leaky relu 5%	128		1024
Dense layer leaky relu 5%	1024		2048
Dense layer leaky relu 5%	2048		2048
Transpose convolution, leaky relu 5%	1*128*16	1*5, 1*2	8
Transpose convolution	1*256*8	1*5, 1*4	8
Transpose convolution	2*1024*8	1*5, 1*4	1
Extra branch			
Dense layer, leaky relu 5%	128		1024
Dense layer, leaky relu 5%%	1024		2048
Dense layer, leaky relu 5%%	2048		6000
Transpose convolution	2*3000*1	2*11, 1*4	1
Softmax	2*12000	1*12000	12000

TABLE I
MODEL ARCHITECTURE

V. Learning Latent Space and Invariant by Stereo Autoencoder

In conclusion with the extraction of the TDoA from the multichannel recordings, we were able to compute the DOA (Direction of Arrival), which allowed us to pinpoint the source location on the video. Since F. Sarano's team is able to identify each animal, we are able to tie each click to an individual, thus giving us a database that can be used to understand more deeply which features are tied to an individual, and which are invariant and define the *Pm* sonar. We now use autoencoders to analyse the data. The goal with the autoencoders is double. As for usual autoencoders, the first aim was to study the features captured by the autoencoders, which could be features describing the individual that emitted the click, or features describing the type of click that has been emitted. We show that with the model we used, it was possible to obtain an embedding space that clusters the examples regarding these features as figures in 3D in Fig. 4.

For this study we chose a stereo autoencoder (Fig. 5) since we already had similar work done with the same network architecture. We thus chose to use the first two channels because they are less noisy than channel 3, and recorded with the same hydrophone, which could help the network to learn. This autoencorder's decoder is composed of two branches. One that will reconstruct the signal, and one that will offset it to match each channel input.

The other goal of the autoencoder is to have an unsupervised way of computing the TDoA. By computing TDoA in this manner, the aim is to obtain better localization results than with usual methods, such as the generalized cross-correlation. Hence, we try to instance parameters of the 3D sonar production model. Another output of this approach is to find invariant in the embedded latent space related to a possible individual signature of each individual, because we can feed the SAE only with the localized / identified / named clicks from the previous process.

VI. Modelisation and simulation of biosonar emission

Obtaining the direction of each click, knowing which click features characterize an individual and which describe the information contain in a click help us improve the sperm whale head model we made to understand his sonar. Knowing the individual features allow us to study one sperm whale and try to improve the other characteristics to better match the recorded clicks, and then fix those characteristics and verify that we still have good results when simulating other sperm whales. Knowing the information contained in clicks is useful to find on which part the sperm whale is able to act in order to encode this information, or change the click in various behaviour such as the one describe in [9].

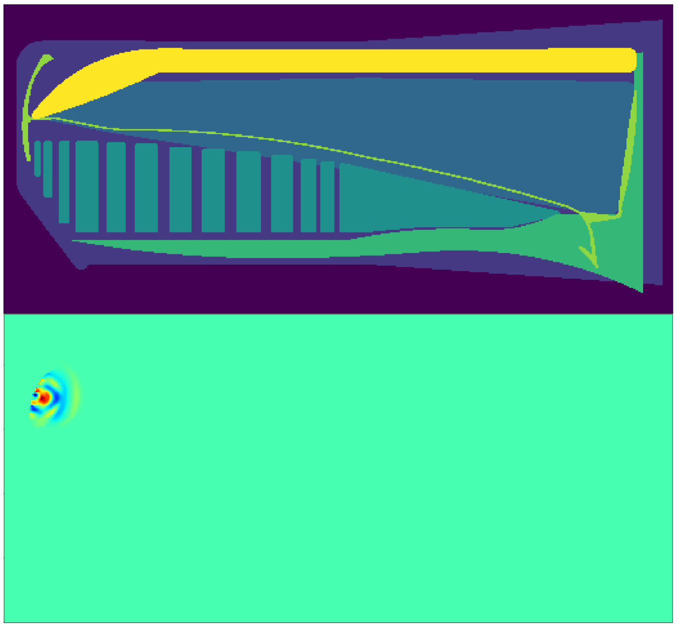

Fig. 6. Top: Sperm whale slice. Bottom: 3D wave propagation simulation

Before trying more complex modelization methods, we tried with simpler ones such as FDTD (Finite Difference Temporal Domain). We design our model base on [10], [11]. Fig. 6 shows a slice of a $540 * 220 * 240\,cm^3$ FDTD, with a $1\,cm$ space step and a $1\,s$. A $20\,ms$ simulation render in 1 hour. The ABC (Absorbing Boundary Condition) use what is discribe in [12].

VII. Discussion and conclusion

Future work will complete AE with SiameseNets [13]. AEs work by reducing the signals to a few characteristics while allowing their reconstruction. Siamese-nets are trained to maintain small distances between representations of clicks belonging to a given group, and large distances with others. We will then group together clicks coming from the same direction at similar times. The obtained representations are visualized in search of interesting invariant like individual acoustic signatures of each whale.

References

[1] Michael I Mandel, *Binaural model-based source separation and localization*, Citeseer, 2010.

[2] M. Fourniol, V. Gies, V. Barchasz, E. Kussener, H. Barthelemy, R. Vauché, and H. Glotin, "Low-power wake-up system based on frequency analysis for environmental internet of things," in *Int. Conf. on Mechatronic, Embedded Systems, App.* IEEE, 2018, pp. 1–6.

[3] P.-T. Madsen, R. Payne, N. Kristiansen, M. Wahlberg, I. Kerr, and B Møhl, "Sperm whale sound production studied with ultrasound time/depth-recording tags," *J. of Exp. Biology*, vol. 205, no. 13, pp. 1899–1906, 2002.

[4] V. Kandia and Y. Stylianou, "Detection of sperm whale clicks based on the Teager–Kaiser energy operator," *Applied Acoustics*, vol. 67, pp. 1144–1163, 2006.

[5] H. Glotin, F. Caudal, and P. Giraudet, "Whale cocktail party: real-time multiple tracking and signal analyses," *Canadian acoustics*, vol. 36, no. 1, pp. 139–145, 2008.

[6] R. Abeille, Y. Doh, P. Giraudet, H. Glotin, J.-M. Prevot, and C. Rabouy, "Estimation robuste par acoustique passive de lintervalle-inter-pulse des clics de physeter macrocephalus: méthode et application sur le parc national de Port-Cros," *Journal of the Scientific Reports of Port-Cros National Park*, vol. 28, 2014.

[7] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.

[8] M Poupard, M Ferrari, J Schluter, R Marxer, P Giraudet, V Barchasz, V Gies, G Pavan, and H Glotin, "Real-time passive acoustic 3d tracking of deep diving cetacean by small non-uniform mobile surface antenna," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8251–8255.

[9] Stefan Huggenberger, Michel André, and Helmut HA Oelschläger, "An acoustic valve within the nose of sperm whales p hyseter macrocephalus," *Mammal review*, vol. 44, no. 2, pp. 81–87, 2014.

[10] Malcolm R Clarke, "Structure and proportions of the spermaceti organ in the sperm whale," *Journal of the Marine Biological Association of the United Kingdom*, vol. 58, no. 1, pp. 1–17, 1978.

[11] John C Goold, James D Bennell, and Sarah E Jones, "Sound velocity measurements in spermaceti oil under the combined influences of temperature and pressure," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 43, no. 7, pp. 961–969, 1996.

[12] Robert L Higdon, "Absorbing boundary conditions for difference approximations to the multidimensional wave equation," *Mathematics of computation*, vol. 47, no. 176, pp. 437–459, 1986.

[13] Sumit Chopra, Raia Hadsell, Yann LeCun, et al., "Learning a similarity metric discriminatively, with application to face verification," in *CVPR (1)*, 2005, pp. 539–546.