High-frequency Near-field Physeter macrocephalus Monitoring by Stereo-Autoencoder and 3D Model of Sonar Organ

Maxence Ferrari, Hervé Glotin, Ricard Marxer, Valentin Barchasz, Véronique Sarano, Valentin Gies, Mark Asch, François Sarano

High-frequency Near-field *Physeter macrocephalus* Monitoring by Stereo-Autoencoder and 3D Model of Sonar Organ

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

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

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
LAMFA, CNRS, U. Picardie
Amiens, France
mark.asch@u-picardie.fr

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

***Abstract*—Passive acoustics allow us 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 ultra high-frequency and small aperture antenna. We propose a novel kind of autoencoder, a Stereo-Autoencoder, and show how it allows to build acoustic manifolds in order to increase our knowledge regarding the characterization of their vocalizations, and possible acoustic individual signature.**

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

I. Introduction

Due to their large size and long dives, sperm whales (*Physeter macrocephalus*, *Pm*) 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 roughly performed via dissections, functional aspects and mechanisms involved are still unclear. We study their acoustic production through data-driven techniques on multichannel 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 dynamics of the Mauritian population. The main originality is that, since 2017, the data protocol is reinforced by a collaboration with H. Glotin with the use of a high sampling rate hydrophone array, Blue JASON, of SMIoT and LIS DYNI, that can record their most intimate acoustic behaviour minimizing their disturbance. We show in this paper the first results of these unique recordings from this endangered species, and the challenges involved in their analyses, clustering and classification at the group-versus-individual level of such complex transients through the use of advanced deep learning methods.

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

II. Material

During the past 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 the recording protocol has been evolving to improve the data collection. Since 2017, on the initiative of H. Glotin, V. and F. Sarano have been using a GoPro Hero 4 mounted on a stereophonic acoustic antenna of our design, based on our JASON SMIoT Toulon ultra high velocity DAQ designed for these extreme recordings. Our protocol has evolved each

We thank H. Vitry and A. Preud'Homme of Marine Megafauna Conservation Organisation, Mauritius, and R. Heuzey of 'Un océan de vie' ONG, France. We thank DGA and Région Hauts de France for the PhD grant of M. Ferrari. This research is partly funded by FUI 22 Abyssound, ANR-18-CE40-0014 SMILES, ANR-17-MRS5-0023 NanoSpike and MARITTIMO EUR. GIAS projects on advanced studies on cetaceans. We thank MI CNRS MASTODONS SABIOD.org and EADM MADICS CNRS scaled bioacoustic research groups, and SEAMED PACA project.

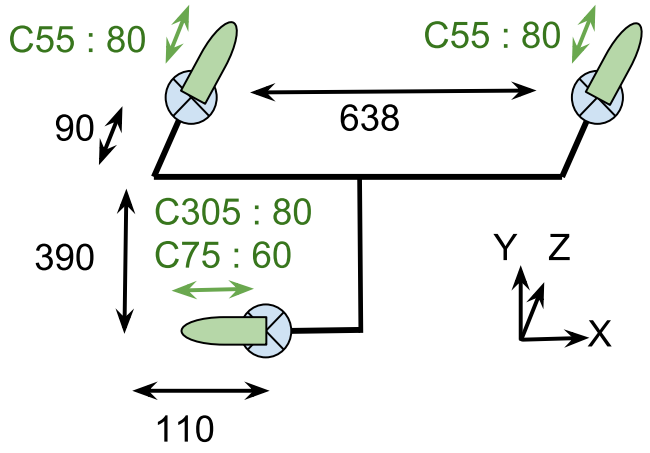

Fig. 1. Blueprint of the 2018 antenna.

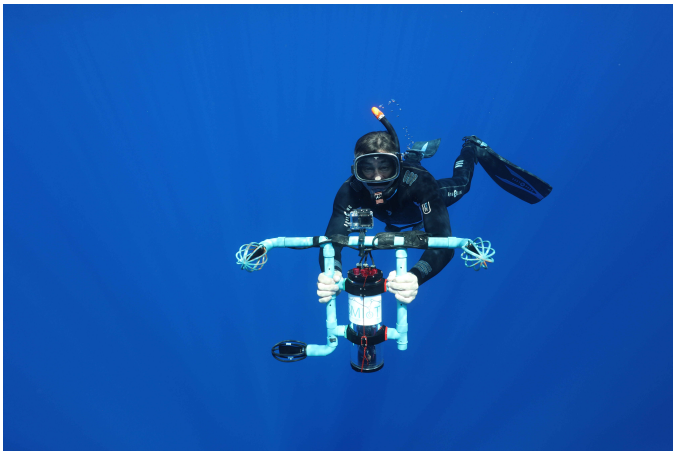

Fig. 2. Franois Sarano holding the 2018 antenna (Image: F. Guerin).

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 sampling rate up to 2 MHz per channel, up to 5 channels. In this paper, we focus on the results of 2018. The antenna was composed of two C55 hydrophones. The third hydrophone was a C305 which has been replaced 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 turned on manually.

III. Click detection

Before performing any of the experiments, we use a simple click detector on all the sound files. We cross-correlate the signal with one period of a 12.5 kHz sinusoid which acts as a band-pass filter (bandwidth of echolocation clicks is 10–15 kHz [3]), followed by a Teager-Kaiser filter [4], [5] and

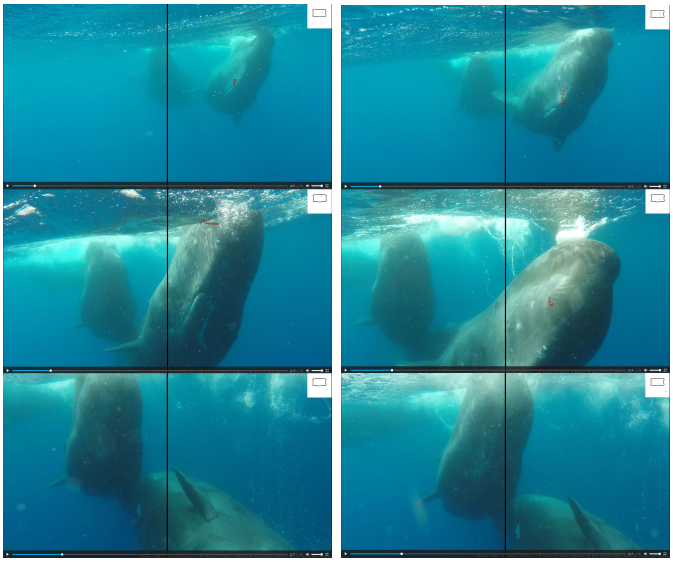

Fig. 3. Six frames from a video where the clicks have been localized. The top right corner shows for the frame the clicks' azimuth / elevation, with the black border being the GoPro screen border. Other videos are available on http://sabiod.univ-tln.fr/workspace/Sarano_2018

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 emanates from the click itself, and one that emanates from the maxima that are between clicks, 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 three times the standard deviation of the Gaussian with the smallest mean [7].

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 Section II, we computed the two independent TDoA (Time Difference of Arrival) using the method described in [8]. In order to obtain the angles from the TDoA, we had to suppose that the sperm whales were far from the antenna. With the elevation and azimuth of each angle, each click origin can be plotted on the video, as Fig. 3 shows. To do so, we converted the elevation and azimuth to XY pixel coordinates 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 videos were 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 earmarked) on the video to make them easier to see. However the antenna does not have means to measure its rotation in space, which means that every oscillation (which is 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

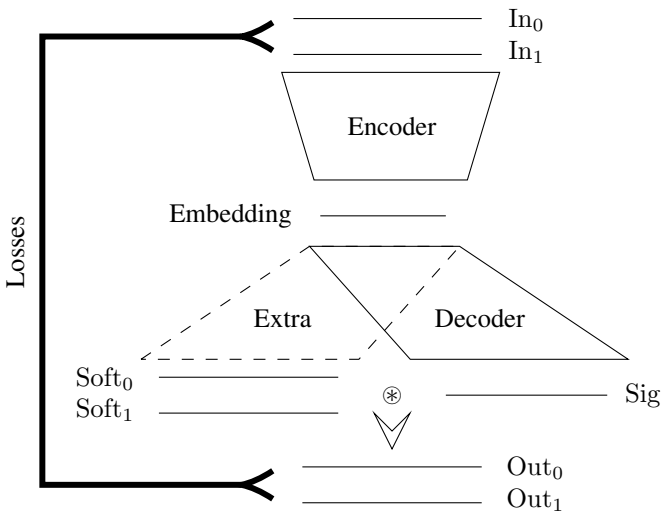

Fig. 4. Stereo Autoencoder (SAE) architecture

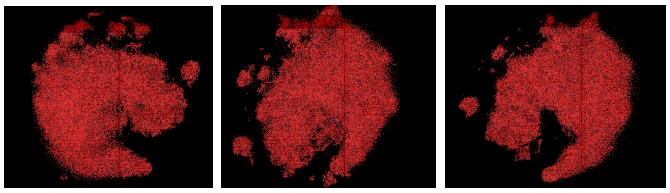

Fig. 5. Various angles of view of the 3D t-SNE (t-distributed stochastic neighbor embedding) showing various acoustic manifolds discovered by the SAE.

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

was in fact in the right place in the first frames in which it was displayed. Since F. and V. Sarano are 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.

V. Learning Latent Space and Invariant by Stereo-Autoencoder (SAE)

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. The visual identification of each animal allows us to tie each click to an individual. This database can be used to understand more deeply which features could be tied to an individual, and which are invariant and define the *Pm* sonar. We now propose a novel stereo-autoencoder to analyse the data. The aim is (i) to study the features captured by the autoencoder, which could be features describing the individual that emitted the click, and (ii) to get features describing the type of click that has been emitted, according to their TDoA, angle of arrival, intensity level difference. We show in the following section that this SAE generates an embedding space that efficiently clusters in 3D the clicks with respect to these features (Fig. 5).

The proposed stereo-autoencoder (SAE) is represented in Fig. 4. This model has simply a double input into an autoencoder with double output. We chose as inputs the first two channels (the least noisy and recorded with the same hydrophones) in order to facilitate the network to learn. The output of this SAE is composed of two branches, one that reconstructs the signal, and one that offsets it to match each channel input.

The other goal of the autoencoder is to have an unsupervised way of computing the TDoA. By computing TDoAs in this manner, we 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 aim of this method is to discover acoustic invariants in the embedded latent space related to a possible acoustic signature of each individual. This is possible because here we can feed the SAE only with the localized and identified clicks after the previous localisation process.

VI. Modelling and Simulation of Biosonar Emission

Obtaining the direction of each click, knowing which click features characterize an individual and which describe the information contained in a click, help us improve the sperm whale head model we made to understand its sonar. Knowing

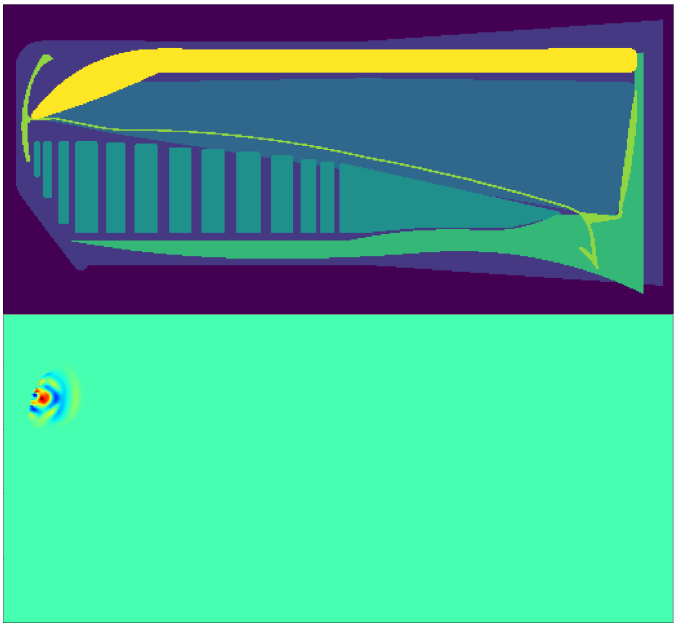

Fig. 6. Top: Sperm whale slice. Bottom: 3D wave propagation simulation

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 the 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 behaviours, such as the one described in [9].

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

VII. Discussion and conclusion

Future work will complete the SAE with SiameseNets [13]. Autoencoders 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 then visualized in search of interesting invariants, such as 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.