SUBMITTED TO INTERSPEECH 2026

Audio-visual diarisation of overlapping click trains in Sperm Whale (*Physeter macrocephalus*) social communication using a compact array

Lara Berkenbaum[1,2,3], François Sarano[2,3], René Heuzey[2,3,4], Axel Preud'homme[2,3,4], Véronique Sarano[2,3], Olivier Adam[2,5], Pascale Giraudet[1,2,3]

[1] Univ. Toulon, Aix Marseille Univ., CNRS, LIS, DYNI, Toulon, France
[2] Univ. Toulon, CIAN, Toulon, France
[3] Longitude 181, France
[4] Un Océan De Vie, France
[5] Alembert Institute & Neurosciences Paris-Saclay Institute, France

lara.berkenbaum@lis-lab.fr

Abstract

The individual attribution (i.e., diarisation) of sperm whale vocalisations during surface interactions constitutes a major methodological challenge due to severe overlaps, multipath propagation, and body shadowing effects. This paper presents a multimodal audio-visual workflow designed to deinterleave continuous click trains of immature males engaged in vocal sparring. Using a portable three-hydrophone array coupled with video, the proposed approach combines spectro-temporal tracking based on inter-click interval dynamics and constant-Q transform with spatial direction-of-arrival modelling projected onto the image plane for optical validation. This combined approach successfully isolates and assigns atypical emissions. This method opens new perspectives for ethological investigations of sperm whale communication, particularly for analysing interactional and dialogue-like acoustic dynamics.

Index Terms: bioacoustics, diarisation, source localisation, array signal processing, sperm whale, animal communication.

1. Introduction

In the sperm whale (*Physeter macrocephalus*), acoustic communication structures social interactions within matrilineal units [1]. Although vocal production primarily relies on broadband impulsive clicks generated by the well-described 'bent horn' mechanism [2,3], the individual attribution of these emission remains a major methodological bottleneck, particularly during surface socialisation phases when several individuals vocalise simultaneously.

These clicks fulfil multiple behavioural functions. The literature has mainly focused on two distinct functional and acoustic categories: on the one hand, echolocation clicks associated with foraging and navigation, characterised by regular emission patterns [4-6], and on the other hand, codas, stereotyped social sequences that function as markers of cultural identity [7,8].

Continuous click trains produced at the surface evade this functional dichotomy. These signals differ from codas by the absence of stereotyped rhythmic patterns and from foraging clicks by their emission context. Indeed, they occur during close-range interactions involving multiple simultaneous emitters, generating a true cocktail-party effect. Such acoustic emissions may play a crucial role in social structuring and perhaps in vocal learning processes, yet they remain poorly described to date. To our knowledge, continuous click trains produced in surface social interaction contexts have not yet been the subject of dedicated acoustic characterisation.

2. Related work

In these compact multi-emitter configurations, typical of surface socialisation phases, where individuals interact at very close range and move at similar depths, classical source separation and individual click assignment approaches reach their limits, thereby constraining fine-scale analyses of interactional dynamics. The proximity of the animals renders spatial separation by time differences of arrival (TDOA) using a compact array (fewer than 4 hydrophones) poorly discriminative and prone to geometric ambiguities [9-13]. Similarly, exploiting multipath delays to separate sources by depth [6,9,14,15] becomes ineffective, as the Lloyd mirror effect merges the direct path with the surface reflection [16], altering the spectral structure of the signal. In the temporal domain, disentanglement algorithms relying on the regularity of the inter-click interval (ICI), commonly used for echolocation clicks trains [6,17], become unstable when faced with overlapping emissions. Furthermore, acoustic discrimination based on the inter-pulse interval (IPI), which correlates with body size [18,19], fails to reliably differentiate individuals of similar morphology and age. Finally, exploiting fine-scale click features, notably the waveform, also exhibits high variability linked to the animal's orientation relative to the hydrophone (on/off-axis effect), modifying the spectral and temporal distribution of the signal [2,6,20].

In this context, we hypothesise that a multimodal audio-video approach can overcome these spatial and temporal limitations. We propose a method combining spectro-temporal click tracking with optical validation. Through a case study of a surface social interaction between two immature males, termed 'vocal sparring' [21, 22], this work aims to (i) describe the audio-video workflow for click deinterleaving and individual assignment in compact social context, (ii) demonstrate the feasibility of non-invasive emitter attribution using a compact and mobile sensor array, and (iii) provide an initial acoustic characterisation of these juvenile social vocalisations, opening the way for a fine-scale interactional analyses of previously undescribed behaviours.

3. Materials and Methods

3.1. Study subject and behavioural context

Data were collected between March and June 2023 off the coast of Mauritius (Indian ocean), as part of the 'La Voix des cachalots' programme (Longitude 181, Un Océan De Vie, CIAN). The study focuses on surface social interactions involving two immature males

(named Ali and Daren, 5 years old;[23]), characterised by continuous and frequently overlapping emissions of click trains ('vocal sparring')[21,22]. This context provides an optimal case study for testing individual assignment due to the spatial proximity of individuals and the acoustic overlap. To evaluate our methodology, we extracted and subdivided two audio recordings from distinct videos. The analysis encompasses six scenes: four featuring simple click trains and two containing complex trains with a higher degree of apparent overlap.

3.2. Audio-video acquisition system: OPALE array

To ensure passive and non-invasive acquisition, the setup relies on the manual deployment of a compact OPALE array [13, 24], maintained at a maximum depth of 5 m by an experienced scientific swimmer. The field trips complied with Mauritian regulations and responsible approach protocols, maintaining a minimal individual-array distance of 10-30 m.

The array includes three operated hydrophones (two SQ26 sensors spaced 50 cm apart horizontally, and one C75 positioned 50 cm below the horizontal plane and 15 cm forward; sampling frequency at 256 kHz, 24-bit resolution) and two synchronised cameras (GoPro Hero 10, 4K, 60 fps, distortion corrected) mounted on the same plane as the horizontal hydrophones (Fig. 1). This triangular geometry enables the estimation of time differences of arrival (TDOA) and directions of arrival (DOA).

Figure 1: *(Left) Surface interaction frame with multiple individuals. (Right) Compact OPALE audio-visual array.*

3.3. Acoustic pre-processing and automatic click detection

The processing chain is inspired by standard passive acoustic methods for odontocetes, primarily Zimmer (2011) [11,13,16,17,24]. Following calibration to standardise the relative sensitivities of the three channels and the application of a band-pass filter (at 3 kHz), the impulsive energy is extracted using the Teager-Kaiser Energy Operator (TKEO) [25,26]. To mitigate the variability of non-stationary noise, the energy trace undergoes adaptive normalisation using a local sliding-window median. An empirical threshold, determined from ROC curve analysis [27], was applied in conjunction with a leak integrator prior to binary thresholding. For each manually validated click (delimited by its onset/offset times), the channel with the highest signal-to-noise ratio was selected as the reference channel for subsequent analyses.

3.4. Spectral characterisation and temporal tracking of click trains

To address overlapping, an incremental multi-target tracking strategy was developed, inspired by data association methods. The adaptive time-frequency tracking algorithm reconstructs continuous trains by combining spectral validation and rhythmic prediction.

3.4.1. Spectral characterisation and dual memory

Each click is realigned on its energy peak, a constant-Q transform (CQT: 500Hz-32kHz, 12 bins/octave) extracts its logarithmic spectral signature [28]. The generated time-frequency patches (9 ms) are converted to dB, flattened, and then normalised (z-score), yielding a normalised spectral signature per click.

3.4.2. Similarity score

Assigning a candidate click to a track relies on a similarity score combining Pearson correlation, robust to temporal misalignment (jitter evaluated across three relative alignments), and cosine similarity (angular distance in the normalised space). The final visual score is a weighted combination. The system employs a dual memory: the candidate click is compared (i) to a dynamically updated template (the emitter's average profile) and (ii) to the last validated click of the track. This manages spectral variations induced by the animal's orientation (on/off-axis) [2,3,20].

3.4.3. Predictive rhythmic constraint (ICI)

In the absence of overlap, the ICI exhibit smooth physiological variations [17]. A temporal acceptance window is predicted based on the sliding median of the track's recent ICIs (25% relative tolerance). A quadratic penalty is applied if the temporal shift deviates from the expected rhythm.

3.4.4. Post-processing fusion

A posteriori phase reconnects interrupted track fragments (e.g., due to perfect superposition of two impulses) by verifying overall rhythmic consistency and spectral similarity at the junction points.

3.5. Spatial localisation and audio-video projection

Subsequent to the temporal and spectral grouping, a rigorous spatial selection is performed. Signals with excessively low energy (distant sources) are excluded. To guarantee unambiguous localisation in this highly proximal context, a strict 10 cm threshold is jointly applied to the loop error and geometric residuals. This dual filtering respectively validates the strict temporal against surface multipath effects and array hardware heterogeneity [11,16,29]. Although this stringent thresholding rejects numerous detections, the prior assignment of these clicks to coherent trains allows the entire sequence to be assigned based on its most reliable directional impulses.

For these validated clicks, inter-sensor TDOAs are estimated via cross-correlation in the frequency domain [11]. Under the assumption of a plane wave front (speed sound 1541 m/s, local conditions), the direction of arrival (DOA: azimuth and elevation) is calculated via non-linear least squares optimisation [14], and then corrected using a rotation matrix to align with camera's frame of reference [11].

Directional analysis is therefore articulated across three levels of interpretation. In the acoustic space (azimuth and elevation), the angular distribution of the clicks is modelled using kernel density estimation (KDE) [30]. This probabilistic approach identifies the dominant emission axis (the mode) and spatial concentration zones, whilst mitigating the effect of outliers inherent to multipath propagation [2]. In the image space, the DOA vectors are mathematically projected onto the calibrated image plane of the synchronised camera, incorporating its field of view (FOV). At the level of biological interpretation, this spatial superposition resolves the ambiguities of the compact array. Dissociation and individual assignment are visually confirmed when the projected detections consistently and stably intersect the position of an animal on screen,

following a frame-by-frame verification of temporal alignment across all remaining detections.

4. Results

Table 1: *Contextual details of the six analysed interaction scenes, including the acoustic tracking configuration, duration, and final individual assignment (ID) for each separated track.*

Scene	Track	Context	Duration (s)	Final ID
1	Single	1 speaker	8.4	Ali
2	Main	1 main spk.	9.7	Daren
	Sec.		< 1.0	Ali
3	Single	1 spk. (+dist.)	15.6	Ali
4	Main	1 spk. (+coda)	10.8	Ali
5	Main	2 speakers	11.2	Ali
	Sec.		4.0	Daren
6	Main	2 speakers	8.9	Daren
	Sec.		8.5	Ali

Table 2: *Acoustic and spatial metrics for each separated track within the analysed scenes. Mean values are reported for the Inter-Click Interval, Constant-Q Transform similarity, and energy. Spatial directions (azimuth and elevation) correspond to the dominant mode extracted via Kernel Density Estimation. 'Tot.' validated clicks; 'Filt.' clicks retained for localisation.*

Sc	Clicks (Tot /Filt)	ICI (s)	Sim. CQT	Energy (dB)	KDE mode (Az°/El°)
1	82/66	0.10	0.68	-31.0	9.7/60.7
2	32/30	0.31	0.72	-31.9	-0.5/64.0
	10/9	N/A	N/A	-35.2	7.1/9.3
3	76/67	0.21	0.87	-29.4	13.3/75.8
4	48/31	0.23	0.78	-31.8	42.0/67.4
5	78/67	0.15	0.80	-33.0	12.9/66.0
	13/10	0.33	0.83	-28.3	4.4/68.0
6	67/59	0.14	0.69	-32.2	-1.3/66.8
	42/20	0.21	0.72	-24.0	-20.9/69.9

4.1. Dataset and overall performance

The analysis encompassed 1354 validated clicks, extracted from two audio-visual recordings (total duration 3.12 min) and segmented into six interaction scenes (four with moderate overlap, two with heavy overlap).

The acoustic and spatial characteristics are summarised in Tables 1 and 2. Figure 2 illustrates the multimodal processing workflow, from the waveform to the spectral evolution (CQT) and the spatial density projection (KDE). The proposed multimodal approach enabled the coherent assignment of click trains to the visible individuals across all scenes, validating the methodological hypothesis.

4.2. Evaluation of deinterleaving metrics

4.2.1. Inter-Click Interval

With averages ranging from 0.104 s to 0.331 s, the ICI constitutes a robust segregation criterion when emission rhythms differ between emitters (Scenes 2 and 5). However, it becomes unstable during extreme overlaps where the inter-emitter delay is shorter than the impulse duration (Scene 6) or during intra-train decelerations (Scene 1).

4.2.2. Spectral similarity (CQT)

Intra-train consistency remains high (0.67 to 0.86), with dominant energy between 50-80 kHz. The low frequencies (<17 kHz) primarily reflect depth variations linked to multipath propagation. Nevertheless, the CQT is sensitive to background noise variations (Scene 4) and is poorly discriminative for perfectly superimposed clicks (Scene 6).

4.2.3. Energy

The energy differential (-24 to -38 dB) acts as a useful secondary criterion during extreme superpositions (Scene 6). However, its variations remain sensitive to directional effects (off-axis) and physical or environmental masking (swell).

4.2.4. Inter-Pulse Interval (IPI)

Estimable only in Scenes 1 and 2 (mean 2.87 and 2.96 ms) (confirming a juvenile stage of approximately 5 years). It proves unusable for separation; besides the similarity in the emitters' sizes, surface reflections and predominantly off-axis emissions degrade the internal structure of the pulses, generating aberrant or bimodal distributions.

4.2.5. Spatial localisation and optical validation

Isolated spatial localisation is insufficient to assign every click of a train to a specific emitter during compact interactions. Times of arrival (TDOA) suffer from geometric errors amplified by surface multipath effects [16] and array uncertainties [29]. Therefore, bearing estimation is only validated on previously clustered clicks that have passed a strict geometric threshold. Angular modelling (KDE) within the acoustic space then smooths out outliers to estimate a dominant emission axis per train. For instance, in Scene 2, the dominant axes of the two emitters are clearly disjoint. In Scenes 5 and 6, although the isopleths partially overlap, the KDE clearly discriminates two distinct modes along the azimuthal axis.

Ultimately, projection onto the video plane enables formal individual assignment: Ali is identified as the sole or primary emitter in Scenes 1, 3, and 5; Daren dominates Scenes 2 and 6; and Scene 4 isolates a background emitter. We note, however, a systematic elevation bias during very shallow emissions (Sc. 3 and 4), which is compensated for by the visual analysis.

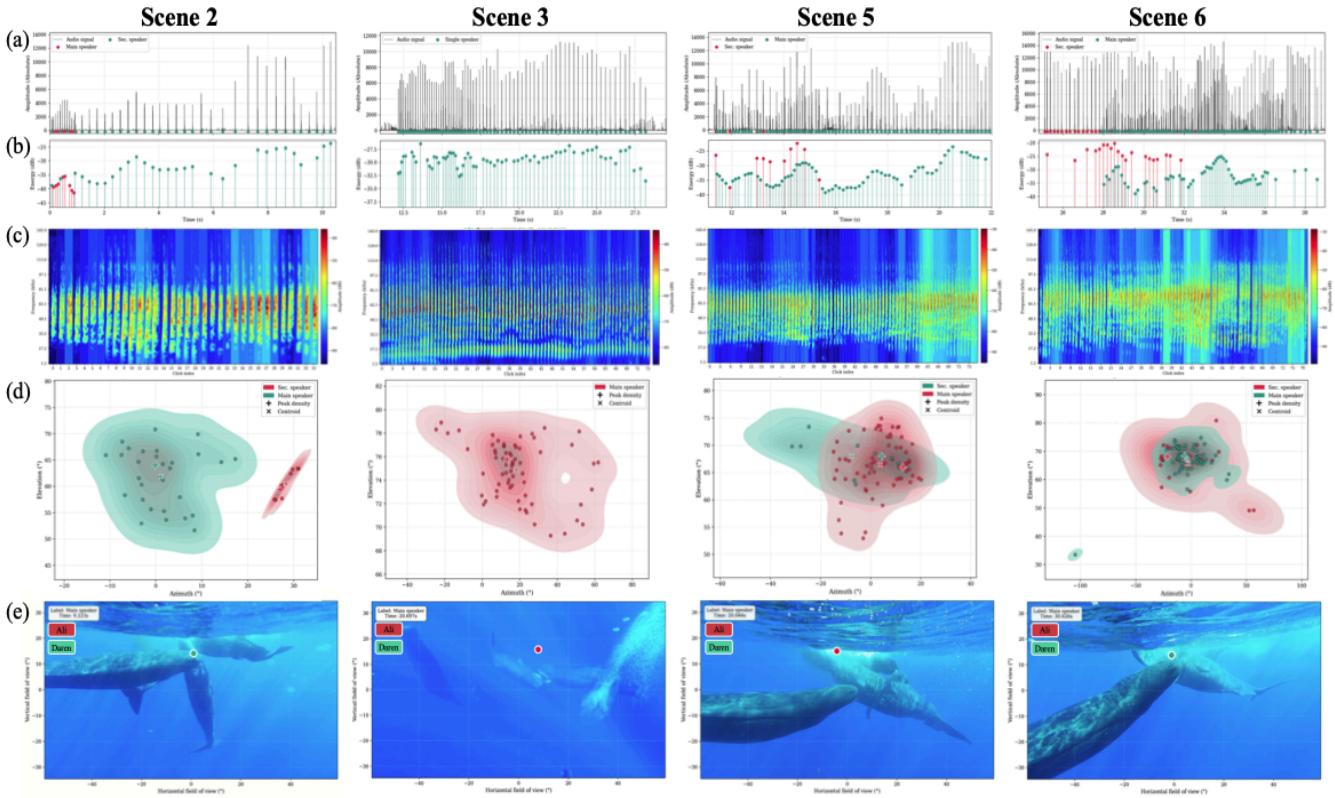

Figure 2: *Multimodal source separation workflow for four interaction scenes (2, 3, 5, 6). Top to bottom: (a) acoustic waveform (all validated clicks), (b) temporal energy profile (filtered subset), (c) CQT spectral evolution (click patches of the main speaker), (d) acoustic spatial distribution (azimuth/elevation) via KDE (filtered subset), and (e) optical validation projecting a selected click of the main speaker onto the video plane. Colour coding indicates individual assignment: teal for Daren, red for Ali.*

5. Discussion

5.1. Methodological contributions to near-field source separation

The multimodal approach structures highly constrained acoustic scenes (temporal overlaps, body shadowing, surface multipath) where classical tools reach their limits. The ICI emerges as a primary criterion, supported by spectral continuity (CQT), with energy intervening occasionally as a last resort. The study confirms that a fully automatic separation remains challenging with a compact array in a tactile social context. Body shadowing effects and directional variations limit the robustness of purely acoustic methods, making the contribution of video validation necessary.

5.2. Spectro-temporal variability and behavioural implications

The major contribution of this methodology is enabling the assignment of atypical signals (broad frequency band, variable rhythms, strong modulations, regularly emitted with an open jaw), suggesting very short-range intra-sex social communication. The CQT analyses indicate that spectral variability is primarily mechanical, driven by multipath propagation, body masking, and off-axis source effects. This can be contrasted with recent hypotheses of voluntary control raised in [31,32], which suggest that spectral changes in codas may result from deliberate contractions of the nasal complex (source-filter mechanism), akin to a complex phonology (coarticulation). Nevertheless, the rhythmic and energetic modulations observed during vocal sparring may reflect a degree of active control of the vocal repertoire and a potentially intentional dialogue structure.

Link to audio-video diarisation of sperm whale vocal sparring: https://cian.lis-lab.fr/pub/datainterspeech2026/

6. Conclusion and perspectives

This work presents a novel multimodal audio-visual workflow that overcomes the methodological bottleneck of individual assignment during intense surface socialisation phases in sperm whales. By combining spectro-temporal tracking and optical validation, the approach successfully isolated and characterised vocal sparring vocalisations in young males for the first time. Far from being merely chaotic signals, these highly modulated click trains suggest an advanced social and tactile function, possibly subject to active vocal control and complex coarticulation. Coupling with video offers an unprecedented opportunity to analyse the behaviour in its entirety, correlating body kinematics (e.g., jaw opening) with acoustic emissions. Better understanding these structured dialogues will open vast perspectives for the study of the species' social intelligence, and particularly the vocal evolution and learning of young males within matrilineal units.

7. Acknowledgements

We thank Longitude 181, Un Océan de Vie and Label Bleu productions for fieldwork, video recordings and data support. Thanks to N. Boodhonee for this valuable participation in the fieldwork. This study was conducted under the official authorization of the Mauritius Authority. The project “La Voix Des Cachalots” benefited from the support of the Mauritian Prime Minister’s Office, MCSEA (Dr. Réza Badal & team), AFRC (Mr. Satish Kadhun), MFDC (Mr. Sachin Jootun & Miss Eliana Timol), and the Tourism Authority (Miss Khoudijah Boodoo). We are grateful to Intelligent Acoustics/SMIoT, V. Gies, V. Barchasz and S. Marzetti for their help with OPALE recorder. This research was supported by ULP-COCHLEA ANR-21-CE04-0020-01, ADSIL AID ANR-20-CHIA-0014, CIAN https://cian.lis-lab.fr, LIS and Forza Media. We also warmly thank W.M.X. Zimmer and D.S. Pace for their support as CSI members.

8. Generative AI Disclosure

During article preparation, the authors used Google Gemini to assist with translation into English, style refinement, and assistance with coding and figure formatting. All AI-assisted outputs were carefully reviewed, validated, and edited by the authors, who assume full responsibility for the originality, accuracy, and scientific integrity of this publication.

9. References

[1] H. Whitehead, “Analyzing animal societies: quantitative methods for vertebrate social analysis,” Chicago, IL, USA: Chicago Press, 2008.

[2] B. Møhl, M. Wahlberg, P.T. Madsen, A. Heerfordt, and A. Lund, “The monopulsed nature of sperm whale clicks,” *J. Acoust. Soc. Am.*, vol. 114, no. 2, pp. 1143-1154, 2003.

[3] W.M.X. Zimmer, P.T. Madsen, V. Teloni, M.P. Johnson, and P.L. Tyack, “Off-axis effects on the multipulse structure of sperm whale usual clicks with implications for sound production,” *JASA V*118, 2005.

[4] P.T. Madsen, M. Wahlberg, and B. Møhl, “Male sperm whale (Physeter macrocephalus) acoustics in a high-latitude habitat: Implications for echolocation and communication,” *Behav. Ecol. Sociobiol.*, vol. 53, 2002.

[5] S.L. Watwood, P.J.O. Miller, M. Johnson, P.T. Madsen, and P.L. Tyack “Deep-diving foraging behaviour of sperm whales (Physeter macrocephalus),” *Journal of Animal Ecology*, vol. 75, pp. 814-825, 2006.

[6] G. Gubnitsky, Y. Mevorach, D. Tchernov and R. Diamant, “Source Separation of Sperm Whales' Echolocation Clicks,” in IEEE Transactions on Audio, Speech and Language Proc, V33, 4471-4485, 2025.

[7] L. Rendell, and H. Whitehead, “Vocal clans in sperm whales (Physeter macrocephalus),” *Royal Society B*, Biol. Sci., V270.1512, 2003.

[8] S. Gero, H. Whitehead, and L. Rendell, “Individual, unit and vocal clan level identity cues in sperm whale codas,” *Royal Society open science*, vol. 3, p. 150372, 2016.

[9] P. Giraudet, and, H. Glotin, “Real-time 3D tracking of whales by echo-robust precise TDOA estimates with a widely spaced hydrophone array,” *Applied Acoustics*, vol. 67, no. 11-12, pp. 1106-1117, 2006.

[10] H. Glotin, F. Caudal, and P. Giraudet, “Whale cocktail party: Real-time multiple tracking and signal analyses,” *International Journal Canadian Acoustics*, vol. 36, no. 1, 2008.

[11] W.M.X. Zimmer, “Passive Acoustic Monitoring of Cetaceans,” *Cambridge University Press*, pp. 1-356, 2011.

[12] M. Ferrari, H. Glotin, R. Marxer, V. Barchasz, V. Sarano, V. Giés, M. Asch, ”High-frequency near-field Physeter macrocephalus monitoring by stereo-autoencoder and 3d model of sonar organ”, in IEEE OCEANS, 2019.

[13] M. Ferrari, H. Glotin, M. Oger, R. Marxer, M. Asch, V. Gies, and F. Sarano, “3D diarization of a sperm whale click cocktail party by an ultra-high sampling rate portable hydrophone array for assessing individual cetacean growth curves,” *Forum Acusticum*, pp. 3239-3243, 2020.

[14] A. Thode, “Tracking sperm whale (Physeter macrocephalus) dive profiles using a towed passive acoustic array,” *J. Acoust. Soc. Am.*, vol. 116, no. 1, pp. 245-253, 1 July 2004.

[15] A. Thode, “Three-dimensional passive acoustic tracking of sperm whales (Physeter macrocephalus) in ray-refracting environments,” *J. Acoust. Soc. Am*., vol. 118, no. 6, pp. 3575-84, Dec. 2005.

[16] E.-M. Nosal, and N. Frazer, “Track of a sperm whale from delays between direct and surface-reflected clicks,” *Applied Acoustics,* vol. 67, pp. 1187-1201, 2006.

[17] P.M. Baggenstoss, “Separation of sperm whale click-trains for multipath rejection,” *J. Acoust. Soc. Am*., vol. 129, no. 6, 3598-609, 2011.

[18] J. Gordon, “Evaluation of a method for determining the length of sperm whales (Physeter catodon) from their vocalizations,” *J. Zool.*, vol. 2, pp. 301-314, 1991.

[19] M. Ferrari, M. Trinh, F. Sarano, V. Sarano, P. Giraudet, H. Glotin, “Age and interpulse interval relation from newborn to adult sperm whale (Physeter macrocephalus) off Mauritius,” *Sci. Rep.*, vol. 14, p. 18474, 2024.

[20] V. Teloni, W.M.X. Zimmer, M. Wahlberg, T.M. Peter, “Consistent acoustic size estimation of sperm whales using clicks recorded from unknown aspects,” *J. Cetacean Res. Manage*, vol. 9, pp. 127-136, 2007.

[21] L. Berkenbaum, F. Sarano, T. Flecher, W.M.X. Zimmer, O. Adam, R. Heuzey, A. Preud’homme, and H. Glotin, “Analysis of sperm whale (Physeter macrocephalus) dialogues, click by click: an ethoacoustic approach,” Poster One Ocean Science Congress 2025 Nice, France, 2025.

[22] L. Berkenbaum, H. Glotin, F. Sarano, O. Adam, R. Heuzey, and A. Preud’homme, “First description of ‘vocal sparring’ behaviour between juvenile male sperm whale (Physeter macrocephalus) off Mauritius,” Poster IBAC 2025 Kerteminde, Danmark, 2025.

[23] V. Sarano, F. Sarano, J. Girardet, A. Preud’homme, H. Vitry, R. Heuzey, M. Sarano, F. Delfour, H. Glotin, O. Adam, B. Madon, and J.L. Jung, “Underwater photo-identification of sperm whales (Physeter macrocephalus) off Mauritius,” *Marine Biology Research*, vol. 18, no.1-2, pp. 131-146, 2022.

[24] H. Glotin, N. Deloustal, A. Paiement, S. Paris, V. Gies, V. Barchasz, J-C. Vinaj, and F. Sarano, “Ethoacoustics of Free Ranging Cetaceans by High Resolution Portable Array: from P.m. to G.m. multipulsed clicks?,” Poster DCLDE 2024 Rotterdam, Netherlands, 2024.

[25] J.F. Kaiser, “On a simple algorithm to calculate the Energy of a signal”, Proc. IEEE Int. Conf. Acousti., Speech, Signal Process. (ICASSP), Albuquerque, NM, USA, pp. 381-384, 1990.

[26] 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.

[27] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861-874, 2006.

[28] J.C. Brown “Calculation of a constant Q spectral transform,” *J. Acousti. Soc. Am.*, vol. 89, no. 1, January 1991.

[29] M. Wahlberg, B. Møhl, P.T. Madsen, “Estimating source position accuracy of a large-aperture hydrophone array for bioacoustics,” *J. Acoust. Soc. Am.*, vol. 109, no. 1, pp. 397-406, 1 January 2001.

[30] B.J. Worton, “Kernel Methods For Estimating The Utilization Distribution In Home-Range Studies,” *Ecology*, vol. 70.1, 164-168, 1989.

[31] G. Beguš, R.L. Sprouse, A. Leban, M. Silva, S. Gero, “Vowel- and Diphthong-Like Spectral Patterns in Sperm Whale Codas,” Open Mind (Camb), vol. 2, no. 9, pp. 1849-1874, Nov. 2025.

[32] G. Beguš, M. Dabkowski, R.L. Sprouse, D.F. Gruber, S. Gero, “The phonology of sperm whale coda vowels”, *bioRxiv*, 2025.