LeAudioJEPA: Generic Audio Representation Learning Identifies Whale in Cocktail Party

Anonymous

Anonymous

ABSTRACT

Acoustic individual identification of Sperm Whales usually relies on temporal Inter-Pulse Intervals (IPI). In this work, we investigate whether if the whale coda sequences could also encode discriminative information about speaker identity. Therefore, we introduce LeAudioJEPA, an audio adaptation of LeJEPA based on self-supervised learning of predictive latent representations. The model is pre-trained on AudioSet and evaluated on a Sperm Whale coda dataset collected off Mauritius between 2013 and 2024, including both mono and multi-speaker contexts involving three identified individuals. We compare LeAudioJEPA with AudioJEPA, as well as supervised CNN and MLP baselines, on a multi-label speaker classification task using Mel spectrograms. Results show that LeAudioJEPA achieves the best average performance, with a particularly strong improvement in multi-speaker conditions. These findings suggest that coda sequence contain exploitable temporal context information for individual recognition, and highlight the potential of predictive latent architectures for low-annotation bioacoustic analysis.

***Index Terms*—** Self-Supervised Learning, LeAudioJEPA, Animal Communication, Bioacoustics, Sperm Whale, Individual signature, Multi-Speaker classification.

1. INTRODUCTION

Individual recognition from animal vocalizations is essential for non-invasive population monitoring and the study of social behavior. The Sperm Whale (*Physeter macrocephalus*) produces among the most powerful biological sounds on Earth: highly broadband echolocation clicks, with dominant energy between 2 and 20 kHz [1]. Their multipulsed biosonar acoustic structure is characterized by the Inter-Pulse Interval (IPI) that can be used to monitor each individual, even their growth [2]. Beyond their biosonar function, the clicks also play a central role in social communication through sequences known as codas, organized into Inter-Click Intervals (ICI) structures [3, 4, 5]. Codas are specific to each clan [6], and may convey individual signatures and contextual interaction patterns. Recent advances in bioacoustics have highlighted the combinatorial and contextual properties of these vocalizations, suggesting existence of structured communication system [7].

Fig. 1. Screenshot from a video of a dialogue between Vanessa and Caroline.

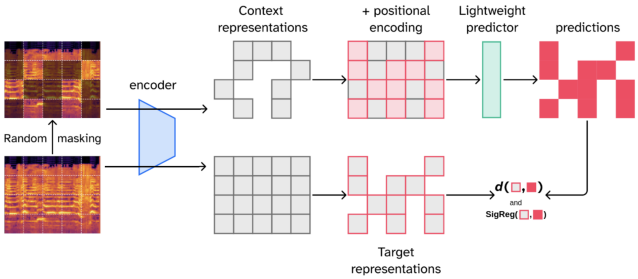

Fig. 2. LeAudioJEPA architecture. Compared with AudioJEPA, the EMA-based teacher-student mechanism is removed and the latent prediction objective is regularized with SIGReg (illust. inspired from [8]).

Over the past 10 years, the audiovisual field program conducted off Mauritius has enabled the recording and identification of numerous natural dialogues between individuals (Fig.1). This effort has produced a unique corpus of underwater communications combining acoustic recordings, visual identification, and behavioral annotations [9, 10, 11, 12]. However, the automatic analysis of these exchanges remains challenging due to acoustic variability, multi-speaker interactions, and the absence of models capable of efficiently capturing the temporal context underlying codas emissions. Moreover, IPI-based methods require high temporal resolution and precise pulse-level analysis in high SNR data. In this work, we investigate if the multi-speaker classification could be predicted from

Anonymous.

979-8-3503-2411-2/25/$31.00 ©2025 IEEE

time-frequency representations of codas, without providing IPI descriptors. Such an approach could enable lighter individual recognition on coda-based representation.

In this context, recent advances in self-supervised learning offer promising new perspectives. Joint-Embedding Predictive Architectures (JEPA) [13], initially developed for computer vision and later adapted to audio, learn robust contextual representations. Nevertheless, existing models adapted to Audio such as AudioJEPA [8] still present limitations in terms of training stability. We introduce LeAudioJEPA, an audio adaptation of LeJEPA for self-supervised predictive latent representation learning. Building on AudioJEPA, LeAudioJEPA replaces the teacher-student architecture and Exponential Moving Average (EMA) mechanism with SIGReg regularization, yielding a simpler and faster pre-training procedure.

Both AudioJEPA and LeAudioJEPA are pre-trained on AudioSet and evaluated on a multi-label speaker classification task using Mel spectrograms of Sperm Whale codas. It is well designed to investigate the contextual nature of Sperm Whale communication. Unlike conventional supervised approaches or probabilistic models relying on strong topological priors, LeAudioJEPA learns contextual representations directly from 30 s chunk of dialogues. The main objective in this paper is to evaluate whether LeAudioJEPA improves multi-speaker classification performance taking advantage of any temporal context compared to a flat MLP, or a short term CNN classifiers.

Experiments are conducted on a dataset collected off Mauritius between 2013 and 2024, including three identified individuals-Caroline, Delphine, and Vanessa in mono and multi-speaker contexts. They demonstrate that LeAudioJEPA in average consistently outperforms traditional supervised baselines (MLP and CNN) as well as AudioJEPA, with a particularly strong improvement in multi-speaker conditions, suggesting that predictive latent representations effectively exploit the contextual structure of coda sequences with several individual interactions. This suggests that contextual acoustic modeling can play a key role in the analysis of sperm whale communication and highlight the potential of self-supervised foundation models for the study of complex animal interactions.

Contributions are: (i) we adapt LeJEPA-style regularized latent prediction to audio spectrogram pre-training, yielding LeAudioJEPA, a simplified alternative to AudioJEPA without an EMA-based teacher-student mechanism; (ii) we evaluate LeAudioJEPA for individual Sperm Whale coda classification on full, mono-speaker, and multi-speaker test conditions; (iii) we show that large scale Mel-spectrogram representations contain individual-discriminative information.

2. RELATED WORK

2.1. Sperm Whale Codas, IPI, and Individual Recognition

In this dataset, codas were automatically detected and identified by a dedicated unsupervised model of ICI structures [5]. The IPI is used to recognition the individuals and thus to label each coda in the whale cocktail parties (Fig.3). Finally, the identity assessment is validated coupling the visual identification [14].

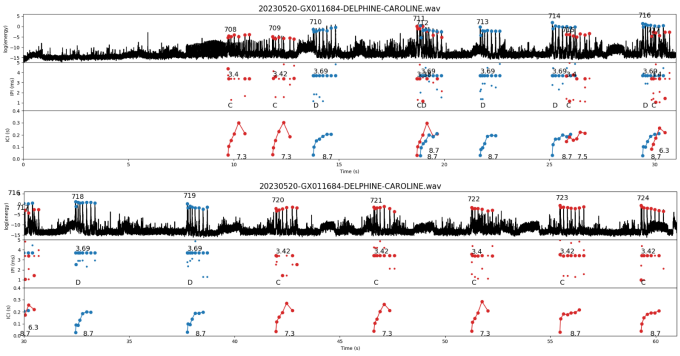

Fig. 3. A 60 s dialogue, Delphine with Caroline, in two consecutive excerpts: (Top) the click amplitude over time, each color is a speaker identity, (Middle) IPI of each click, (Bottom) ICI, highlighting temporal structure of codas.

2.2. Deep learning for audio representation learning

Deep learning has progressively replaced hand-crafted features in audio analysis with learned representations from time-frequency inputs such as log-Mel spectrograms. Convolutional Neural Networks were first widely adopted, followed by Transformer-based architectures, which better capture long-range spectro-temporal dependencies.

Recent advances have focused on Self-Supervised Learning (SSL), enabling large-scale pre-training without manual annotations. In speech, methods such as wav2vec 2.0 [15] and HuBERT [16] rely on masked prediction objectives, while approaches like AST [17] and AudioMAE [18] extend masked modeling to general audio using spectrogram patches and Transformer backbones.

Most existing SSL methods either reconstruct input features or rely on contrastive objectives. In contrast, Joint-Embedding Predictive Architectures (JEPA) [13] learn representations by predicting latent embeddings of masked regions, focusing on higher-level structure rather than low-level reconstruction. This paradigm is particularly relevant for bioacoustic signals, where discriminative information may lie in temporal organization and contextual patterns. Our work builds on AudioJEPA [8] and introduces LeAudioJEPA (Fig.2), an audio adaptation of LeJEPA [19] in which the EMA-based teacher-student mechanism is replaced by SIGReg regularization. This

yields a simpler pre-training scheme while preserving the same downstream encoder architecture.

3. MATERIALS

Two datasets are used: a large-scale generic audio corpus for self-supervised pre-training, and a domain-specific dataset for downstream evaluation on multi Whale classification.

Pre-training dataset: We use a subset of AudioSet [20], a large-scale collection of human-labeled audio events covering a wide variety of acoustic scenes. AudioSet consists of 10 s audio clips sampled at 32 kHz, annotated with a hierarchical ontology of sound events. In this work, we use 15% of the unbalanced training set, corresponding to approximately 833 h of audio. This dataset provides a diverse acoustic environment that enables the model to learn general-purpose audio representations.

Bioacoustic dataset:

The dataset used in this study comes from the *Voix du Cachalot* program, a long-term data collection effort conducted over 594 days between 2013 and 2024 [14]. It focuses on a social group of sperm whales from the clan of Irene Gueule Tordue off the coast of Mauritius. Each spring, the whales were recorded during surface socialization using synchronized underwater acoustic recordings and video data acquired with GoPro cameras and the OPALE system [21]. The downstream evaluation is conducted on the dataset depicted above, collected between 2013 and 2024, which focuses on the study of a social group of Sperm Whales off the coast of Mauritius. The selected dataset contains approximately 4 hours of underwater acoustic recordings, over 80 audio files of 2 min duration each in average. A subset of the recorded clicks and codas was manually annotated and attributed to individual whales. Speaker identity was established through manual inspection and cross-validated using synchronized video recordings. In total, 1123 codas were automatically annotated by unsupervised process, and checked. Among the 16 identified individuals, only 3 were retained in order to ensure a sufficient number of annotations per class for reliable model training. This results in a dataset of 451 codas corresponding to three individuals: Delphine, Vanessa, and Caroline (Tab.1). Each sample consists of a 30 s audio segment centered on a target coda to capture sufficient temporal context, including multiple codas and potential speaker interactions (Fig.4). Spectrograms are generated from these segments, but edge effects (when codas occur near the beginning or end of recordings) can produce duplicate samples. These duplicates are explicitly removed to avoid biasing the training process.

Train-Test Split: The train-test split is performed at the recording-file level rather than at the coda level to avoid temporal overlap between 30 s windows and to reduce the risk of exploiting recording-specific background conditions. The final split (Tab.1) is nearly balanced across 2013-2024.

Fig. 4. A sample of the dataset: a 30 s Mel spectrogram of a dialogue of Vanessa and Caroline. Each Coda is a group of large frequency band events. Only the 3rd and 6th groups are mono–speaker codas, others are interlinked.

Table 1. Speaker distribution in the 30 s chunks dataset.

Speaker config.	Full	Train	Test
Caroline only	115 (25%)	81 (22%)	34 (38%)
Delphine only	84 (19%)	74 (20%)	10 (11%)
Vanessa only	67 (15%)	50 (14%)	17 (19%)
Total mono speaker	266 (59%)	205 (56%)	61 (68%)
Delph. + Vane.	83 (18%)	69 (19%)	14 (16%)
Caro. + Vane.	49 (11%)	43 (12%)	6 (7%)
Caro. + Delph.	46 (10%)	38 (11%)	8 (9%)
Caro.+Delph.+Vane.	7 (2%)	7 (2%)	0 (0%)
Total multi speaker	185 (41%)	157 (44%)	28 (32%)
TOTAL Full dataset	451	362	89

4. METHOD

4.1. Model Architecture

LeAudioJEPA (Fig.2) is built upon AudioJEPA (Fig.5) by removing the teacher-student architecture and the Exponential Moving Average (EMA) mechanism, and by introducing the SIGReg regularization loss. This modification simplifies the training procedure and significantly reduces pre-training time, by approximately a factor of 2 in our experiments: ppre-training of LeAudioJEPA takes approximately 5.5 h on a single NVIDIA A40 GPU, compared to 11 h for AudioJEPA under identical conditions.

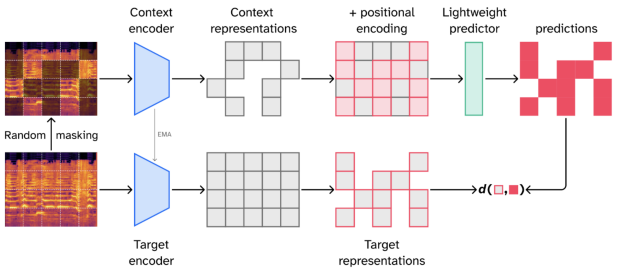

Fig. 5. AudioJEPA predicts latent representations of masked spectrogram patches from visible context representations [8].

For a fair comparison, both Audio-JEPA and LeAudioJEPA share the same backbone architecture. The encoder is a Vision Transformer (ViT) [22] with 12 layers and 12 attention heads

per layer. Input Mel spectrograms are divided into 16×16 patches and embedded into a latent space of dimension 768. The predictor module is also implemented as a ViT with 6 layers and 12 attention heads.

4.2. Training on the Upstream task with AudioSet

Both AudioJEPA and LeAudioJEPA are first pre-trained on the same subset of AudioSet, corresponding to 15% of the unbalanced training set. Audio clips of 10 seconds are converted into Mel spectrograms with $n_{\text{mels}} = 96$ and 512 time bins.

Pre-training is performed using a random patch masking strategy, where 40% to 60% of the spectrogram patches are masked uniformly across time and frequency. Models are trained with a batch size of 256 using the AdamW optimizer (weight decay 5%. Two stages learning rate: (i) first 1000 steps from 10^{-7} to 10^{-4}, (ii) then a cosine decay from 10^{-4} to 0). After pre-training, only the encoder is retained for the downstream multi-speaker classification task.

The AudioJEPA model is trained to predict the latent representations of randomly masked audio patches from the visible context. Let $\hat{\mathbf{z}}_{i,\ell} \in \mathbb{R}^D$ denote the predicted latent embedding for the ℓ-th masked patch of sample i, and let $\mathbf{z}_{i,\ell} \in \mathbb{R}^D$ denote the corresponding target embedding produced by the target encoder, the loss over the embedding dimension is:

$$\mathcal{L}_{\text{MSE}}(i,\ell) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{z}_{i,\ell,d} - z_{i,\ell,d}\right)^2. \tag{1}$$

The final loss is the average of this patch-wise reconstruction error over all samples and masked prediction patches:

$$\mathcal{L}_{\text{AudioJEPA}} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{\ell=1}^{L}\mathcal{L}_{\text{MSE}}(i,\ell). \tag{2}$$

For the LeAudioJEPA variant, as LeJEPA, we keep the same latent prediction objective and we add a SIGReg (Sketched Isotropic Gaussian Regularization) (Fig.6) term [19] for 256 random projection directions over the concatenation of predicted and target latent embeddings, with $\lambda = 0.01$:

$$\mathcal{L}_{\text{LeAudioJEPA}} = (1-\lambda)\mathcal{L}_{\text{AudioJEPA}} + \lambda\mathcal{L}_{\text{SIGReg}}. \tag{3}$$

4.3. Training the Downstream Multi-label Classification

For each annotated coda, a 30 s audio segment centered on the coda is extracted. A high-pass filter at 1 kHz is applied to the signal. Mel spectrograms are then computed using 128 Mel bands, a window size of 40 ms, and a hop size of 5 ms. To account for multi-speaker contexts, labels are assigned at the speaker level within each 30 s segment. A speaker is included in the target label set if at least 20% of the duration of one of their annotated codas overlaps with the selected 30-second window. This criterion allows codas located near the boundaries of the segment to be included only when a sufficient portion of the event is observed.

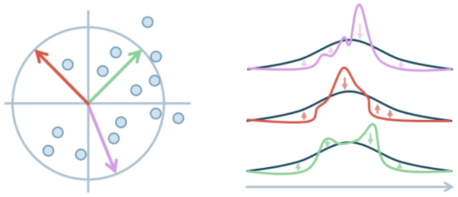

Fig. 6. To prevent trivial collapse, the SIGReg regularization term enforces Gaussian-distributed latent embeddings: (Left) latent embeddings are projected onto multiple random directions, (Right) normality test is applied to each one-dimensional projection: this encourages the full embedding distribution to match an isotropic Gaussian [23] and feature diversity.

Spectrograms are temporally downsampled via bilinear interpolation from 5993 to 1024 time bins, while keeping the 128 Mel-frequency bins unchanged. The resulting spectrograms are standardized using the mean and standard deviation computed on the training set, which are then applied to both training and test samples. A fixed threshold (0.5) converts predicted probabilities into binary labels for all models.

Baselines: We consider 2 fully supervised baselines trained from scratch on the same log-Mel spectrograms, an MLP and a CNN. The MLP consists of 2 fully connected layers with 512 and 256 units, ReLU activations, and dropout rate 25%, applied to flattened spectrograms. The CNN contains 4 convolutional blocks with 16, 32, 64, and 128 channels; each block applies a 3×3 convolution, batch normalization, ReLU activation, and 2×2 max pooling. The resulting feature map is flattened and passed through a 256-unit fully connected layer with dropout probability 30%, followed by a 3-output linear classifier. As (Le)AudioJEPA, both baselines are trained as multi-label classifiers over the 3 individuals using a binary cross-entropy loss with class-dependent positive weighting to mitigate class imbalance.

JEPA-based models: a linear classification head ($768 \rightarrow 3$) is added on top of the frozen encoder in the AudioJEPA and LeAudioJEPA, preceded by a layer normalization. All downstream models are trained as multi-label classifiers (sigmoid output) by this weighted binary cross-entropy loss:

$$\mathcal{L} = \frac{1}{C}\sum_{c=1}^{C}\left[-w_c y_c \log\sigma(z_c) - (1-y_c)\log(1-\sigma(z_c))\right], \tag{4}$$

where z_c is the logit for class c, $y_c \in \{0,1\}$ is the target label, and $w_c = N_c^{\text{neg}}/N_c^{\text{pos}}$ compensates for class imbalance. All models are trained for 50 epochs with AdamW, a learning rate of 10^{-3}, cosine decay, and a batch size of 64.

Model complexity, trainable and non-trainable parameters: AudioJEPA and LeAudioJEPA use the same frozen downstream encoder and classification head, while LeAudio-

JEPA reduces upstream pre-training complexity by removing the teacher-student architecture (Tab.2).

Table 2. Number of parameters per model (in Million).

Model	Pretrain upstream	Non-trainable downstream	Trainable downstream	Total downstream
MLP	0	0	67	67
CNN	0	0	16	16
AudioJEPA	187	85	0.004	85
LeAudioJEPA	85	85	0.004	85

5. EXPERIMENTS AND RESULTS

We evaluate all models on three variants of the same test set: the full, the mono-speaker, and the multi-speaker subset. For each experiment, performance is reported over the same 20 random seeds using identical train-test splits (Tab.3)

5.1. Full Test Set: Mono or Multi-Speaker Conditions

This experiment considers the complete test set, including both mono-speaker and multi-speaker conditions. On this full test set, LeAudioJEPA achieves the best average performance (both on AUC and F1 at 0.5), improving over AudioJEPA and the fully supervised CNN and MLP baselines.

5.2. Mono-Speaker versus Multi-Speaker Only Contexts

To better understand the behavior of supervised baselines compared to JEPA-based models, we further evaluate all models under two separate contexts. The first test set includes only mono-speaker samples, while the second includes only multi-speaker samples. The mono-speaker samples provide a simpler setting, where each segment contains only one identified speaker. In this context, performance differences are smaller and the MLP baseline remains competitive. In contrast, LeAudioJEPA shows an advantage on multi-speaker context, suggesting that its predictive representations better exploit contextual and interaction-related context.

6. DISCUSSION

Statistical significance is assessed across all the paired runs using a Friedman test followed by Wilcoxon signed-rank post-hoc comparisons. The Friedman test shows a significant model effect for both global metrics, with $p_{\mathrm{FDR}} = 1.32 \times 10^{-9}$ for average F1 and $p_{\mathrm{FDR}} = 7.08 \times 10^{-11}$ for global AUC. Significant effects are also obtained for all per-class metrics after correction. LeAudioJEPA significantly improves over AudioJEPA on the global metrics and for Caroline and Vanessa, while AudioJEPA remains significantly better for Delphine. The difference between CNN and LeAudioJEPA for Delphine is not significant. LeAudioJEPA achieves the best average performance on the full multi-label speaker classification task, with its clearest advantage in multi-speaker settings. In mono-speaker conditions, performance remains closer to that of supervised baselines, suggesting that the main benefit of the proposed model lies in exploiting contextual information within coda sequences rather than simply improving isolated-speaker classification. This is consistent with JEPA-based learning, which predicts latent representations from partially observed context. In 30 s windows where multiple codas and speakers may co-occur, discriminative information likely depends on temporal organization and acoustic context. LeAudioJEPA appears to better exploit these structures on average, especially in multi-speaker conditions.

Performance remains heterogeneous across individuals. For Delphine, supervised baselines and AudioJEPA remain highly competitive and in several settings outperform LeAudioJEPA, possibly due to specific individual interactions or recording conditions. This result, as well as the multi-speaker evaluation, should be interpreted with caution given the limited number of test samples, but in average over all conditions, LeAudioJEPA demonstrates that Sperm Whale multi-speaker classification can be processed from low-resolution time-frequency representation on 30 s chunk cocktail party.

Table 3. Speaker identification Area Under the Receiver Operating Characteristic curve (AUC), and F1 (at 0.5), mean $\pm$ STD over 20 seeds, in all speaker contexts, then in mono-only and multi-only speaker test sets.

Metric	Random	MLP	CNN	AudioJEPA	LeAudioJEPA
Mono+Multi					
AUC(Caro.)	50.4 ± 5	53.1 ± 6	52.7 ± 5	61.2 ± 7	$\mathbf{69.2 \pm 2}$
AUC(Delph.)	50.8 ± 5	$\mathbf{90.6 \pm 3}$	80.8 ± 2	87.1 ± 2	78.5 ± 4
AUC(Vane.)	50.1 ± 6	59.5 ± 2	64.7 ± 6	77.8 ± 5	$\mathbf{84.2 \pm 1}$
AUC(All)	50.4 ± 5.4	67.7 ± 4	66.1 ± 4	75.4 ± 4	$\mathbf{77.3 \pm 2}$
F1(Caro.)	53.4 ± 5	60.4 ± 4	59.7 ± 1	57.5 ± 7	$\mathbf{69.3 \pm 3}$
F1(Delph.)	41.9 ± 5	$\mathbf{69.6 \pm 5}$	65.2 ± 3	67.9 ± 3	63.2 ± 7
F1(Vane.)	44.2 ± 5	57.7 ± 2	44.7 ± 7	57.0 ± 1	$\mathbf{71.2 \pm 5}$
F1(All)	46.5 ± 5	62.6 ± 4	56.6 ± 4	60.8 ± 4	$\mathbf{67.9 \pm 5}$
Mono only					
AUC(All)	52.3 ± 8	$\mathbf{82.6 \pm 3}$	74.3 ± 3	79.1 ± 3	78.2 ± 3
F1(All)	39.7 ± 6	$\mathbf{61.8 \pm 4}$	55.0 ± 5	59.7 ± 4	59.7 ± 6
Multi only					
AUC(All)	48.6 ± 13	45.7 ± 4	44.9 ± 6	64.2 ± 12	$\mathbf{71.4 \pm 5}$
F1(All)	56.0 ± 12	56.7 ± 4	50.0 ± 4	55.3 ± 6	$\mathbf{70.9 \pm 7}$

7. CONCLUSION

We introduced LeAudioJEPA, an adaptation of the AudioJEPA architecture incorporating the principles of LeJEPA for self-supervised audio representation learning. The model was evaluated on a multi-label speaker classification task using sperm whale coda spectrograms. Results show that LeAudioJEPA achieves the best average performance on the full test set and provides the clearest gains in multi-speaker settings. This

suggests that LeAudioJEPA, with predictive latent representations, better captures possible inter-individual interactions resulting into temporal structures of possibly interlinked sequences of codas. The frequency content of the clicks is poorly represented in our experiments and is not known to be speaker discriminant, thus our findings suggest that temporal representations of codas may contain exploitable individual information, even when IPI are not provided. Future work will focus on scaling the evaluation to a larger number of individuals, assessing cross-session and inter-annual generalization, and further analyzing the representations learned by LeAudioJEPA by heatmaps for example. In the short term, our approach may contribute to the development of underwater communication strategies.

8. ACKNOWLEDGMENTS

Permission to record was granted by the Mauritius Prime Minister's Office on Feb. 2017, and before through the documentaries (permits MFDC/PP/3/2015 by the Mauritian Gov. and the Mauritius Film Dev. Corp. (MDMC); MFDC/PP/68/2015-/58/2018-/46/2019-/10/2020). The videos were recorded according to the respect of the official Charter for responsible approach and observation of marine mammals and the Maritime zones regulations (Conduct of Marine Scientific Res. n57.2017) promulgated by the Mauritian Gov. We thank the authority of Mauritius who facilitated the program, the Prime Minister Office of the Republic of Mauritius, the dpt for Continental Shelf, Maritime Zones Admin. and Exploration (CSMZAE, Dr. R. Badal), the Albion Fisheries Res. Center (AFRC; Chief Scientific officer Mr. S. Kadhun), the Mauritius Film Dev. Corp. (MFDC; Mr. S. Jootun and Miss E. Timol) and the Tourism Authority (TA; Miss K. Boodoo, Dir.).

9. REFERENCES

[1] W. M. X. Zimmer, *Passive Acoustic Monitoring of Cetaceans*, Cambridge Univ., 2011.

[2] M. Ferrari, M. Trinh, F. Sarano, V. Sarano, P. Giraudet, et al., "Age and IPI relation from newborn to adult Sperm Whale off Mauritius," *Nat. Scientific Reports*, 2024.

[3] H. Whitehead and L. Weilgart, "Patterns of visually observable behaviour and vocalizations in groups of female Sperm Whales," *Behaviour*, vol. 118, pp. 275–96, 1991.

[4] P. Sharma, S. Gero, R. Payne, D. Gruber, and et. al, "Contextual and combinatorial structure in Sperm Whale vocalisations," *Nature Communications*, vol. 15, 2024.

[5] Anonymous, "in submission," 2026.

[6] L. Rendell and H. Whitehead, "Vocal clans in Sperm Whales (P.m.)," *P. Royal Soc. London*, pp. 225–31, 2003.

[7] P. Madsen et al., "Sperm Whale sound prod. studied with ultrasound time/depth-recording tags," *Exp. Bio.*, 2002.

[8] L. Tuncay et al., "Audio-JEPA: Joint-Embedding Predictive Archit. for Audio Rep. Learning," in *ICME*, 2025.

[9] L. Berkenbaum, P. Giraudet, et al., "Exploring coda repertoires in two recently separated Sperm Whale social units off Mauritius," DCLDE, https://univ-tln.hal.science/hal-04937640v1, 2024.

[10] V. Sarano et al., "Underwater photo-identification of Sperm Whales off Mauritius," *Mar. Bio. R.*, V18, 2022.

[11] F. Sarano et al., "A focal animal 6-points Likert scale to rate intra-unit interactions in sperm whales (P.m.) off Mauritius Island," in *World MM C.*, 2019.

[12] F. Sarano et al., "Nursing behavior in Sperm Whales," *Animal Behavior & Cog.*, vol. 10, pp. 105–31, 2023.

[13] Yann LeCun, "A path towards autonomous machine intelligence," Tech. Rep., OpenReview, 2022.

[14] F. Sarano, V. Sarano, and P. Giraudet, "The cachalot voice," `https://www.longitude181.org/programme-cetaces-cachalots`, 2023.

[15] A. Baevski et al., "WAV2VEC2.0 a Framework for S. S. Learning of Speech Representations," in *NeurIPS*, 2020.

[16] W. Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, 2021.

[17] Y. Gong et al., "AST: Audio Spectrogram Transformer," in *Interspeech*, 2021.

[18] D. Niizumi and et al, "Masked spectrogram modeling using masked autoencoders for audio representation learning," in *HEAR Workshop*, 2022.

[19] R. Balestriero and Y. LeCun, "LeJEPA: Provable and scalable self-supervised learning without the heuristics," *arXiv:2511.08544*, 2025.

[20] J. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[21] Maxence Ferrari et al., "High-frequency Near-field Physeter macrocephalus Monitoring by Stereo-Autoencoder and 3D Model of Sonar Organ," in *IEEE OCEANS*, 2019.

[22] A. Dosovitskiy and et al, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[23] L. Maes et al., "Leworldmodel:stable end-to-end joint-embedding predictive architecture from pixels," *arXiv:2603.19312*, 2026.