LeAudioJEPA: Generic Audio Representation Learning for Identifying Whales in Cocktail Parties

Adam Chareyre, Pascale Giraudet, Maha Mellal, Sébastien Paris, Randall Balestriero, Véronique Sarano, François Sarano, Hervé Glotin

CIAN RESEARCH REPORT

CIAN-RR-20260814

LeAudioJEPA: Generic Audio Representation Learning For Identifying Whales in Cocktail Parties

Adam Chareyre[1,4,*], *Pascale Giraudet[1,4], Maha Mellal[1,4], Sébastien Paris[1,4], Randall Balestriero[2,*], Véronique Sarano[3,4], François Sarano[3,4,*], Hervé Glotin[1,4,*]*

1. Univ Toulon, Aix Marseille Univ, CNRS, LIS, NATAL, Toulon, France
2. Brown University, Department of Computer Science, USA
3. Longitude 181, Toulon, France
4. Univ Toulon, CIAN, Toulon, France

*Corresponding authors: adam.chareyre.1999@gmail.com, randallbalestriero@gmail.com, saranofrancois@gmail.com, glotin@univ-tln.fr

CIAN Research Report 20260814

Similar content is accepted for publication in the 36th Annual IEEE Workshop on Machine Learning for Signal Processing 2026.

Int. Center of Artificial Intelligence in Natural Acoustics

Research Fed. RNSR 202424556S
av. de l'Université
CS 60584, 83041,Toulon, FR

https://cian.lis-lab.fr

LeAudioJEPA: Generic Audio Representation Learning for Identifying Whales in Cocktail Parties

Adam Chareyre[1,4,*], Pascale Giraudet[1,4], Maha Mellal[1,4], Sébastien Paris[1,4], Randall Balestriero[2,*], Véronique Sarano[3,4], François Sarano[3,4,*], and Hervé Glotin[1,4,*]

[1]Univ Toulon, Aix Marseille Univ, CNRS, LIS, NATAL, Toulon, France
[2]Brown University, Department of Computer Science, USA
[3]Longitude 181, Valence, France
[4]Univ Toulon, CIAN, Toulon, France
[*]Corresponding authors: adam.chareyre.1999@gmail.com, randallbalestriero@gmail.com, saranofrancois@gmail.com, glotin@univ-tln.fr

August 24, 2026

Abstract

Acoustic individual identification of sperm whales usually relies on temporal Inter-Pulse Intervals (IPI). In this work, we investigate whether sperm whale coda sequences could also encode discriminative information about speaker identity. To address this question, 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 2015 and 2024, including both mono-speaker and multi-speaker contexts involving three identified individuals. We compare LeAudioJEPA with AudioJEPA, as well as supervised MLP, CNN, and ViT baselines trained from scratch, on a multi-label speaker classification task using Mel spectrograms. Results show that LeAudioJEPA achieves the best average macro-F1 score across full, mono-speaker, and multi-speaker test conditions over 20 seeds. Statistical testing over the 20 paired random seeds confirms that LeAudioJEPA significantly outperforms all evaluated trained baselines after FDR correction. These findings suggest that coda sequences contain exploitable temporal context information for individual recognition, and highlight the potential of predictive latent architectures for low-annotation bioacoustic analysis.

Keywords: 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 studying social behavior. The sperm whale (*Physeter macrocephalus*) produces some of 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), which can be used to monitor individuals and even track their growth [2]. Beyond their biosonar function, the clicks 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 the existence of a structured communication system [4].

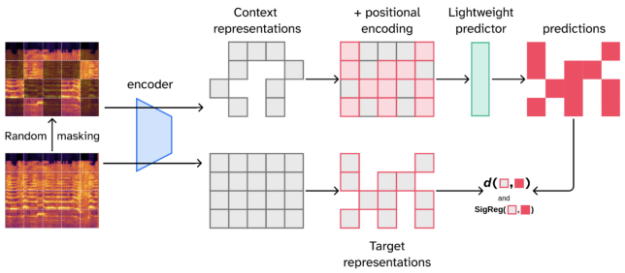

Figure 1: LeAudioJEPA: compared with AudioJEPA [7], the EMA-based teacher-student mechanism is removed and the latent prediction objective is regularized by SIGReg.

Over the past 10 years, the audiovisual field program conducted off Mauritius has enabled the recording and identification of numerous natural dialogues between individuals. This effort has produced a unique corpus of underwater communications combining acoustic recordings, visual identification, and behavioral annotations [8, 9, 10, 11]. 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 coda emissions. Moreover, IPI-based methods require high temporal resolution and precise pulse-level analysis in high-SNR data. In this work, we investigate whether multi-label speaker identities can be predicted from time-frequency representations of codas, without providing IPI descriptors. Such an approach could enable lightweight individual recognition from coda-based representations.

In this context, recent advances in self-supervised learning offer promising new perspectives. Joint-Embedding Predictive Architectures (JEPA) [12], initially developed for computer vision and later adapted to audio, learn robust contextual representations. Nevertheless, existing models adapted to audio such as AudioJEPA [7] still present limitations in terms of training stability. We introduce LeAudioJEPA, an audio adaptation of LeJEPA for self-supervised predictive latent representation learning. Importantly, LeAudioJEPA retains the AudioJEPA backbone; our contribution lies in the pre-training methodology rather than in the encoder architecture. We replace AudioJEPA's EMA-based teacher-student mechanism with SIGReg-based collapse prevention, resulting in a simpler and more computationally efficient pre-training procedure.

Both AudioJEPA and LeAudioJEPA are pre-trained on AudioSet and evaluated on a multi-label speaker classification task using log-Mel spectrograms of sperm whale codas. This setup is 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 chunks of dialogue. The main objective of this paper is to evaluate whether LeAudioJEPA improves multi-label speaker classification by exploiting temporal context, compared with supervised MLP, CNN, and ViT baselines trained from scratch.

Experiments use data collected off Mauritius between 2015 and 2024 from three identified individuals (Caroline, Delphine, and Vanessa) in mono- and multi-speaker contexts. Over 20 paired seeds, LeAudioJEPA achieves the highest mean macro-F1 in all three evaluation conditions and significantly outperforms all trained baselines after FDR correction.

Contributions are: (i) a LeJEPA-style pre-training scheme for the AudioJEPA backbone that replaces the EMA teacher-student mechanism with SIGReg, reducing upstream parameters from 183.6M to 97.6M and pre-training time from 11 h to 5.5 h; and (ii) evaluation on mono- and multi-speaker sperm-whale coda classification, showing that large-scale log-Mel pre-training yields speaker-discriminative representations.

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]. IPI is used to recognize individuals and thus to label each coda in whale cocktail-party recordings (Fig. 2). Finally, the identity assessment is validated by coupling IPI-based acoustic assignments with visual identification [13].

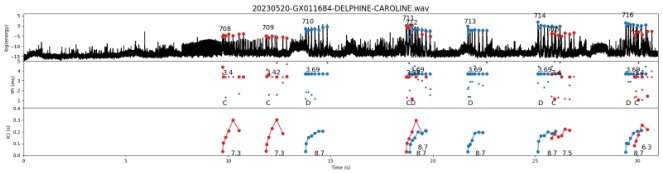

Figure 2: A 30 s dialogue, Delphine with Caroline: (Top) the click amplitude over time, each color is a speaker identity, (Middle) IPI of each click, (Bottom) ICI, highlighting temporal structure of codas.

2.2 Deep learning for audio representation learning

Deep audio models increasingly learn representations directly from time-frequency inputs. SSL methods such as wav2vec 2.0 [14], HuBERT [15], AST [16], and AudioMAE [17] reduce reliance on labels through masked or predictive objectives. Recent work in cetacean bioacoustics has further shown that self-supervised representations can improve robustness under limited annotations and noisy acoustic conditions [18]. JEPA [12] instead predicts latent embeddings of masked regions, emphasizing higher-level contextual structure. Building on AudioJEPA [7], LeAudioJEPA follows LeJEPA [19] by replacing the EMA teacher-student mechanism with SIGReg while keeping the same downstream encoder architecture (Fig. 1).

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-speaker 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 2015 and 2024 [13]. 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 selected subset contains approximately 4 hours of underwater acoustic recordings, over 130 audio files of 2 min duration each on average. A subset of the recorded clicks and codas was manually annotated

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

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 using an unsupervised process and manually 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 non-duplicated 30 s samples centered on codas corresponding to three individuals: Delphine, Vanessa, and Caroline (Table 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. 3). Spectrograms are generated from these segments, but edge effects (when codas occur near the beginning or end of recordings) can produce duplicate samples that are explicitly removed to avoid biasing the training process.

Figure 3: A sample of the dataset: a 30 s log-Mel spectrogram of a dialogue of Vanessa and Caroline. Each coda is a group of broadband click events. Only the 3rd and 6th groups are mono–speaker codas, others are interlinked.

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 preserves the recording-file-level separation and covers recordings from 2015 to 2024, while maintaining a comparable speaker distribution between training and test sets (Table 1).

4 Method

4.1 Model Architecture

LeAudioJEPA (Fig. 1) modifies AudioJEPA [7] 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: pre-training of LeAudioJEPA takes approximately 5.5 h on a single NVIDIA A40 GPU, compared to 11 h for AudioJEPA under identical conditions.

For a fair comparison, both AudioJEPA 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 log-Mel spectrograms are divided into 16×16 patches and embedded into a latent space of dimension $K = 768$. 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 log-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. We also evaluated a structured block-masking strategy using contiguous spectro-temporal regions. This alternative did not improve downstream performance over uniform random patch masking, and we therefore retain the 40–60% random masking strategy in the reported experiments. Models are trained with a batch size of 256 using the AdamW optimizer with a weight decay of 5%. The learning-rate schedule has two stages: (i) a linear warm-up during the first 1000 steps from 10^{-7} to 10^{-4}, followed by (ii) 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 $\hat{\mathbf{z}}_{i,\ell} \in \mathbb{R}^K$ denote the predicted latent embedding for the ℓ-th masked patch of sample i, and let $\mathbf{z}_{i,\ell} \in \mathbb{R}^K$ denote the corresponding target embedding produced by the target encoder. For N samples and L masked patches per sample, the loss is:

$$\mathcal{L}_{\text{AudioJEPA}} = \frac{1}{NLK} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \sum_{k=1}^{K} \left(\hat{z}_{i,\ell,k} - z_{i,\ell,k}\right)^2 . \tag{1}$$

Following LeJEPA [19], LeAudioJEPA keeps the predictor-based latent embedding objective used during pre-training, but removes the EMA teacher-student

mechanism. Collapse prevention is instead handled by adding a SIGReg (Sketched Isotropic Gaussian Regularization) term (Fig. 4) to the prediction loss. Let $\mathcal{A} = \{a_m\}_{m=1}^{M}$ denote the set of random projection directions used by SIGReg. We use M=256 directions over the concatenation of predicted and target latent embeddings, with $\lambda = 0.025$:

$$\mathcal{L}_{\text{LeAudioJEPA}} = (1-\lambda)\mathcal{L}_{\text{AudioJEPA}} + \lambda\mathcal{L}_{\text{SIGReg}}. \tag{2}$$

We additionally evaluated a LeAudioJEPA configuration using $M = 1024$ random projections together with a [-5,5] integration domain. Because these two hyperparameters were modified jointly, their individual effects cannot be isolated. This configuration did not provide a meaningful downstream improvement while increasing pre-training cost. We therefore retain $M = 256$ projections in the reported model.

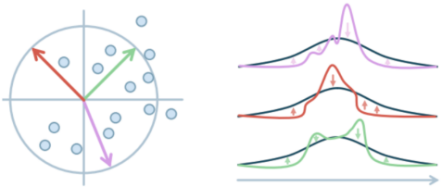

Figure 4: To prevent trivial collapse, the SIGReg regularization term enforces Gaussian-distributed latent embeddings: (left) latent embeddings are projected onto multiple random directions, (right) normality test is applied to each one-dimensional projection: this encourages the full embedding distribution to match an isotropic Gaussian [19] thereby promoting feature diversity.

4.3 Training the Downstream Multi-label Classification

For each annotated coda, a 30 s audio segment centered on the coda is extracted. Log-Mel filterbank spectrograms are computed from the 30 s audio segments using 128 Mel bands, a window size of 40 ms, a hop size of 5 ms, and a lower frequency bound of 1 kHz. 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.

Spectrograms are temporally downsampled from 5993 to 1024 time bins while retaining 128 Mel bins. The resulting 29.3 ms temporal resolution is substantially coarser than the typical 3–3.8 ms female sperm-whale IPI scale [2], so pulse-level IPI timing is not explicitly preserved, although coarser click-related cues may remain. 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 of 0.5 is chosen a priori for all classes and models. We additionally evaluated class-specific thresholds estimated from out-of-fold predictions using 5-fold GroupKFold on training recordings, maximizing Matthews correlation coefficient (MCC) independently for each class. The resulting thresholds were frozen before test evaluation.

Baselines: We compare against supervised MLP, CNN, and ViT models trained from scratch on the same log-Mel spectrograms. The MLP uses two hidden layers of 512 and 256 units. The CNN comprises six convolutional blocks with 32–256 channels, batch normalization, ReLU, max pooling, and alternating rectangular time- and frequency-oriented kernels, followed by adaptive average pooling and a linear classifier. The ViT uses the same 12-layer, 768-dimensional encoder architecture as the JEPA models but is randomly initialized and trained end-to-end. All baselines use weighted binary cross-entropy to account for class imbalance.

JEPA-based models: for AudioJEPA and LeAudioJEPA, the pre-trained encoder is frozen and produces frame-level latent representations of dimension 768. A lightweight trainable classification head is then added on top of these frozen features. This head first applies attentive temporal pooling over the frame-level embeddings, followed by layer normalization and a final linear classifier mapping the pooled 768-dimensional representation to the three speaker logits. All downstream models are trained as multi-label classifiers with sigmoid outputs using the weighted binary cross-entropy loss:

$$\mathcal{L} = \frac{1}{C} \sum_{c=1}^{C} \left[-w_c y_c \log \sigma(z_c) - (1 - y_c) \log(1 - \sigma(z_c)) \right], \tag{3}$$

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. For the JEPA-based models, all reported seeds correspond to independent downstream training runs using fixed pre-trained checkpoints.

Model complexity, trainable and non-trainable parameters: AudioJEPA and LeAudioJEPA use the same frozen downstream encoder and the same trainable classification head, while LeAudioJEPA reduces upstream pre-training complexity by removing the teacher-student architecture (Table 2).

5 Experiments and Results

We evaluate all models on three variants of the same test set: the full, the mono-speaker subset, and the multi-speaker subset. For each experiment, macro-F1 is reported over the same 20 random seeds using the same train-test split (Table 4). A random baseline is also reported for reference but is excluded from the statistical tests. For each test condition, it generates binary predictions at random using the same decision format as the trained models, providing a chance-level comparison for the multi-label speaker identification task.

Table 2: Number of parameters per model, in millions. The upstream column reports the number of parameters instantiated during pre-training. Downstream columns report frozen and trainable parameters during the supervised speaker classification stage.

Model	**Pretrain upstream**	**Non-trainable downstream**	**Trainable downstream**	**Total downstream**
MLP	0	0	67	67
CNN	0	0	2.3	2.3
ViT	0	0	86	86
AudioJEPA	183.6	86	0.005	86
LeAudioJEPA	97.6	86	0.005	86

Table 3: Ablation of the SIGReg regularization weight λ. Macro-F1 on the full test set is reported over 20 seeds.

λ	0	0.01	0.015	0.02	**0.025**	0.03	0.05
Mean (%)	32.5	67.9	59.5	67.5	**71.8**	59.4	60.4
Std. (%)	13.8	4.1	3.0	2.3	**4.0**	1.4	2.5

5.1 Ablation on the SIGReg Weight

We first study the influence of the SIGReg regularization weight λ during LeAudioJEPA pre-training. Table 3 shows that performance is sensitive to λ: removing SIGReg ($\lambda = 0$) reduces macro-F1 to 32.5 ± 13.8, whereas $\lambda = 0.025$ gives the best result (71.8 ± 4.0), confirming the need for moderate regularization.

5.2 Full Test Set: Mono and Multi-Speaker Conditions

This experiment considers the complete test set, including both mono-speaker and multi-speaker conditions. On the full test set, LeAudioJEPA achieves the best macro-F1 (71.8 ± 4.0), ahead of CNN (66.9 ± 4.9) and the other baselines (Table 4). Validation-tuned class-specific thresholds did not improve over 0.5, which is therefore used throughout.

5.3 Mono-Speaker versus Multi-Speaker Only Contexts

LeAudioJEPA also ranks first on the mono-speaker subset (64.7 ± 4.4) and multi-speaker subset (77.3 ± 6.9). CNN is competitive in the multi-speaker condition (74.1 ± 8.7), suggesting that both local spectro-temporal structure and predictive contextual representations contribute to identification.

Table 4: Speaker identification macro-F1 score at a fixed threshold of 0.5, reported as mean ± standard deviation over 20 random seeds, for the full test set, the mono-speaker subset, and the multi-speaker subset.

Metric	**Random**	**MLP**	**CNN**
F1(Mono+Multi)	46.5 ± 2.9	61.3 ± 2.1	66.9 ± 4.9
F1(Mono only)	39.7 ± 5.9	60.5 ± 2.8	54.9 ± 3.3
F1(Multi only)	56.0 ± 12.1	55.6 ± 2.2	74.1 ± 8.7
Metric	**ViT**	**AudioJEPA**	**LeAudioJEPA**
F1(Mono+Multi)	60.6 ± 5.0	60.0 ± 2.6	$\mathbf{71.8 \pm 4.0}$
F1(Mono only)	55.3 ± 4.7	58.7 ± 1.7	$\mathbf{64.7 \pm 4.4}$
F1(Multi only)	61.4 ± 9.8	54.8 ± 5.2	$\mathbf{77.3 \pm 6.9}$

6 Discussion

Statistical significance was assessed on the 20 paired macro-F1 scores obtained from the 20 random seeds, using a Friedman test followed by Wilcoxon signed-rank post-hoc comparisons with FDR correction, considering only the five trained models. The Friedman test showed a significant model effect on the mean macro-F1 score over 20 random seeds ($p = 1.62 \times 10^{-9}$). LeAudioJEPA significantly outperformed all trained baselines, including CNN (+4.9 macro-F1, $p_{FDR} = 9.65 \times 10^{-3}$) and AudioJEPA (+11.8 macro-F1, $p_{FDR} = 8.00 \times 10^{-6}$).

These results indicate that predictive latent pre-training transfers effectively to sperm whale speaker identification despite the limited annotated dataset. The largest descriptive gain appears in multi-speaker segments, where LeAudioJEPA reaches 77.3 ± 6.9 macro-F1, while CNN remains competitive. This suggests that both local spectro-temporal structure and longer contextual cues contribute to speaker discrimination. The absence of improvement from validation-tuned thresholds may reflect the limited amount of labeled training data, which makes class-specific threshold estimates more variable across folds.

The λ ablation further shows that SIGReg is critical to LeAudioJEPA pre-training: removing the regularization ($\lambda = 0$) leads to a large performance drop, while an intermediate weight of $\lambda = 0.025$ provides the best downstream result.

Because the temporally downsampled log-Mel inputs do not preserve pulse-level IPI timing explicitly, the proposed approach should be viewed as complementary to IPI-based identification rather than a replacement. The learned representations may nevertheless capture coarser spectro-temporal patterns that correlate indirectly with click structure. A direct analysis of feature–IPI relationships would therefore be valuable future work to disentangle explicit timing cues from broader coda-context information. These results suggest that coda sequences contain speaker-discriminative temporal context even without explicit pulse-level IPI descriptors, although indirect correlations with coarser click-

structure cues cannot be excluded. Limitations include the small multi-speaker test subset, the restriction to three individuals, and the absence of cross-session or cross-year generalization tests.

7 Conclusion

We introduced LeAudioJEPA, an audio adaptation of LeJEPA that removes AudioJEPA's EMA teacher-student mechanism and uses SIGReg-regularized latent prediction. On 30 s log-Mel-spectrogram segments of sperm whale codas, LeAudioJEPA obtains the best mean macro-F1 on the full, mono-speaker, and multi-speaker test sets, and significantly outperforms all trained baselines when evaluated over the 20 paired macro-F1 scores obtained from the random seeds. These results suggest that coda sequences contain speaker-discriminative temporal context even without explicit IPI descriptors. Future work will scale evaluation to more individuals, test cross-session and inter-annual generalization, and analyze whether the learned representations encode timing, spectral structure, or interaction context.

8 Acknowledgments

Data were collected by R. Heuzey, A. Preud'homme, N. Boodhonee, Label Bleu production and ODV production, Longitude 181, under permits No.57/2017 from Mauritian authorities, in accordance with local regulations for MMO. Thank to the Mauritian Prime Minister's Off., Dr. Rezah Badal, Chief Scientific off Mr. Satish Kadhun (AFRC), Mr. Sachin Jootun, Miss Eliana Timol (MFDC), Dir. Miss Khoudijah Boodoo (Tourism Aut.). We thank Longitude 181, Un Océan de Vie, Label Bleu, and all field contributors. A.C.'s PhD is co-funded by Agence Innovation Defense AID, ULP-COCHLEA ANR-21-CE04-0020-01 and Biodiversa2021-488 EUROPAM. H.G. thanks also ANR-20-CHIA-0014.

References

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

[2] Ferrari, Trinh, F. Sarano, Giraudet, et al. Age and IPI Relation from Newborn to Adult Sperm Whale off Mauritius. *Scientific Reports*, 2024.

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

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

[5] Giraudet et al. Individual coda repertoires and signatures in a sperm whale social unit. *Sub. to JASA*, 2026.

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

[7] Tuncay et al. Audio-JEPA: Joint-Embedding Predictive Archit. for Audio Rep. Learning. In *ICME*, 2025.

[8] Berkenbaum, 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.

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

[10] 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.

[11] F. Sarano et al. Nursing behavior in Sperm Whales. *Animal Behavior & Cog.*, 10:105–31, 2023.

[12] LeCun. A path towards autonomous machine intelligence. Technical report, OpenReview, 2022.

[13] F. Sarano, V. Sarano, Giraudet, and Glotin. The Cachalot Voice https://cian.lis-lab.fr/pub/CachalotVoice2023.pdf, 2023.

[14] Baevski et al. wav2vec 2.0 a Framework for Self Supervised Learning of Speech Representations. In *NeurIPS*, 2020.

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

[16] Gong et al. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021.

[17] Niizumi et al. Masked spectrogram modeling using masked autoencoders for audio representation learning. In *HEAR Workshop*, 2022.

[18] Chareyre et al. Self-Supervised vs Supervised Representation Learning for Fin Whale Vocalization Detection. In *39th Conference on NeurIPS workshop: AI for animal communication*, 2025.

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

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

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

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