Dolph2Vec: Self-Supervised Representations of Dolphin Vocalizations

Chiara Semenzin[1] **Faadil Mustun**[1] **Roberto Dessì**[2] **Pierre Orhan**[3] **Yair Lakretz**[1]
Gonzalo de Polavieja[4] **Germán Sumbre**[1]

[1]École Normale Supérieure, Paris, France [2]Not Diamond, San Francisco, USA [3]Institut du Cerveau, Paris, France [4]Champalimaud Foundation, Lisbon, Portugal

Abstract

Self-supervised learning (SSL) has opened new opportunities in bioacoustics by enabling scalable modeling of animal vocalizations without the need for expensive manual annotation. However, current SSL models in this domain prioritize broad generalization across species and are not optimized for uncovering the fine-grained structure of individual communication systems. In this work, we collect and release a novel dataset of over five years of longitudinal recordings, from five known dolphins in a semi-naturalistic marine environment—an unprecedented resource for studying dolphin communication. We adapt the Wav2Vec2.0 [1] architecture to this domain and introduce *Dolph2Vec*, the first large-scale, species-specific SSL model trained exclusively on this data. We benchmark our model on two biologically relevant tasks: signature whistle classification and whistle detection. *Dolph2Vec* significantly outperforms general-purpose baselines in both tasks. Beyond performance, we show that learned embeddings and codebook structure capture interpretable acoustic units aligned with dolphin whistle categories and possibly sub-whistle structure, enabling fine-grained analysis of communication patterns. Our findings demonstrate how SSL can serve as both a model and a scientific tool to explore hypotheses in animal communication research. Our code is available at: https://github.com/chiarasemenzin/Dolph2Vec/

1 Introduction

Animal communication provides critical insights into cognition, social organization, and survival [2, 3]. Among vocal species, dolphins show an intriguing vocalization repertoire, dominated by whistles [4]. These include *signature whistles* (SWs), individually distinctive calls akin to "names" [5], and *non-signature whistles* (NSWs), whose communicative role is not well understood. Whistles are learned, mimicked, and exchanged in social interactions, yet their functional structure remains elusive [6, 7].

Recent advances in deep learning [8] have transformed bioacoustics, enabling large-scale analyses of raw vocalizations [9–12]. In particular, self-supervised learning (SSL) eliminates the need for costly annotations, especially in a field such as animal communication where ground truth is unknown. [13–15]. However, existing SSL models are trained on heterogeneous, multi-species data. While such generalist models enable broad transfer, they dilute the fine-grained, species-specific structure needed to understand a communication systems in depth [16–19].

Our work. We address this gap by introducing the first large-scale, species-specific dataset of dolphin vocalizations: ∼180,000 whistles recorded over five years from a stable pod in a semi-naturalistic marine enclosure. Building on this resource, we adapt Wav2Vec2.0 [1] to dolphin

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

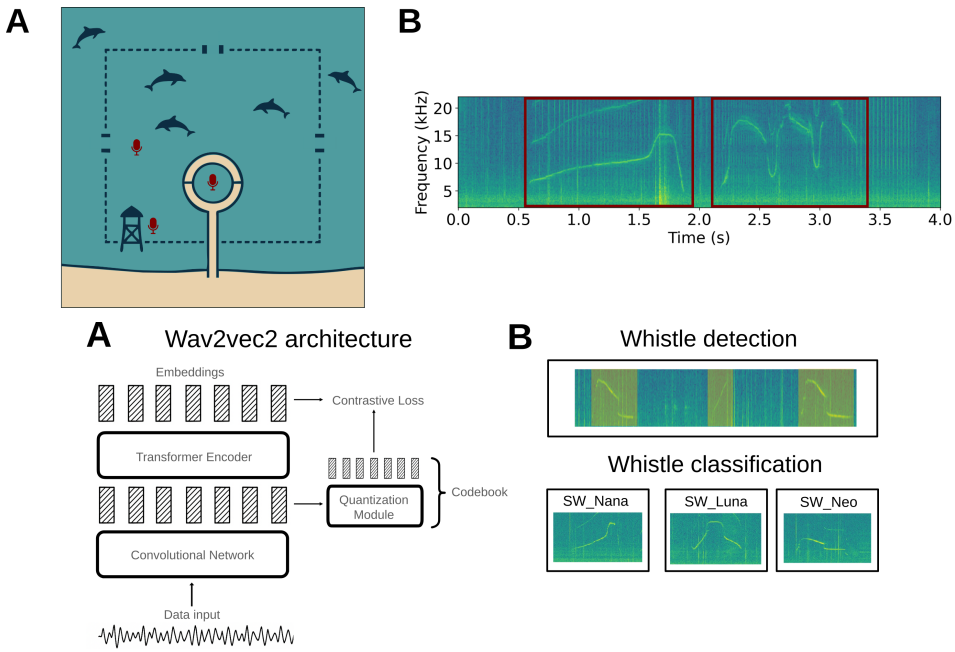

Figure 1: Top: Semi-naturalistic recording setup with hydrophones. Bottom: Dolph2Vec architecture and downstream tasks.

acoustics, introducing *Dolph2Vec*, the first SSL model trained exclusively on dolphin vocalizations. We evaluate Dolph2Vec on two biologically motivated tasks—whistle classification and whistle detection— and show that it outperforms both acoustic baselines and general-purpose SSL models.

Beyond downstream accuracy, we demonstrate that Dolph2Vec embeddings disentangle whistle categories and that its learned codebook units exhibit partial specialization for signature whistles. This suggests that *Dolph2Vec* captures recurring sub-whistle acoustic motifs, offering new opportunities to probe the compositional structure of dolphin communication. Taken together, *Dolph2Vec* serves both as a state-of-the-art representation learning model and a scientific tool for hypothesis generation in animal communication.

Contributions. (1) A new large-scale dataset of dolphin whistles (around 180k whistles). (2) *Dolph2Vec*, a Wav2Vec2-based SSL model adapted to dolphin acoustics. (3) State-of-the-art results on biologically relevant benchmarks. (4) Evidence that embeddings and codebook units are able to capture interpretable, biologically relevant categories.

2 Dataset and Method

Dataset. We collected a new large-scale dataset of bottlenose dolphin whistles in a semi-naturalistic marine enclosure over five years (Fig. 1). Recordings cover four identified dolphins plus one transient female visitor, yielding ∼180,000 whistles—up to three orders of magnitude larger than previously available datasets (e.g., [19]). Data were obtained with three hydrophones positioned around the reef, enabling longitudinal tracking of the same individuals. A subset of ∼8,000 whistles was annotated into signature and non-signature categories using ARTwarp [20] alignment and expert validation, supporting both supervised evaluation and the study of whistles over time. To our knowledge, this is the first publicly released dolphin dataset combining ecological realism, long-term continuity, and whistle-level annotations, thus filling a critical gap in computational bioacoustics.

Model. We adapt Wav2Vec2.0 to dolphin acoustics, introducing *Dolph2Vec* [1]. Raw audio is passed through the network, which consists of a convolutional feature encoder, a quantization module, and a Transformer-based context network. The encoder processes raw audio into latent representations, which are discretized by the quantization module into learned codewords drawn from a codebook. These discrete units serve as targets in a contrastive SSL task, where a context network captures

temporal dependencies to learn high-level speech features without labels. A diversity loss promotes balanced codebook usage. Unlike human speech (16 kHz), dolphin whistles reach higher frequencies; we therefore adjust kernel sizes and strides to match 44.1 kHz recordings.

Downstream tasks. We evaluate on two benchmarks (Fig. 1, right): (1) *Whistle detection*, predicting presence/absence of whistles as well as whistle category as described below. (2) *Whistle classification*, assigning isolated whistles to 10 previously described categories (5 signature, 5 non-signature). For both, we use simple linear classifiers on frozen embeddings, reporting mean accuracy (classification) and mAP (detection).

Baselines. We compare against hand-crafted acoustic features (MFCCs, spectral features, spectrogram means) and state-of-the-art SSL models: AVES [21] and BioLingual [17]. These serve as strong references for general-purpose bioacoustic representation learning.

3 Results

Performance of *Dolph2vec*. We evaluate *Dolph2Vec* on whistle classification and detection. Table 1 shows that it consistently outperforms acoustic baselines (MFCCs, spectrograms) and general-purpose SSL models (AVES, BioLingual). *Dolph2Vec* reaches 82% accuracy on classification, improving by $+5$ points over the best baseline, and matches state-of-the-art detection at 67.8 mAP.

Feature / Model	Classification Acc.	Detection mAP
MFCCs	47.2	53.3
Mean Spectrogram	61.6	65.5
AVES-bio	76.3	63.9
BioLingual	74.5	67.6
Dolph2Vec (ours)	**82.0**	**67.8**

Table 1: Performance on whistle classification and detection tasks.

Interpretability. UMAP projections of embeddings reveal clear clusters corresponding to signature whistle categories, indicating stronger disentanglement compared to the baselines (Fig. 2). Analysis of codebook activations shows partial specialization for individual whistles, suggesting that *Dolph2Vec* learns recurring acoustic sub-units beyond whole-whistle categories, consistent with biological hypotheses of sub-whistle structure [22].

Codebook specialization. Beyond performance, *Dolph2Vec*'s learned codewords exhibit partial specialization for individual signature whistles. By computing the conditional probability $P(\text{SW} \mid q_i)$ of whistle categories given quantized codebook indices, we observe that many units become associated with specific whistle types, whereas a randomly initialized model shows no such structure (Fig. 3). This indicates that the model captures recurring acoustic motifs aligned with biologically meaningful categories.

To quantify this effect, we measure conditional entropy $H(\text{SW} \mid q_i)$ and mutual information $I(q_i; \text{SW})$. Training reduces entropy ($2.13 \rightarrow 1.85$) and increases mutual information ($0.43 \rightarrow 0.70$), confirming that codebook representations encode label-relevant information (Table 2).

Model	Conditional Entropy	Mutual Information
Dolph2Vec (random init)	2.13	0.43 (17%)
Dolph2Vec (trained)	1.85	0.70 (28%)

Table 2: Information-theoretic metrics showing specialization of Dolph2Vec codebook units.

4 Discussion and Conclusion

Limitations. While *Dolph2Vec* achieves state-of-the-art results on dolphin-specific tasks, its specialization may reduce transfer to other species or ecological domains. The current version of the

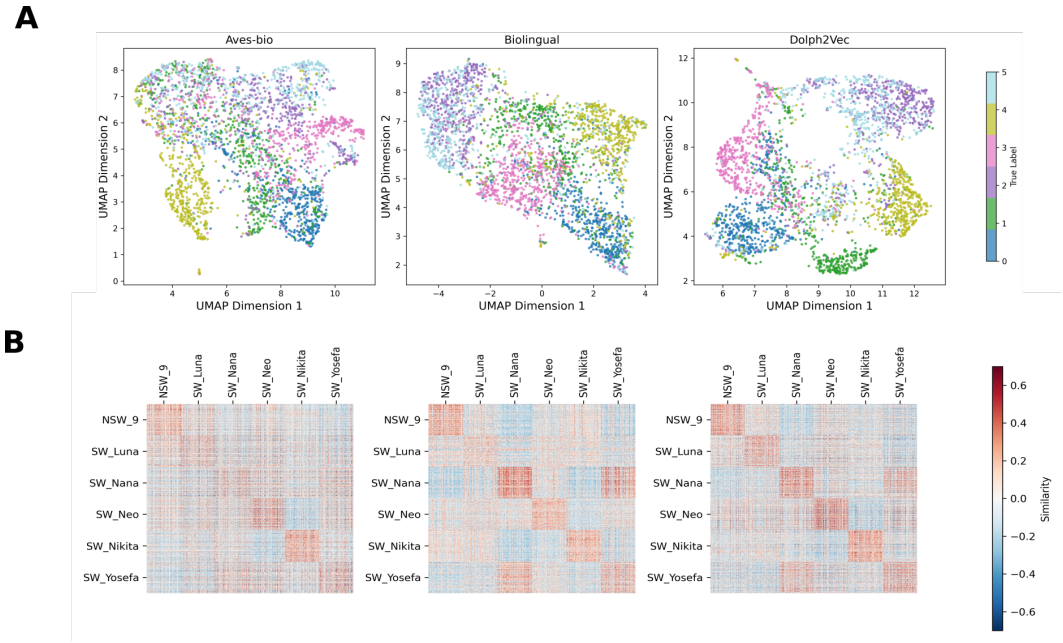

Figure 2: UMAP projection of embeddings showing clear separation by signature whistle category.

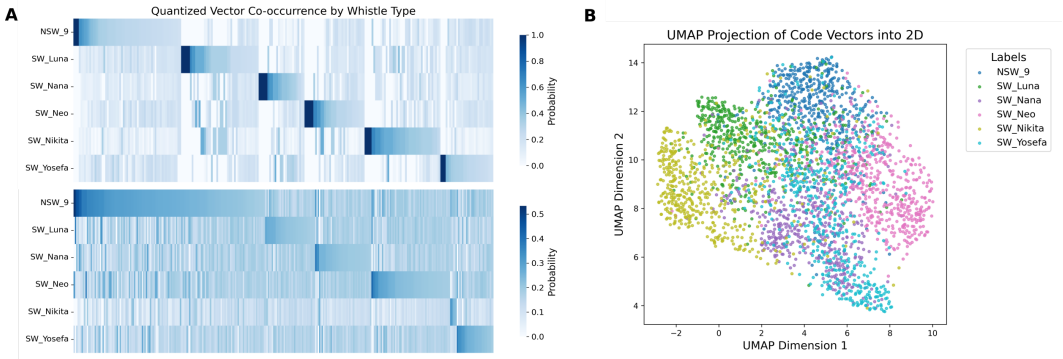

Figure 3: Codebook activations by signature whistle category. Trained Dolph2Vec units show partial specialization compared to random initialization.

model focuses solely on acoustic features, omitting behavioral and environmental context that are crucial for fully interpreting communicative function. In addition, while codebook units show partial specialization for signature whistles, not all units are clearly interpretable, and their relation to sub-whistle acoustic building blocks remains speculative. Future work should integrate multimodal data (e.g., movement, social context), test transfer across species, and systematically analyze how learned codebook units combine to form higher-order communication patterns.

Conclusion. We release a new large-scale dataset of ∼180k dolphin whistles recorded over five years, providing an unprecedented resource for studying communication in a semi-naturalistic setting. On this foundation, we introduce *Dolph2Vec*, the first species-specific self-supervised model of dolphin vocalizations. *Dolph2Vec* not only outperforms acoustic and general SSL baselines on whistle classification and detection, but also learns interpretable representations: embeddings that disentangle signature whistle categories and codebook units that capture recurring acoustic motifs. These findings suggest that self-supervised models can recover both coarse (whistle-level) and fine-grained (sub-whistle) structure, bridging performance with interpretability.

Overall, *Dolph2Vec* demonstrates how machine learning can serve as both a high-performance tool and a scientific probe in animal communication research. By releasing both the dataset and model, we

aim to foster interdisciplinary collaboration, encourage new benchmarks, and advance understanding of non-human communication systems.

Acknowledgments

We thank Emmanuel Chemla and Emanuele Rossi for their helpful feedback on this work.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 945304 — Cofund AI4theSciences hosted by PSL University (Chiara Semenzin) as well as a fourth-year fellowship granted by PSL-Neuro, funded by PSL University (Faadil Mustun). We also acknowledge support from the CNRS through the International Emerging Actions (IEA) and International Research Project (IRP) programs.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[2] Weronika Penar, Angelika Magiera, and Czesław Klocek. Applications of bioacoustics in animal ecology. *Ecological complexity*, 43:100847, 2020.

[3] Michael A Pardo, Kurt Fristrup, David S Lolchuragi, Joyce H Poole, Petter Granli, Cynthia Moss, Iain Douglas-Hamilton, and George Wittemyer. African elephants address one another with individually specific name-like calls. *Nature Ecology & Evolution*, 8(7):1353–1364, 2024.

[4] Peter L Tyack. Dolphins whistle a signature tune. *Science*, 289(5483):1310–1311, 2000.

[5] Vincent M Janik. Acoustic communication in delphinids. *Advances in the Study of Behavior*, 40:123–157, 2009.

[6] Diana Reiss and Brenda McCowan. Spontaneous vocal mimicry and production by bottlenose dolphins (tursiops truncatus): evidence for vocal learning. *Journal of Comparative Psychology*, 107(3):301, 1993.

[7] Vincent M Janik. Cetacean vocal learning and communication. *Current opinion in neurobiology*, 28:60–65, 2014.

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[9] Guy Oren, Aner Shapira, Reuven Lifshitz, Ehud Vinepinsky, Roni Cohen, Tomer Fried, Guy P Hadad, and David Omer. Vocal labeling of others by nonhuman primates. *Science*, 385(6712):996–1003, 2024.

[10] Peter C Bermant. Biocppnet: automatic bioacoustic source separation with deep neural networks. *Scientific Reports*, 11(1):23502, 2021.

[11] Clea Parcerisas, Elena Schall, Kees te Velde, Dick Botteldooren, Paul Devos, and Elisabeth Debusschere. Machine learning for efficient segregation and labeling of potential biological sounds in long-term underwater recordings. *Frontiers in Remote Sensing*, Volume 5 - 2024, 2024.

[12] Elena Schall, Idil Ilgaz Kaya, Elisabeth Debusschere, Paul Devos, and Clea Parcerisas. Deep learning in marine bioacoustics: a benchmark for baleen whale detection. *Remote Sensing in Ecology and Conservation*, 10(5):642–654, 2024.

[13] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR, 17–23 Jul 2022.

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[15] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, 2022.

[16] Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. Beans: The benchmark of animal sounds. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[17] David Robinson, Adelaide Robinson, and Lily Akrapongpisak. Transferable models for bioacoustics with human language supervision. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1316–1320, 2024.

[18] David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. NatureLM-audio: an audio-language foundation model for bioacoustics. In *The Thirteenth International Conference on Learning Representations*, 2025.

[19] Laela Sayigh, Mary Ann Daher, Julie Allen, Helen Gordon, Katherine Joyce, Claire Stuhlmann, and Peter Tyack. The watkins marine mammal sound database: an online, freely accessible resource. In *Proceedings of Meetings on Acoustics*, volume 27. AIP Publishing, 2016.

[20] Volker B Deecke and Vincent M Janik. Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. *The Journal of the Acoustical Society of America*, 119(1):645–653, 2006.

[21] Masato Hagiwara. Aves: Animal vocalization encoder based on self-supervision, 2022.

[22] Faadil Mustun, Chiara Semenzin, Dean Rance, Emiliano Marachlian, Zohria-Lys Guillerm, Agathe Mancini, Inès Bouaziz, Elisabeth Fleck, Nadav Shashar, Gonzalo G de Polavieja, et al. Whistle variability and social acoustic interactions in bottlenose dolphins. *bioRxiv*, pages 2024–10, 2024.