Dolph2Vec: Self-Supervised Representations of Dolphin Vocalizations

Anonymous authors
Paper under double-blind review

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.

1 Introduction

Bioacoustics—the study of sound production, perception, and function in animals—is foundational for understanding animal behavior, ecology, and conservation (2; 3). A key application is the study of animal communication, which reveals social structures, cognitive abilities, and survival mechanisms (4; 5).

Among vocal species, dolphins are especially intriguing due to their sophisticated communication system. Their most studied vocalizations are tonal whistles (6), which include signature whistles (SWs)—individually distinctive sounds functioning as acoustic labels akin to names (7)—and non-signature whistles (NSWs), whose function remains unknown. Whistles are learned, mimicked (8; 9), and exchanged in sequences that maintain social bonds and coordination (10; 11). Despite these advances, our understanding of the function of dolphin whistles remains limited.

In recent years, deep learning (12) has become a pivotal tool in bioacoustics (13–16) by enabling scalable analysis of large audio datasets. Self-supervised learning (SSL) is particularly powerful for extracting structure from raw, unlabeled recordings (17–19). By designing proxy tasks that derive supervisory signals from the data itself, SSL removes the need for costly manual annotations—a major advantage in animal communication studies, where labeling is especially ambiguous due to the lack of ground truth.

Despite growing interest in SSL for animal vocalizations, most bioacoustic models remain general-purpose. Typically trained on large, heterogeneous datasets spanning a handful of vocalizations from many species—including diverse background and non-animal sounds (20–23)—they achieve broad generalization for tasks such as species detection and classification but dilute the species-specific structure needed to understand communication systems.

To address this limitation, we introduce the first large-scale, species-specific dataset of dolphin vocalizations: about 180,000 whistles collected over five years from a stable pod in a semi-naturalistic setting—up to three orders of magnitude larger than prior datasets. This longitudinal resource captures vocal behavior over time, enabling analysis of individual identity, social dynamics, and potential drift due to age or group reorganization.

Building on this dataset, we release *Dolph2Vec*, the first self-supervised model pre-trained exclusively on dolphin vocalizations. We adapt the Wav2Vec2.0 architecture (1) to dolphin whistle acoustics and benchmark it on a novel downstream task reflecting biologically relevant questions in dolphin communication. Unlike generalist models, our Wav2Vec-based architecture also enables hypothesis testing via learned codebooks, providing interpretable units grounded in the structure of the vocal signal.

More broadly, this work highlights the reciprocal value of combining deep learning with animal communication research. Animal vocalization datasets offer a rich testbed for developing and stress-testing machine learning models on species with acoustically rich, high-frequency, and continuously varying signals. Conversely, machine learning—particularly self-supervised models—provides a transformative approach to studying non-human communication, uncovering latent structure directly from raw data. These models can serve both as analytical tools and as hypothesis-generating engines in animal acoustic research. By demonstrating the power of SSL to reveal structure in bioacoustic data, we aim to strengthen the growing intersection of machine learning and animal communication and inspire new approaches to investigating the evolution and mechanisms of animal communication.

2 Related work

Animal studies Dolphins produce three main sound types—echolocation clicks, burst pulses, and whistles—of which the latter two are central to communication (24; 25). A key element is the signature whistle (SW), an individually distinctive and stereotyped call used for identification and group cohesion (26). SWs, along with non-signature whistles (NSWs), constitute the majority of dolphin vocal output, with SWs accounting for up to 70% of whistles emitted in the wild (7). Recent work indicates SWs may include transient frequency modulations conveying information beyond identity (27), suggesting greater structural complexity than previously assumed, though their functional roles remain unclear.

Research on dolphin communication has largely relied on either behavioral studies of captive individuals in controlled environments (28; 29), or acoustic data from free-ranging dolphins (26; 30). The latter remains challenging to obtain, and, to the best of our knowledge, long-term, consistent datasets from the same individuals in the wild have not been released. Datasets typically provide short-term recordings of isolated instances of a mix of dolphin sounds (23; 31), or lack ecological realism due to captivity constraints. In addition, these datasets are often not publicly available (32). In contrast, our dataset consists of longitudinal recordings from a dolphin population living in a large, naturalistic marine environment we refer to as *semi-captive* (i.e. enclosed from boats but with openings to the sea). To our knowledge, this is the first publicly available dolphin dataset combining semi-captivity, longitudinal data, and whistle-level annotations, providing a new resource for studying both individual-specific and social aspects of dolphin acoustic communication.

Deep learning for animal studies The growing availability of acoustic data (33) has enabled deep learning across bioacoustics (14–16), with spectrogram-based convolutional networks (34) widely used for detection, classification, and clustering. For dolphins, supervised whistle-classification approaches have been proposed (35; 36), but these rely entirely on labeled data and cannot uncover structure in an unsupervised fashion. While some studies have leveraged very large audio datasets to improve performance (37; 38), they still require vast amounts of annotated data, which is costly and labour-intensive to obtain.

Self-supervised learning (SSL) directly addresses this constraint by exploiting the abundance of unlabeled acoustic data. Instead of relying on human-provided labels, SSL models learn meaningful representations by defining proxy tasks that capture inherent audio patterns (see

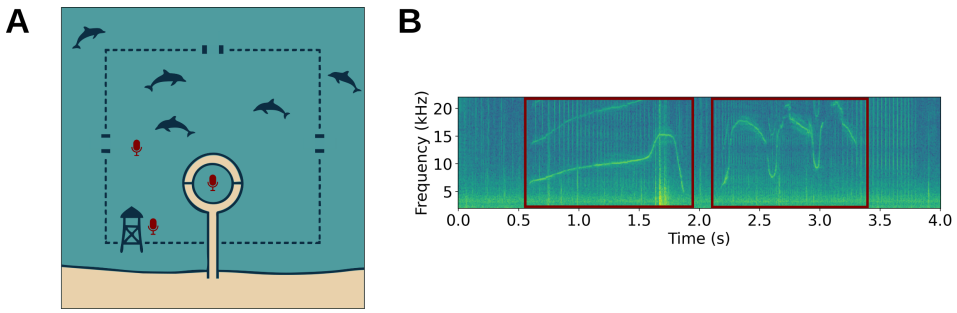

Figure 1: A) Schematic of the data collection setup, showing the dolphin area with three hydrophones and a rope with openings which allow dolphin passage to the open sea, while preventing boat access. B) A representative spectrogram of dolphin signature whistles showing distinct frequency patterns. The first example belongs to an individual named Nana, and the second one to Nikita.

Sec. A for details). This allows researchers to bypass annotation constraints and exploit large collections of raw recordings.

Models such as AVES (39), Nature-LM (22), and BioLingual (21) show that SSL can achieve strong downstream performance in species detection and classification. Nature-LM trains a generative audio–language model, while BioLingual uses a dual-tower audio–text approach with over a million synthetic captions—a powerful but less scalable strategy for homogeneous single-species datasets like ours. Our work instead uses an encoder-only architecture, well-suited for extracting representations for downstream tasks. Most similar to our work is AVES, which trains a HuBERT (18) model with several pre-training data mixes. While this model performs well for multi-species classification tasks, our focus is on an animal-specific model that shows good performance while also enabling testing theories grounded in biological studies of animal communication, an aspect overlooked in AVES (40; 41).

Transferability across domains remains unresolved. SSL models pretrained on human speech support species identification and call-type classification (42; 43), yet species-specific embeddings outperform general audio for birdsong (44; 45). WhaleLM (46) shows that SSL can also capture biologically relevant features in whale communication, while Gubnitsky et al. (47) stress species-specificity with a click detector for sperm whale codas. Our work contributes to this debate by introducing individual dolphin signature whistle identification with a dolphin-specific SSL model. *Dolph2Vec* combines large-scale pretraining with interpretable analyses to yield more generalizable and biologically informative embeddings.

3 Experimental Setup

In this section, we present the setup used for our data collection pipeline, the unique properties of our dataset, the pre-training of *Dolph2Vec*, as well as the data used for downstream tasks.

3.1 Data Collection

We present a novel dataset of bottlenose dolphin vocalizations collected in a semi-captive yet ecologically valid setting. Recordings were made in a natural marine enclosure in the Red Sea, where a resident pod of *Tursiops truncatus ponticus* coexists with human caregivers and visitors. Dolphins on the reef are untrained and free to enter and leave the area without restriction. This distinctive setup enables natural vocal behavior to be recorded while supporting long-term tracking of the same individuals (27; 48). Fig. 1**A** provides a schematic overview of the data-collection setup; a photo is shown in Appendix B.

The dataset consists of longitudinal acoustic recordings from four previously identified dolphins (48). In 2019, a fifth individual (*Tursiops aduncus*), an extralimital female from the Indian Ocean, joined the pod sporadically. Her signature whistle was identified from temporal

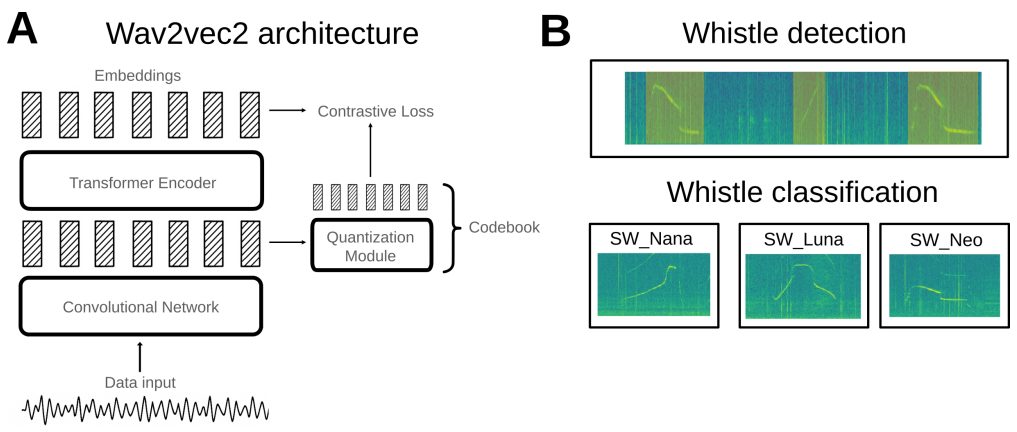

Figure 2: A) The Wav2Vec2.0 architecture used in *Dolph2Vec*. Raw audio is encoded into latent representations by a convolutional feature encoder, discretized via a quantization module with a learned codebook, and contextualized using a Transformer network. B) Downstream tasks: Top—whistle detection on spectrograms with highlighted whistles; Bottom—whistle classification of three distinct signature whistles from different individuals.

production patterns using the Signature Identification (SIGID) method (49). Additional information on the software and equipment used for data collection is provided in Appendix B.

Currently, the only large, publicly available dataset of dolphin sounds contains 566 dolphin whistles (23). Our dataset contains around 180,000 whistles, almost continuously recorded for 5 years, from the same and known pod of dolphins living in natural conditions. Furthermore, it includes whistle category labels for a subset of the data (around 8,000 whistles), enabling detailed analysis of vocal behavior over time. To support reproducibility, we will publicly release all data, providing a valuable resource for animal communication and computational bioacoustics research.

3.2 Pre-training Data

Our pre-training dataset contains 100 hours of audio resampled at 44.1 kHz, spanning all recordings from October 2019 to April 2024 (excluding 2022 due to technical issues which prevented stable recordings). This corresponds to roughly 180,000 individual whistles, estimated from the labeled subset using empirical whistle durations and inter-whistle intervals (both 1 s). Under this assumption, 100.73 hours of recordings yield about 50 hours of whistling. The total amount of whistles is roughly 300 times more than other existing dolphin datasets.

To select pretraining data, we use a custom convolutional neural network (CNN) based on the VGG16 architecture (50) pre-trained on ImageNet (51) to extract recordings containing vocalizations. The CNN takes spectrograms as input and was fine-tuned on over 8,500 whistles identified by a custom algorithm leveraging spectral features and dynamic time warping (DTW) to known whistle templates (27). The resulting dataset comprises 33,267 segments (max duration 246 s), truncated to 20 s during training.

3.3 *Dolph2Vec* - Architecture

Unlike written language, which provides discrete symbols as natural supervision, audio is a continuous signal without predefined units. *Dolph2Vec* follows the Wav2Vec2.0 architecture (1), illustrated in Fig. 2A. It 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.

3.4 *Dolph2Vec* - Training

We pre-trained a modified Wav2Vec2.0 model (1) on our dolphin vocalization dataset for 400K steps using 32 A100 GPUs with two steps of gradient accumulation and a batch size of 4 sound files per device (total $64 \times 4 = 256$). We used two codebooks with 320 vectors each and trained with the AdamW optimizer (52). Additional details on the pre-training setup and hyperparameters are in Sec. D.

Because our recordings are sampled at 44.1 kHz rather than the standard 16 kHz of speech, we modified the feature encoder to preserve the original model's temporal resolution. Specifically, we increased the first convolutional layer's kernel size from 10 to 30 and stride from 5 to 15, matching the receptive-field granularity of the original Wav2Vec2.0. All other architectural parameters follow the base configuration; see (1) for details.

3.5 Downstream tasks

We test the models on two downstream tasks of increasing granularity. First, we follow the detection task proposed in (39), and identify the presence or absence of specific whistle types in fixed-length audio segments. Secondly, we perform classification by assigning a label to isolated whistle sounds. Task representation is in Fig. 2**B**. For both tasks, we trained a logistic regression model with 5-fold stratified cross-validation using scikit-learn (53). We use the lbfgs solver and a maximum of 1500 iterations to ensure convergence. We use stratified splits to maintain the class distribution across folds. We test L2 regularization parameters with value 0.1, 1.0 and 10 and report the best score. Classification performance is measured by average accuracy across folds, and detection by mean average precision (mAP), following (20).

Detection Detection is the task of predicting whether a whistle is present in 0.5-second audio segments, by also classifying its whistle category. Each segment receives binary labels for all known whistle types; segments without any are labeled non-whistle. The dataset was built by segmenting recordings and assigning labels based on annotations from the classification task, supplemented with manually labeled data from (23). The final dataset consists of 660 segments containing at least one labeled whistle, along with additional segments containing no whistles to serve as background.

Classification For classification, we constructed a dataset of whistles labeled with the whistle type, containing 10 classes (5 signature whistles and 5 non-signature whistles). The classes were obtained by first classifying all whistles into categories using ARTwarp (27) (54), an unsupervised neural network algorithm which incorporates dynamic time warping (55). The automatically assigned labels were then manually corrected by expert annotators following visual inspection of spectrograms. Since the original dataset was highly imbalanced, with four classes having fewer than 300 samples each, we excluded these four categories. We then randomly sampled 500 instances from each of the remaining classes, creating a balanced dataset consisting of six classes. We use this dataset to train a linear regression model with stratified 5-fold cross validation.

3.6 Baseline Models

Acoustic Baselines As acoustic baselines, we evaluated traditional hand-crafted features including spectral features (spectral centroid, spectral bandwidth, spectral contrast, and spectral rolloff), Mel-frequency cepstral coefficients (MFCCs), and mean spectrogram representations.

BioLingual As a baseline, we include BioLingual, a contrastive language-audio model based on the CLAP-LAION architecture (56) which was trained on AnimalSpeak (21), a large-scale dataset comprising over one million captioned bioacoustic recordings from 25,000 species. Using audio-text alignment, BioLingual enables zero-shot retrieval and classification across taxa. We evaluate its performance on dolphin vocalizations using its pre-trained audio encoder without additional tuning.

Feature Type	Whistle Classification	Whistle Detection
Chance level	16.7	8.3
Spectral Features	34.2 ± 0.01	44.7 ± 4.44
MFCCs	47.2 ± 0.02	53.3 ± 3.72
Mean Spectrogram	61.6 ± 0.02	65.5 ± 3.74
AVES-core (39)	74.0 ± 0.01	64.5 ± 3.44
AVES-bio (39)	76.3 ± 0.01	63.9 ± 2.03
BioLingual (21)	74.5 ± 0.01	67.6 ± 4.33
Dolph2Vec (ours)	**82.0** ± 0.01	**67.8** ± 2.85

Table 1: Accuracy on the whistle classification dataset and mAP on the whistle detection one. Scores computed using stratified 5-fold cross validation.

AVES We also include AVES, a self-supervised transformer-based audio model adapted from HuBERT (57) and pre-trained on a large corpus of unannotated audio comprising animal vocalizations, human speech, and environmental sounds. AVES learns discrete latent targets through clustering and predicts masked waveform segments. We evaluate two variants: AVES-core, pre-trained on general audio datasets including FSD50K (58) and the balanced subset of AudioSet (37); and AVES-bio, pre-trained on a curated subset of AudioSet and VGGSound (59) containing only animal vocalizations. We use the AVES encoder without further fine-tuning.

4 Results

4.1 *Dolph2Vec* - The First Large-Scale Species-Specific Self-Supervised Model

During training loss decreases steadily, with both contrastive and diversity losses contributing to this trend, as shown in Fig. 6. The declining contrastive loss indicates improved discrimination of latent representations, while the diversity loss ensures utilization of the full representational space. This confirms effective convergence of the pre-training process.

4.2 *Dolph2Vec* Matches State-of-the-Art Performance on Whistle Detection

Following standard practice in the field of self-supervised learning (60–62), after training *Dolph2Vec*, we froze its weights and use the model to extract embeddings for downstream tasks. Performance on these tasks acts as a measure of quality of model representations. Audio was resampled to 44.1 kHz models. Although the AVES models were originally pre-trained on 16 kHz inputs, we found that 44.1 kHz yielded better results on our data and was more appropriate given the frequency characteristics of our dolphin whistles dataset. For BioLingual, we retained the original 48 kHz sampling rate used during its pre-training to ensure compatibility and optimal performance.

Table 1 reports performance on the whistle classification and detection tasks across all feature types and embedding models. BioLingual and *Dolph2Vec* achieve the highest detection scores, with BioLingual at 67.6 mAP and *Dolph2Vec* slightly higher at 67.8 mAP, indicating that *Dolph2Vec* matches state-of-the-art performance on the whistle detection task.

4.3 *Dolph2Vec* Achieves New State-of-the-Art in Whistle Classification

Traditional acoustic features, used as baselines, achieved limited classification accuracy (Table 1, with spectral features reaching only 34.2% and MFCCs slightly higher at 47.2%. Mean spectrograms performed best among the baselines, achieving 61.6% accuracy.

Embedding-based models yielded significantly stronger results. AVES-core (39), AVES-bio (39), and BioLingual (21) all surpassed 70% classification accuracy. Our model, *Dolph2Vec*,

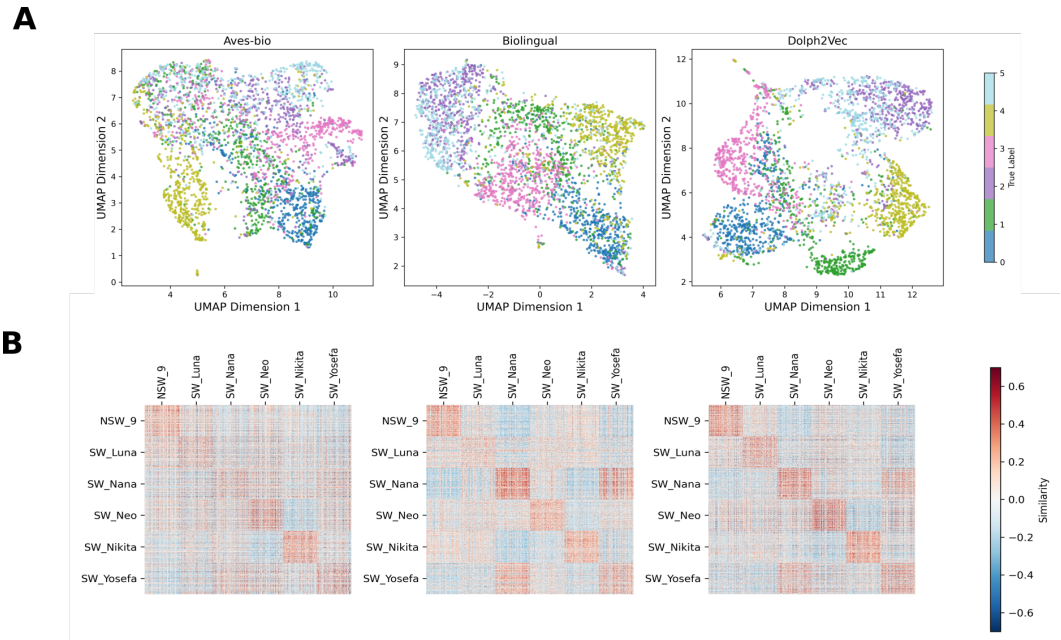

Figure 3: A) UMAP projection of learned embeddings from AVES-bio, BioLingual and *Dolph2Vec*, colored by their true label (Signature Whistle Category) B) RSA matrices of the three models.

achieved the highest overall performance, reaching 82% accuracy, demonstrating the strongest representation quality across both tasks. Notably, SW_Yosefa and SW_Nana are consistently confused by all models, reflecting their actual acoustic similarity and highlighting the biological realism of the benchmark. BioLingual was trained on the largest amount of data between all models while using an audio-text contrastive objective. While this showed remarkable performance on multi-species classification tasks (21), its inferior score on our task suggests that the model might not be able to pick up narrow features that are necessary when transferring to single-species benchmarks. This suggests a trade-off between broad and narrow performance when modeling bioacoustic data. Investigating this trade-off in relation to pre-training data is a valuable avenue for future work.

4.4 Strong Disentanglement of Signature Whistle Representations in *Dolph2Vec* Embeddings

To examine how well our model disentangles signature whistles from different individuals, we qualitatively and quantitatively analyze embeddings from AVES-bio, BioLingual, and *Dolph2Vec* using dimensionality-reduction techniques. We cluster UMAP projections of each model's representations with Gaussian Mixture Models (GMMs), which provide soft assignments and accommodate non-spherical cluster shapes, making them well suited to high-dimensional embeddings (Fig. 3).

Dolph2Vec embeddings show the clearest visual separation of the six ground-truth whistle categories (Fig. 3**A**). We further evaluate clustering performance with Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI): *Dolph2Vec* achieves the highest scores (ARI = 0.3565, NMI = 0.4226), outperforming BioLingual (ARI = 0.2963, NMI = 0.3480) and AVES-bio (ARI = 0.1984, NMI = 0.2488). These results confirm that domain-specific self-supervised pretraining yields more structured and separable dolphin vocalization representations than general-purpose models.

To further characterize representational structure, we computed Representational Similarity Analysis (RSA) matrices between *Dolph2Vec* and the two baselines (63). RSA correlates pairwise similarity scores across models, capturing how their representational structures align.

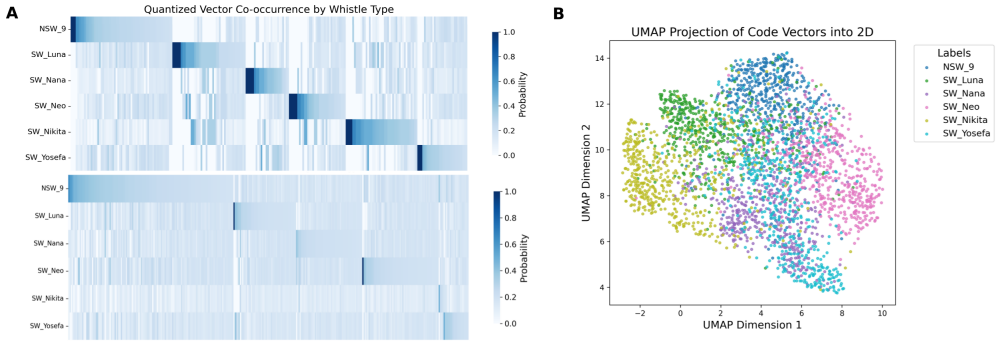

Figure 4: A) Codebook activations by signature whistle category in *Dolph2Vec* trained (top) and *Dolph2Vec* randomly initialized (bottom). B) 2D Projection of mean codevectors by SW category.

We computed RSA on unseen test data used for classification, reporting similarity matrices in Fig. 3**B**. *Dolph2Vec* shows stronger within-category similarity and clearer diagonal blocks (red) than baselines, indicating higher within-category consistency and better between-category differentiation. Cross-model Spearman correlations revealed that *Dolph2Vec* representations differed meaningfully from AVES-bio ($r_s = 0.35$, $p < 10^{-5}$) and BioLingual ($r_s = 0.31$, $p < 10^{-4}$), suggesting each model captures distinct statistical regularities.

4.5 Dolph2Vec Codebook Units Exhibit Partial Specialization for Signature Whistles

To test whether the discrete latent representations capture signature whistle (SW) information, we use *Dolph2Vec* to compute quantized latents q_t and codebook indices q_i across the full evaluation set, then calculate co-occurrence between annotated SW labels and codebook indices. Fig. 4**A** shows the conditional probability $P(\text{SW} \mid q_i)$: many discrete latents in *Dolph2Vec* (top) specialize for specific whistle types, unlike the randomly initialized model (bottom), which nonetheless retains some structure—consistent with prior findings on generalization in random networks (64). This may explain why specialization mainly occurs in the first codebook, while the second stays closer to its random state; all analyses in Fig. 4**A** therefore use one code set.

We then compute conditional entropy $H(\text{SW} \mid q_i)$ and mutual information $I(q_i; \text{SW})$, comparing against an untrained *Dolph2Vec* (Table 2) to quantify how strongly the learned latent space reflects class-specific encoding. Training *Dolph2Vec* markedly reduces conditional entropy and increases mutual information, indicating more informative and structured codebook representations.

The partial specialization of codebook indices—some highly specific, others shared across SW types—suggests they capture acoustic structure at a sub-whistle level. Fig. 4**B** supports this, showing that SW-type–averaged codevectors do not form distinct clusters, implying the learned codebook represents recurring acoustic features rather than whole whistle types. This aligns with hypotheses that meaningful information may occur at the sub-whistle level (27; 65; 8). We propose these learned units can act as fine-grained building blocks to investigate order effects and generate new hypotheses about dolphin acoustic communication

Model	*Conditional Entropy*	*Mutual Information*
Dolph2Vec-Random	2.13	0.43 (17%)
Dolph2Vec	1.85	0.70 (28%)

Table 2: Information-theoretic metrics comparing untrained and trained codebooks.

4.6 Perturbation of temporal features

To test how temporal structure contributes to individual-identity classification, we compared performance on original versus temporally shuffled vocalizations. Shuffling the feature-encoder output along the time axis preserves local acoustic content but disrupts global call sequence. Accuracy dropped from 82.0% on unshuffled input to 75.1% on shuffled input, indicating a modest but significant reliance on temporal structure. This suggests identity-relevant information is mainly encoded in short-timescale acoustic features, consistent with findings in human speech where models such as Audio-MAE (66) and WavLM (67) achieve above-chance performance even on temporally ablated inputs (40). The small effect further implies temporal structure is not pivotal for categorizing signature whistles. As future work, we propose systematically perturbing the frequency dimension (e.g., pitch shifting or spectral warping) to test its contribution, clarifying spectral versus temporal encoding strategies and informing hypotheses on key acoustic features in dolphin communication.

5 Limitations

While *Dolph2Vec* surpasses general-purpose models on a dolphin-specific task, its specialization compromises performance on multi-species or cross-ecological applications. Optimal performance on downstream tasks in a broad range of bioacoustic domains may be achieved by fine-tuning general models on large-scale, species-specific datasets, combining cross-species representational breadth with domain-specific granularity. The model focuses exclusively on acoustic features, omitting behavioral and environmental context critical for interpreting communicative function. Future integration of multimodal data—such as the individuals' movements, social dynamics, or environmental cues—will be necessary to ground acoustic signals in biologically meaningful events.

6 Conclusion

This work introduces the first large-scale dataset of dolphin vocalizations—over five years of longitudinal recordings from a pod of five dolphins in a naturalistic marine environment. With roughly 180,000 estimated whistles, it enables communication-focused research at a scale and resolution previously unavailable, bridging the gap between ecological realism and machine learning scalability.

We show that *Dolph2Vec*, a domain-adapted self-supervised model trained on this dataset, achieves state-of-the-art performance on new whistle classification and detection tasks. A large-scale, species-specific model can thus deliver both high performance and scientific insight. Analysis of *Dolph2Vec*'s internal structure reveals interpretable patterns in dolphin vocal behavior, including possible sub-whistle acoustic units—offering new ways to test hypotheses in animal communication.

Future work should explore domain-specific pretraining enhancements such as augmentations tailored to dolphin vocal features, adjusting the convolutional extractor and codebook to better match species-specific acoustics, and studying how human or background sounds in pretraining data affect performance. On the interpretability front, perturbing features such as frequency or duration could test classification robustness and clarify spectral vs. temporal encoding strategies. Another promising avenue is examining whether learned codebook units act as discrete building blocks in dolphin vocal sequences, shedding light on compositionality in dolphin communication.

Beyond technical advances, our findings highlight the mutual benefits of combining animal studies and deep learning. By releasing both our dataset and pre-trained model, we aim to catalyze cross-disciplinary research and promote integrative approaches to non-human communication, inspiring broader efforts to build species-specific resources and interpretable computational tools.

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] Julia Fischer, Rahel Noser, and Kurt Hammerschmidt. Bioacoustic field research: a primer to acoustic analyses and playback experiments with primates. *American journal of primatology*, 75(7):643–663, 2013.

[3] Daniel T Blumstein, Daniel J Mennill, Patrick Clemins, Lewis Girod, Kung Yao, Gail Patricelli, Jill L Deppe, Alan H Krakauer, Christopher Clark, Kathryn A Cortopassi, et al. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology*, 48(3):758–767, 2011.

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

[5] 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.

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

[7] Laela S Sayigh, H Carter Esch, Randall S Wells, and Vincent M Janik. Facts about signature whistles of bottlenose dolphins, tursiops truncatus. *Animal Behaviour*, 74(6):1631–1642, 2007.

[8] 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.

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

[10] Peter Tyack. Population biology, social behavior and communication in whales and dolphins. *Trends in ecology & evolution*, 1(6):144–150, 1986.

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Richard C Connor and Rachel A Smolker. 'pop'goes the dolphin: A vocalization male bottlenose dolphins produce during consortships. *Behaviour*, 133(9-10):643–662, 1996.

[25] Vincent M Janik and Laela S Sayigh. Communication in bottlenose dolphins: 50 years of signature whistle research. *Journal of Comparative Physiology A*, 199:479–489, 2013.

[26] Vincent M Janik. Whistle matching in wild bottlenose dolphins (tursiops truncatus). *Science*, 289(5483):1355–1357, 2000.

[27] 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.

[28] Brenda McCowan and Diana Reiss. Quantitative Comparison of Whistle Repertoires from Captive Adult Bottlenose Dolphins (Delphinidae, *Tursiops truncatus)*: a Reevaluation of the Signature Whistle Hypothesis. *Ethology*, 100(3):194–209, jan 1995.

[29] Jennifer L Miksis, Peter L Tyack, and John R Buck. Captive dolphins, *Tursiops truncatus*, develop signature whistles that match acoustic features of human-made model sounds. *The Journal of the Acoustical Society of America*, 112(2):728–739, 2002.

[30] Denise L Herzing. Vocalizations and associated underwater behavior of free-ranging atlantic spotted dolphins, stenella frontalis and bottlenose dolphins, tursiops truncatus. *Aquatic Mammals*, 22:61–80, 1996.

[31] Francesco Di Nardo, Rocco De Marco, Alessandro Lucchetti, and David Scaradozzi. A wav file dataset of bottlenose dolphin whistles, clicks, and pulse sounds during trawling interactions. *Scientific Data*, 10:650, 2023.

[32] Laela S Sayigh, Vincent M Janik, Frants H Jensen, Michael D Scott, Peter L Tyack, and Randall S Wells. The sarasota dolphin whistle database: A unique long-term resource for understanding dolphin communication. *Frontiers in Marine Science*, 9:923046, 2022.

[33] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank Van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

[34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[35] Xunbin Deng, Mingzhang Zhou, Haixin Sun, and Yu Jiang. A novel dolphin whistle classifier based on attention-densenet. In *2023 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pages 1–6, 2023.

[36] Frants Havmand Jensen, Piper Wolters, Louisa van Zeeland, Evan Morrison, Gracie Ermi, Scott Smith, Peter L. Tyack, Randall S. Wells, Sam McKennoch, Vincent M. Janik, and Laela S. Sayigh. *Automatic Deep-Learning-Based Classification of Bottlenose Dolphin Signature Whistles*, pages 2059–2070. Springer International Publishing, Cham, 2024.

[37] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[38] Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

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

[40] Jules Cauzinille, Benoît Favre, Ricard Marxer, Dena Clink, Abdul Hamid Ahmad, and Arnaud Rey. Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures. In *Interspeech 2024*, pages 132–136. ISCA; ISCA, 2024.

[41] Gašper Beguš, Andrej Leban, and Shane Gero. Approaching an unknown communication system by latent space exploration and causal inference, 2024.

[42] Peter C. Bermant, Leandra Brickson, and Alexander J. Titus. Bioacoustic event detection with self-supervised contrastive learning. *bioRxiv*, 2022.

[43] Eklavya Sarkar and Mathew Magimai. Doss. Comparing self-supervised learning models pre-trained on human speech and animal vocalizations for bioacoustics processing, 2025.

[44] Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

[45] Burooj Ghani, Vincent J. Kalkman, Bob Planqué, Willem-Pier Vellinga, Lisa Gill, and Dan Stowell. Impact of transfer learning methods and dataset characteristics on generalization in birdsong classification. *Scientific Reports*, 15(16273), 2025.

[46] P. Sharma, S. Gero, R. Payne, et al. Contextual and combinatorial structure in sperm whale vocalisations. *Nature Communications*, 15(3617), 2024.

[47] G. Gubnitsky, Y. Mevorach, S. Gero, et al. Automatic detection and annotation of eastern caribbean sperm whale codas. *Scientific Reports*, 15(12790), 2025.

[48] Amir Perelberg, Frank Veit, Sylvia E van der Woude, Sophie Donio, and Nadav Shashar. Studying dolphin behavior in a semi-natural marine enclosure: Couldn't we do it all in the wild? *International Journal of Comparative Psychology*, 23(4), 2010.

[49] Vincent M Janik, Stephanie L King, Laela S Sayigh, and Randall S Wells. Identifying signature whistles from recordings of groups of unrestrained bottlenose dolphins (tursiops truncatus). *Marine Mammal Science*, 29(1):109–122, 2013.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015.

[51] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[54] 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.

[55] John R Buck and Peter L Tyack. A quantitative measure of similarity for tursiops truncatus signature whistles. *The Journal of the Acoustical Society of America*, 94(5):2497–2506, 1993.

[56] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[57] 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.

[58] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

[59] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[60] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[61] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

[62] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[63] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[64] Amir Rosenfeld and John K. Tsotsos. Intriguing properties of randomly weighted networks: Generalizing while learning next to nothing. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 9–16, 2019.

[65] Stephanie L King, Laela S Sayigh, Randall S Wells, Wendi Fellner, and Vincent M Janik. Vocal copying of individually distinctive signature whistles in bottlenose dolphins. *Proceedings of the Royal Society B: Biological Sciences*, 280(1757):20130053, 2013.

[66] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.

[67] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[68] Philippe Schlenker, Camille Coye, Shane Steinert-Threlkeld, Nathan Klinedinst, and Emmanuel Chemla. Beyond anthropocentrism in comparative cognition: Recentering animal linguistics. *Cognitive Science*, 46(12), 2022.

[69] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[70] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5528–5531, 2011.

[71] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[72] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[73] John Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32, 1957.

[74] Paulius Micikevicius, Sharan Narang, Jonah Alben, Greg Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

[75] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *Proceedings of ICLR Conference Track*, Toulon, France, 2017.

[76] Chris Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of ICLR Conference Track*, Toulon, France, 2017.

A Additonal related work: self-supervised learning

The traditional supervised learning approach for dolphin vocalization embeddings has long been criticized for enforcing a human-biased perspective (68). This bias stems from linking each vocalization directly to expert annotations or predefined features assumed by humans to be important. For instance, analyses of dolphin whistles have often focused either on the

identity of the putative emitter or on the assumption that the whistle envelope is the primary carrier of information. However, this reliance on human interpretation and predefined notions risks overlooking crucial communication cues within the signal or misidentifying their significance, potentially missing the true underlying structure and meaning of the vocalizations.

To remedy this problem, unsupervised approaches may provide a new perspective that avoids human biases. Recent progress in natural language processing has demonstrated that the meaning and structure of language could be re-discovered through an unsupervised, machine-learning-based approach. Self-supervised approaches such as (69–72) have first proposed embeddings of words as vectors. These approaches are based on the distributional hypothesis (73): a word is defined by the context of its use. These unsupervised, token-based approaches are not directly applicable to domains where the unit of computation is less clear, like speech processing. Instead, speech-processing models like Wav2Vec2.0 (1) or HuBERT (18) simultaneously extract the unit of computation (speech units) and perform the contextual processing. The Wav2Vec2.0 architecture is composed of two processing blocks (Fig 2A): First convolutional layers extract the speech units through local computations. Next, a transformer block performs contextual processing. Learning is achieved by a masking objective, where the model should unmask speech units, with unmasking evaluated through a contrastive objective.

B Data-collection setup

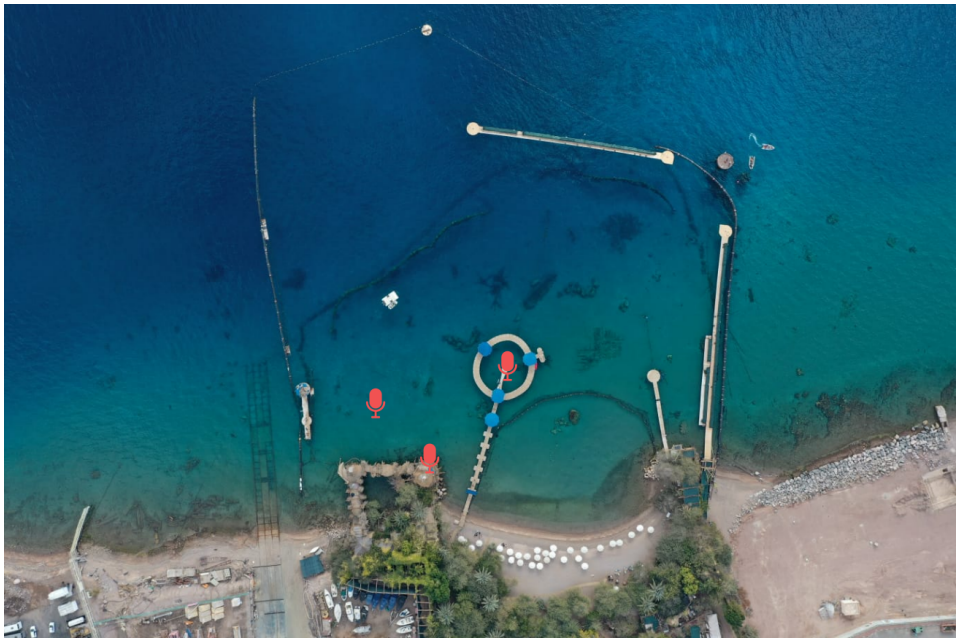

Figure 5: An aeral photo of our data collection setup.

The dataset was collected at Dolphin Reef in Eilat, Israel, a coastal site on the northern Gulf of Aqaba. This location serves as the natural habitat for a resident pod of bottlenose dolphins that freely move between the reef and open sea. Human-dolphin interactions occur only when initiated by the dolphins and are entirely voluntary. Hydrophones were placed in the locations as shown in Fig 5 Acoustic recordings were obtained using a set of 3 Brüel & Kjær® 8104 hydrophones connected to a 1704 preamplifier and a National Instrument® PCI-4474 acquisition card installed in a HP Z400 linux computer, sampling at 96 kHz, controlled by a custom-made code in C++. Recordings were conducted daily for one-hour periods at different times during the day (around 14 hours a day). The recordings were acquired between November 1, 2019, and March 12, 2024. Data acquisition was automated using scheduled crontab commands.

C Dataset properties

Studying a small, stable pod of five dolphins across several years provides advantages rarely available in animal communication research. The individuals' sex, family history, and kinship relations are well documented (27; 48), enabling integration of acoustic analysis with detailed social and historical context. The dataset thus combines individual-level identification with long-term recordings, supporting investigation of both fine-grained whistle structure and long-range social-linguistic patterns.

The pre-training dataset comprises 33,267 audio segments automatically extracted with a custom convolutional neural network. Segments average 12.91 seconds in length ($sd = 19.15$s), ranging from 2 to 246 seconds, and each contains at least one whistle. Whistles typically last about 1 second, with an average interval of 0.5 seconds between whistles, yielding an estimated total of roughly 180,000 individual whistles.

For the downstream classification task, a subset of about 8,000 whistles was annotated by domain experts through spectrogram inspection. These annotations distinguish signature whistles (SW) that serve as individual identifiers from non-signature whistles (NSW). The distribution of labeled data across categories is shown in Table 3.

Category	Count
SW_Luna	2,934
SW_Neo	2,239
SW_Nikita	888
NSW_9	658
SW_Yosefa	626
SW_Nana	521
SW_Dana	335
SW_Shy	81
NSW_3	45
NSW_6	27

Table 3: Distribution of annotated whistle categories.

To balance categories, all classes with at least 500 examples were subsampled to 500 instances each (SW_Luna, SW_Neo, SW_Nikita, NSW_9, SW_Yosefa, SW_Nana), ensuring an even distribution for downstream evaluation.

D Pre-training setup

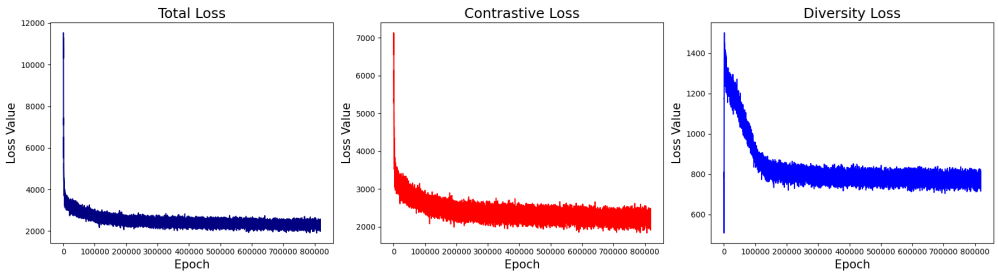

Figure 6: Pretraining losses over total training steps.

Training was conducted using the AdamW optimizer (52) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-6}$. The learning rate was set to 5×10^{-4} with a linear decay scheduler and 32,000 warmup steps. Weight decay was fixed at 0.01. Training proceeded for a maximum of 400,000 steps with a per-device batch size of 4 on 32 GPUs. Mixed precision training was used to reduce memory consumption (74). The Gumbel quantization module (75; 76) employed a

temperature schedule starting at 2.0 and decaying multiplicatively by a factor of 0.999995 at each step, with a floor of 0.5. Fig. 6 shows the total loss, contrastive loss and diversity loss steadily decreasing, confirming convergence during training.

E Additional baselines

Feature Type	**Whistle Classification**
Chance level	16.7
Dolph2Vec2 (random init)	37.9
Wav2Vec2-base (pre-trained)	47.0
Dolph2Vec2-shuffled	75.1

Table 4: Accuracy of additional baselines on our dolphin classification task.

We report performance on the signature whistle classification task for additional baselines in Table 4. Wav2Vec2 refers to the base model pretrained on human speech at 16 kHz (1). *Dolph2Vec*-random init denotes the *Dolph2Vec* model with randomly initialized weights, i.e., before any self-supervised pretraining. *Dolph2Vec*-shuffled is a variant of *Dolph2Vec* in which the temporal structure of the learned representations is disrupted by shuffling the output of the feature encoder along the time axis.

F Binary Whistle Detection

Feature Type	**Detection (mAP)**
AVES-bio	99.93 ± 0.13
AVES-core	99.92 ± 0.08
Dolph2Vec	99.81 ± 0.14
Mean Spectrogram	99.82 ± 0.10
BioLingual	99.37 ± 0.35
MFCCs	98.17 ± 0.54
Spectral Features	95.94 ± 2.69

Table 5: Detection performance (mAP) on binary whistle vs. non-whistle task. Scores reported as mean ± standard deviation across stratified 5-fold cross-validation.

Table 5 reports performance on a binary whistle detection task, where the objective is to distinguish between whistle and non-whistle audio segments. Unlike the main detection task described in Section 3.5, which involves identifying specific whistle types in a multi-label setting, this task simplifies the problem to a single binary classification per segment. All models achieve near-ceiling performance, with mAP scores above 95, indicating that distinguishing whistle sounds from background noise is relatively easy. AVES-bio and AVES-core achieve the highest scores (99.93 and 99.92, respectively), followed closely by *Dolph2Vec* (99.81) and spectrogram-based features (99.82). BioLingual and other baseline features also perform well but slightly below the top models. Due to this performance saturation, the binary detection task provides limited insight into model differences and is included here for completeness.

G Codebook size analysis

We evaluated the hypothesis that a smaller codebook might better capture the structure of dolphin whistles by representing them as combinations of a limited set of sub-whistle units. To test this, we pre-trained several Wav2Vec models with varying codebook sizes: 32, 128, and 320 codewords per codebook. Model performance was then assessed across multiple downstream tasks and unsupervised metrics (codebook entropy, clustering quality).

The results indicate that the configuration reported in the main text, two codebooks with 320 codewords each, achieves the best balance of performance and representation quality. It outperforms smaller codebooks in downstream classification and detection tasks, while also producing superior entropy and clustering results. This suggests that a larger codebook provides a more accurate and flexible representation of dolphin whistle structure.

H Second codebook

Figure 7 shows that the second codebook is visually similar between the trained and randomly initialized models. Table 1 confirms this, with nearly identical conditional entropy and mutual information values. While the distribution of categories varies, no specialized or class-specific activation patterns emerge, indicating limited functional differentiation.

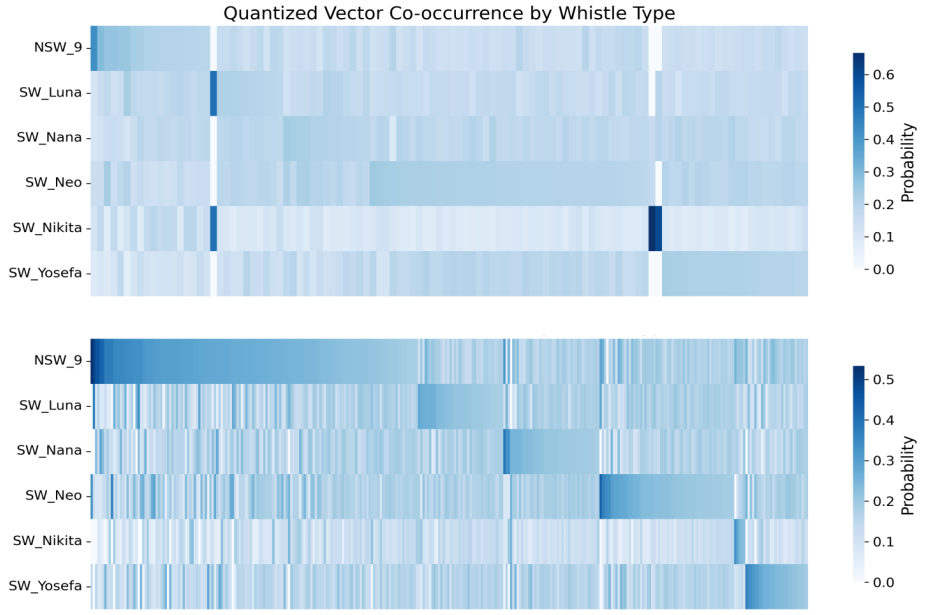

Figure 7: Second Codebook activations by signature whistle category in Dolph2Vec trained (top) and Dolph2Vec randomly initialized (bottom).

	Conditional Entropy	Mutual Information
Dolph2Vec	2.5027	0.0687
Dolph2Vec (random init)	2.4675	0.0907

Table 6: Information-theoretic metrics for the second codebook.