WhAM: Towards A Translative Model of Sperm Whale Vocalization

Orr Paradise[1,2] **Pranav Muralikrishnan**[*1] **Liangyuan Chen**[*1] **Hugo Flores García**[3]

Bryan Pardo[3] **Roee Diamant**[4,2] **David F. Gruber**[5,2] **Shane Gero**[6,2]

Shafi Goldwasser[1,2]

[1]UC Berkeley [2]Project CETI [3]Northwestern University [4]Haifa University
[5]City University of New York [6]Carleton University
[*]Equal contribution

Abstract

Sperm whales communicate in short sequences of clicks known as codas. We present WhAM (Whale Acoustics Model), the first transformer-based model capable of generating synthetic sperm whale codas from any audio prompt. WhAM is built by finetuning VampNet, a masked acoustic token model pretrained on musical audio, using 10k coda recordings collected over the past two decades. Through iterative masked token prediction, WhAM generates high-fidelity synthetic codas that preserve key acoustic features of the source recordings. We evaluate WhAM's synthetic codas using Fréchet Audio Distance and through perceptual studies with expert marine biologists. On downstream classification tasks including rhythm, social unit, and vowel classification, WhAM's learned representations achieve strong performance, despite being trained for generation rather than classification. Our code is available at `https://github.com/Project-CETI/wham`

1 Introduction

Understanding the communication of sperm whales (*Physeter macrocephalus*) is among the most fascinating questions in animal behavioral studies.

Sperm whales communicate using *codas*—short sequences of clicks that vary in number, rhythm, and tempo [Watkins and Schevill, 1977, Weilgart and Whitehead, 1993, Sharma et al., 2024a]. They live in stable, female-led social units that form larger vocal clans based on *dialect* [Rendell and Whitehead, 2003]. That is, the dialect of a social unit determines its clan, with social units associating exclusively with other units from their clan [Gero et al., 2016a]. Furthermore, dialects are believed to be learned socially rather than inherited genetically [Cantor and Whitehead, 2015, Rendell et al., 2012].

The complexity of these learned vocal patterns has motivated new computational approaches to understanding codas and their functionality. Leitao et al. [2024] modeled codas as (variable-length) Markov chains, revealing new patterns of inter-clan social learning. Beguš et al. [2025] study vowel-like spectral properties of codas, which were initially suggested by interpreting the codebook of a Generative Adversarial Network (GAN). Most recently, Sharma et al. [2024b] trained a transformer on click timings (inter-click intervals), which is able to predict codas in an exchange based on long-term dependencies, as well as future diving behavior. These studies collectively highlight how

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

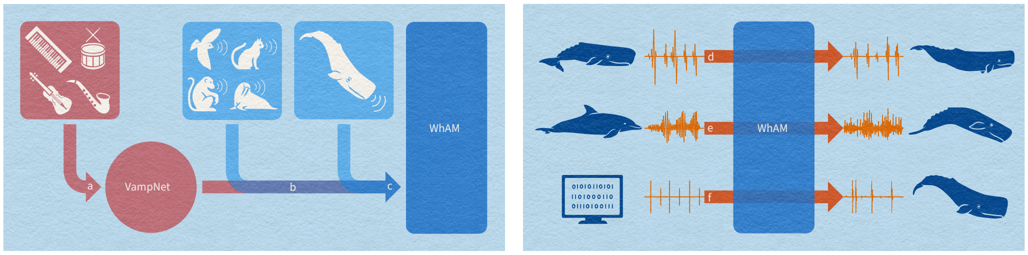

Figure 1: Left: WhAM is trained by finetuning VampNet [García et al., 2023], an audio-to-audio transformer pretrained on a large music dataset (a). Namely, we perform **domain adaptation** (b) on animal vocalizations followed by **species-specific finetuning** (c) on a novel sperm whale coda dataset. Right: WhAM synthesizes **context-aware variations** (d) of input codas and **acoustically translates** (e) natural and (f) artificial audio into coda-like audio. Illustration © Alex Boersma 2025.

machine learning—particularly transformer architectures—can decode patterns imperceptible to traditional methods.

Transformers [Vaswani et al., 2017] originated in natural language translation, where they revolutionized the field by enabling high-quality, context-aware machine translation. Whereas transformers have since become ubiquitous across machine learning (e.g. Chen et al. 2021, Khan et al. 2022, Moussad et al. 2023), in this work we propose again to use transformers towards translation—of animal communication.

While transformers have been used in settings where parallel data is nonexistent [Conneau and Lample, 2019] and for translation from audio [Kano et al., 2021], applying these advances to animal communication presents deep challenges. Even merely defining the problem has been the subject of studies spanning theoretical computer science [Goldwasser et al., 2023], biology [Yovel and Rechavi, 2023, Amphaeris et al., 2023], linguistics [Berwick and Chomsky, 2016, Amphaeris et al., 2022], and philosophy [Suzuki et al., 2020, Hobaiter et al., 2022].

Existing approaches to modeling sperm whale codas have made significant advances. Bermant et al. [2019] developed effective methods for coda detection and classification, while generative models based on GANs [Beguš et al., 2023, Kopets et al., 2024] have shown the potential for synthesizing coda-like audio. The aforementioned timing-based analyses of Leitao et al. [2024], Sharma et al. [2024b] have yielded new insight into the social and behavioral aspects of sperm whale communication.

Our work will address challenges left open by these works: While GAN-based models can generate coda-like audio [Beguš et al., 2023, Kopets et al., 2024], they cannot easily condition on a given context. Timing-based approaches [Leitao et al., 2024, Sharma et al., 2024b] capture important temporal patterns but may miss features only present in the raw audio, such as the recently discovered vowel-like spectral patterns [Beguš et al., 2025]. Moreover, current methods train separate models for classification [Bermant et al., 2019] and generation, despite the intuition that a model capable of realistic generation should also learn representations useful for classification. Lastly, none of these tackled the issue of *translating across acoustic domains*.

To address these challenges, we introduce the Whale Acoustics Model (WhAM, Figure 1), a new approach to modeling sperm whale codas that unifies three capabilities:

- **Acoustic translation:**[1] WhAM can translate an audio prompt (e.g. other animal vocalizations or even noise) into the acoustic style of sperm whale codas, acting as a form of cross-domain style transfer.
- **Generation:** WhAM can generate novel "pseudocodas" that are perceptually similar to real codas, as evaluated by expert listeners.

---

[1] We emphasize that translation is in the acoustic sense; semantic translation remains a distinct and more ambitious goal.

- **Classification:** WhAM's learned representations are useful for a range of classification tasks, including rhythm type [Sharma et al., 2024a], social unit classification [Best, 1979, Christal and Whitehead, 2001, Gero et al., 2016b], and the recently discovered vowel-like spectral patterns of Beguš et al. [2025]—despite being trained primarily for generation.

Contributions. This paper presents the first unified model of sperm whale codas capable of acoustic translation, generation, and classification. Notably, WhAM demonstrates that meaningful bioacoustic features emerge from purely generative training, aligning with recent work on self-supervised (non-generative) modeling of animal vocalizations [Hagiwara, 2023].

WhAM serves as a proof of concept, applying advances in neural audio modeling to bioacoustics in a novel and unifying way. To facilitate further research, we will release the model and its training and evaluation code. Remarkably, WhAM achieves strong results after just five days of training on a single GPU. While the dataset is small compared to those used for large audio models [Borsos et al., 2023, Agostinelli et al., 2023], our results suggest that scaling up could yield even greater improvements.

Finally, WhAM was developed in close collaboration with marine biologists and underwater acousticians with domain expertise in sperm whale vocalizations. The model was evaluated through perceptual studies conducted by an interdisciplinary team of specialists. To our knowledge, this is the first paper to evaluate the perception of experts on synthetically generated codas, igniting a crucial discourse for validating the utility of generative models in bioacoustics research. Code, model, and data are available at `https://github.com/Project-CETI/wham`

Outline. Section 2 reviews related work. Section 3 details our methodological framework. Section 4 presents experimental results and expert analysis. Section 5 discusses future work.

2 Related work

Audio Generation. The vast majority of studies on deep generative audio models focus on human speech or music (e.g. van den Oord et al. 2016, Dong et al. 2018, Dhariwal et al. 2020, Lakhotia et al. 2021, Agostinelli et al. 2023). Some works are dedicated to generating the vocalizations of animals (bioacoustics) such as birds [Bhatia and Kinnunen, 2022, Guei et al., 2024], mice [Reilly et al., 2023], cetaceans [Bergler et al., 2022, Zhang et al., 2022, Honghui and Lanhao, 2022, Kim et al., 2024], and in particular sperm whales [Beguš et al., 2023, Kopets et al., 2024]. However, to our knowledge, all techniques for bioacoustic generation are based on generative adversarial networks (GANs). Unlike our transformer-based WhAM, GANs do not allow for conditioning on context in the form of an audio prompt. We emphasize that WhAM enables translation of input sounds into the *acoustic style* of sperm whale vocalizations, operating purely at the signal level. This is distinct from *semantic* translation between communication systems, which remains a far more ambitious goal requiring a deep understanding of animal cognition and communication (e.g. Goldwasser et al. 2023, Yovel and Rechavi 2023, Amphaeris et al. 2023).

Animal Vocalization Modeling. Deep learning techniques have been applied towards other, non-generative, ends in bioacoustics research. Learned audio representations have been used for species recognition Chen et al. [2014], Hafemann et al. [2014], Xu et al. [2019], Kahl et al. [2021], Xie et al. [2023] and automatic annotation (i.e., vocalization detection and classification) of bioacoustic data Bergler et al. [2019], Coffey et al. [2019], Bermant et al. [2019], Premoli et al. [2021]. AVES [Hagiwara, 2023] utilizes HuBERT's [Hsu et al., 2021] self-supervised learning framework towards state-of-the-art performance in species classification and detection tasks. While AVES demonstrates the power of learned audio representations, its encoder-only architecture limits it to analysis tasks, contrasting with WhAM's generative capabilities. As we show in Section 4.3, while AVES outperforms WhAM on classification tasks as expected given its specialized design, WhAM still learns meaningful representations as a byproduct of its generative training, outperforming baseline approaches despite having a different primary objective.

Sperm whale communication. Understanding sperm whale communication has been a central challenge in marine biology for over six decades (Backus and Schevill 1966, Watkins and Schevill 1977, Whitehead and Weilgart 1991, Andreas et al. 2022; see also Appendix B). Recent computational approaches have focused on analyzing click timing patterns within codas and do not directly address the acoustic properties of individual clicks within codas [Sharma et al., 2024a, Leitao et al., 2024, Sharma et al., 2024b]. WhAM extends this computational trajectory by enabling systematic manipulation of click acoustics, potentially allowing a quantitative analysis of acoustic variations between clan dialects and investigation of features that make codas recognizable. While WhAM's synthetic codas may not yet match the quality needed for playback experiments, WhAM represents progress towards stimuli generation in a responsible behavioral study (Tyack 1983, Deecke 2006, King and Jensen 2023; see also Appendix A).

3 Methods

3.1 Masked Acoustic Token Modeling with VampNet

VampNet García et al. [2023] is an audio-to-audio generative model, pretrained on 797k music tracks from thousands of artists. VampNet consists of three neural models: a tokenizer, a coarse-token model, and a coarse-to-fine model. For simplicity of presentation we will avoid the distinction between coarse and fine tokens, instead decomposing VampNet into an *Acoustic Tokenizer* and a *Masked Acoustic Token Model*. The reader is referred to García et al. [2023] for full details of the model, and Appendix E.2 for a specification of hyperparameters used in training WhAM.

Acoustic Tokenizer. The tokenizer takes as input an N_{sec}-second audio snippet sampled at N_{sam} Hz, and outputs a sequence of ℓ discrete tokens from a finite vocabulary Σ. A jointly-trained detokenizer will convert token sequences back into audio:

$$\begin{aligned} T &: \mathbb{R}^{N_{\mathrm{sec}} \times N_{\mathrm{sam}}} \to \Sigma^{\ell} \\ T^{-1} &: \Sigma^{\ell} \to \mathbb{R}^{N_{\mathrm{sec}} \times N_{\mathrm{sam}}}. \end{aligned}$$

VampNet uses a residual vector quantization approach known as the Descript Audio Codec (DAC, Kumar et al. 2023). At a high level, audio is tokenized in a temporal and hierarchical fashion, such that each interval of samples is replaced with a "stack" of tokens; this means that neighboring stacks of tokens correspond to contiguous intervals of samples in the audio. For example, the first five token stacks $(\sigma_1, \ldots, \sigma_5)$ could correspond to the first 0.5 seconds of audio.

Masked Acoustic Token Model (MATM). A bidirectional transformer M is trained to perform the cloze task on acoustic token sequences. That is, each audio snippet in the pretraining dataset is tokenized, and then a bidirectional transformer is trained to predict a random subset of masked tokens.

$$M : (\Sigma \cup \{\texttt{[MASK]}\})^{\ell} \to \Sigma^{\ell}$$

A pretrained MATM can be finetuned in various ways. Following García et al. [2023], we finetune using Low Rank Adaptation (LoRA, Hu et al. 2022).

Generation. After training a tokenizer T, detokenizer T^{-1} and a (possibly finetuned) MATM M, VampNet can be used to generate variations of given "prompt" audio snippets. This is done in the natural way, by randomly masking the tokenized audio; importantly, the masking scheme used in generation time does *not* need to be uniformly random. For example, the scheme can leave (classically-detected) *beats* unmasked, so as to preserve the rhythm of the prompt. Rather than generating all masked tokens simultaneously (e.g. as in BERT, Devlin et al. 2019), VampNet uses *iterative parallel decoding* [Chang et al., 2022] wherein tokens are gradually "unmasked" in a sequence of forward passes through the model.

3.2 Data

WhAM is trained by finetuning VampNet (Section 3.1) on various datasets.

FSD. The Freesound Dataset [Font et al., 2013] consists of 50k human-labeled recordings. We used recordings with the *animal* tag, which totaled 7h45m of audio.

AudioSet. A dataset of two million human-labeled audio clips taken from YouTube [Gemmeke et al., 2017a]. Of these, we used audio with the *animal* tag, totaling at about 5 hours.

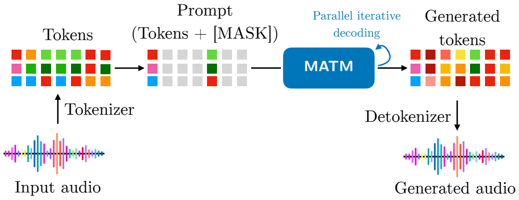

Figure 2: Overview of VampNet's generation pipeline. Input audio is first converted into a grid of tokens by the Tokenizer. These tokens are then partially masked to create a prompt. The Masked Acoustic Token Model (MATM) uses parallel iterative decoding to generate new tokens, which are finally converted back into audio by the Detokenizer. The colored squares represent acoustic tokens, with grey squares indicating masked positions.

WMMS. The Watkins Marine Mammal Sound Database [Sayigh et al., 2016] totaling 4h8m. It includes audio collected over seven decades in at least 67 sites around the world. Sperm whales are among the 51 species recorded.

BirdSet. An avian bioacoustics dataset curated for classification tasks [Rauch et al., 2025], totalling about 6,800 hours. Due to computational limits, we used a 110-hour subset of audio dense with vocalizations.

DSWP. A dataset of 2507 annotated codas (1h26m) collected over thirteen years in a 2000km^2 area off the coast of Dominica. It consists of codas recorded using far-field boat-based hydrophones and noninvasive animal-borne tags.

CETI. A more recent dataset of sperm whale vocalizations consisting of 7653 annotated codas (4h33m) collected similarly to DSWP.

The training of WhAM is split into two phases: (1) *Domain adaptation*, in which the base VampNet is finetuned on FSD+AudioSet+WMMS for 500k iterations; (2) *species-specific finetuning*, in which domain-adapted VampNet is finetuned on DSWP+CETI for an additional 500k iterations. Both phases follow the same (LoRA) finetuning procedure, but we find this split to be conceptually useful. Additional details are deferred to Appendix E.1

4 Results

We evaluate WhAM through three complementary analyses. First, we assess the quality of WhAM's synthetic codas through quantitative metrics, specifically the Fréchet Audio Distance (FAD, Kilgour et al. 2019) between generated and natural codas. Second, we conduct a perceptual study with expert marine biologists to evaluate how well our synthetic codas preserve the characteristic features of natural sperm whale vocalizations. Finally, we evaluate WhAM's learned representations on downstream classification tasks to investigate whether our model captures meaningful acoustic features of sperm whale communication.

4.1 Fréchet Distance of Audio Translation

A key aspect of WhAM is its ability to "translate" audio inputs into the acoustic style of sperm whale codas. To evaluate this capability quantitatively, we measure the Fréchet Audio Distance (FAD, Kilgour et al. 2019) between natural and WhAM-generated synthetic codas. FAD measures the similarity between two audio datasets by comparing embeddings of the audio signals; lower FAD indicates greater acoustic similarity between the datasets.

FAD is computed using a given audio embedding model. We chose BirdNET [Kahl et al., 2021] based on a principled calibration experiment that compared the sensitivity of five embedding models to the rhythmic patterns crucial to coda structure (Appendix D.1). We normalize FAD values by dividing by the maximum distance, scaling all values to $[0, 1]$. Figure 3 portrays WhAM's translation ability using audio prompts from three domains:

1. *Natural codas (S. Whale)*: A disjoint set of codas produced by sperm whales. The FAD between disjoint sets of natural codas is 0.21 (rather than zero) due to variance in recording

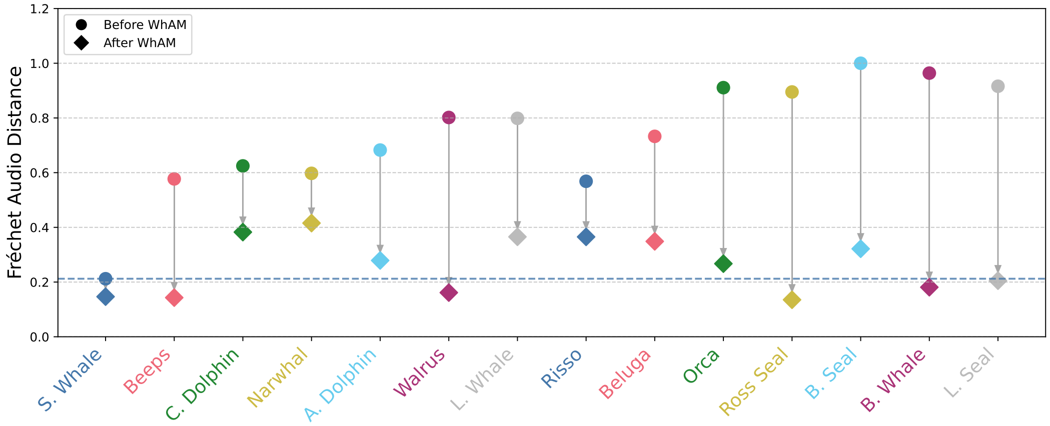

Figure 3: Normalized Fréchet Audio Distance between sperm whale codas and various audio sources, before and after translation through WhAM. Lower FAD indicates greater acoustic similarity to natural codas. The horizontal line at 0.21 represents the baseline FAD between disjoint sets of natural codas. Full names of animals along with the number of samples from each can be found in Table 3.

conditions, individual whales, and coda types. This establishes a baseline below which FAD fails to distinguish audio sources from natural variation: We therefore say that generated outputs are *FAD-indistinguishable* when their FAD falls below 0.21. When passing natural codas through WhAM, we expect a slight decrease in FAD as WhAM regularizes inputs toward the mean embedding of its training distribution (a large dataset of diverse codas).

2. *Animal sounds*: Vocalizations from 12 species of marine mammals. Figure 3 shows that WhAM consistently reduces the acoustic distance to natural codas, effectively translating these diverse inputs into the acoustic style of sperm whale codas. WhAM-generated outputs of four species are FAD-indistinguishable from natural codas.
3. *Digital "beeps"*: Artificial audio generated by initializing an array of zeros and randomly selecting points to assign a peak amplitude of 1. Remarkably, beeps and natural codas have approximately the same post-WhAM FAD. This may be because beeps' sparse structure (mostly silence with isolated peaks) gives WhAM freedom to infill patterns close to the mean embedding of codas, while natural codas with minimal silence constrain the model's regularization but start closer to the target distribution.

The results demonstrate WhAM's remarkable translation capability: five diverse audio sources (four non-whale species and digital beeps) become FAD-indistinguishable from natural sperm whale codas after processing. This success across varied inputs suggests that WhAM has learned a robust representation of the essential acoustic properties that define sperm whale communication.

4.2 Expert Perceptual Study

To evaluate the perceptual quality of WhAM's synthetic codas, we conducted a comprehensive study with domain experts to assess how well our generated outputs match natural sperm whale vocalizations with respect to a human-expert distinguisher. This study measures both audio-only and spectrogram-based discrimination, while also gathering qualitative insights about specific acoustic features that distinguish synthetic from natural codas. Additional details are deferred to Appendix E.5.

Expert backgrounds. Five academic experts participated in the perceptual study. Three identified as marine biologists, and two as underwater acoustics specialists. They listed between 3 and 20 years of experience working with coda audio (field recordings), manual detection and classification, and the development of automatic detection systems. All experts had experience analyzing coda audio and spectrograms, which are the two media through which the experiment was carried out.

Experiment design. We designed a four-task study to be completed sequentially by each expert:

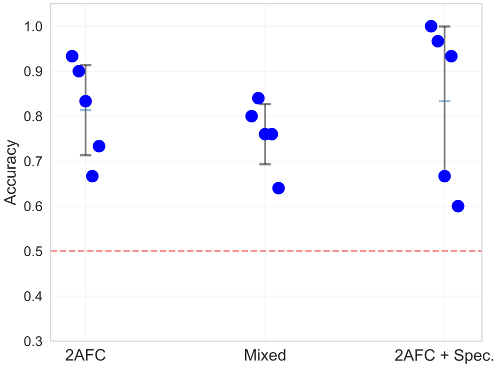

Figure 4: Expert performance on audio-only 2AFC (Task 1), mixed classification (Task 2), and spectrogram-assisted 2AFC (Task 3). Error bars show standard deviation across experts. While all tasks elicited above-chance performance (dashed line), spectrogram analysis showed the greatest variability between experts ($\sigma = 0.17$). Task 1 and 3 had 30 pairs each, Task 2 had a collection of 25 samples.

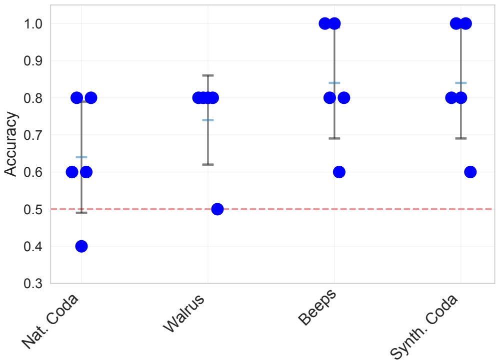

Figure 5: Accuracy in mixed classification (Task 2) for different input domains. Natural codas (left) were misclassified as synthetic 36% of the time. The remaining columns depict performance on synthetic codas generated by WhAM from walrus vocalizations, non-coda acoustic impulses, and codas (respectively). There were five synthetic codas from each domain, plus ten natural codas for a total of 25 items.

1. **Audio-only two-alternative forced choice (2AFC):** Experts compared pairs of codas (one natural, one synthetic) in audio-only conditions, and were asked to identify the synthetic coda. Synthetic codas were generated by WhAM using the paired natural coda as input.
2. **Mixed Collection Classification:** Experts classified clips as natural or synthetic. Clips were either natural codas, or synthetic codas generated from different sources: natural codas, digital beeps,[2] or walrus vocalizations [Sayigh et al., 2016]. False positives (natural misclassified) and false negatives (synthetic undetected) were measured.
3. **Spectrogram-assisted 2AFC:** Experts repeated the first task while visualizing audio with software of their choice. The experts were given the exact same samples as in the first task, ensuring direct comparability between audio-only and spectrogram-aided performance. This task mirrored real-world analysis workflows while quantifying the perceptual "advantage" of multimodal inspection.
4. **Qualitative assessment:** Experts were given five representative samples of synthetic codas. They were then asked questions about how well synthetic codas captured / missed characteristics of natural codas, whether any non-natural patterns appeared in synthetic codas, and which features did they use to distinguish between codas in each of the previous tasks.

Fleiss's κ quantified inter-expert agreement [Fleiss, 1971], and accuracy was calculated relative to ground-truth labels. Task order was chosen towards minimizing bias (audio-first to avoid visual priming), with background information collected in a final section. The samples used in the experiment are attached to this submission as supplementary material. Experimental details are in Appendix E.5.

Quantitative analysis. Experts achieved 81% accuracy ($\kappa = 0.41$), in audio-only 2AFC (Task 1), rising marginally to 83% ($\kappa = 0.41$) with spectrograms visualized (Task 3). This 2% improvement suggests WhAM's synthetic codas lack glaring spectro-temporal artifacts detectable by trained analysts. As expected, accuracy with spectrograms was generally better per-expert, with one expert's performance dramatically increasing from 66% to 93% (another expert even achieved a perfect score). Surprisingly, one expert's performance *decreased* from 83% in Task 1 to 66% in Task 3; comments in the qualitative section did not suggest an explanation.

Performance varied substantially across tasks and among experts (Figure 4). The most experienced expert ranked highest in both 2AFC tasks, but not in mixed classification. These variations reflect

---

[2] i.e., an artificial sequence of clicks

diverging expert strategies—some focused on inter-click patterns, others on spectral properties: "rhythm" to quote one expert, versus "DC offsets" and "inter-pulse structures" [Møhl et al., 2003] to quote others.

Misclassification rates in Task 2 (Figure 5) revealed WhAM's efficacy in acoustic translation: on average, experts correctly flagged walrus-to-coda audio only 75% of the time—less than digital beeps or coda-to-coda outputs of WhAM. For one expert, walrus-to-coda audio was detected only 50% of the time (random chance).

Fleiss's κ values (0.41–0.44) indicated moderate agreement across tasks, with experts showing greatest consensus on mixed classification ($\kappa = 0.44$). Performance on spectrogram-aided 2AFC performance was the most diverse—one expert achieved perfect performance while another approached chance (60%).

Qualitative feedback. Synthetic codas successfully replicated key acoustic features of natural codas. Most experts noted preservation of *rhythm*, referred to as inter-click intervals (ICI); that is, clicks occur "at the right time" in synthetic codas. Additionally, one expert answered that "spectral components" were overall preserved in synthetic codas.

That said, experts identified missing components which can be partitioned into three categories:

- *Within a single click:* Some clicks "came on and disappeared too strongly," had "varying amplitude [within a single coda]," and "inverted peaks." On a spectral level, an expert answered that clicks were too "broadband" compared to natural clicks which have a low-frequency bias.
- *Rhythmic/temporal:* One expert noted that the timing of clicks fit echolocation moreso than codas. (See Appendix B for how they differ.)
- *Recording-level anomalies:* One expert noted a "DC offset" which they described as the unrealistic background noise on synthetic codas. Similarly, another noted that background noise in synthetic codas oscillated too much.

Based on this assessment, we present in Appendix C a guide to the similarities and differences between natural and synthetic codas, supplemented by annotated spectrograms.

4.3 Utility of embeddings for downstream tasks

We test whether WhAM's internal representations capture meaningful features of sperm whale vocalizations through three downstream classification tasks. For each task, we train a small (two-layer) classifier head that takes coda embeddings as input. We compare WhAM to naive random-embedding and majority-class baselines, as well as AVES [Hagiwara, 2023], a self-supervised model achieving state-of-the-art performance on bioacoustic classification tasks. Full details of the experimental setup are deferred to Appendix E.4.

The downstream tasks are:

1. *Coda detection*: Given a snippet of audio, determine whether it contains a coda. The classifier is trained on positive (coda) and negative (no coda) snippets, with negative examples drawn from the same recording conditions to ensure the model learns coda features rather than recording artifacts.
2. *Rhythm type*: Given a snippet of audio, classify its temporal pattern. Rhythm of inter-click intervals serves as a key axis for classification of sperm whale codas in cetacean research Schulz et al. [2011], Bermant et al. [2019], Sharma et al. [2024a].
3. *Social unit classification*: The lowest level of sperm whale social structure are called social units (SU) and have stable, matrilineally-related membership of females and their young [Christal et al., 1998]. Importantly, all SUs in DSWP+CETI belong to the same vocal clan and thus share a common repertoire of coda types, making this more of a speaker identification task than dialect classification.[3]

---

[3] By analogy to human language, consider the task of classifying speakers by city of origin. It would be easier to distinguish between speakers from cities that use different dialects (simply classify the dialect). Importantly, in our data, all speakers use the same dialect.

Table 1: Classification accuracies (%) of different audio embeddings. Each classifier head was trained using three different random seeds, with mean±stderr reported. The Random baseline uses a randomly initialized AVES model (training only the classifier), while Majority predicts the most common class.

Task	WhAM	Baseline		AVES	BirdNET	CLAP
		Rand.	Maj.			
Detection	91.3±0.2	60.9	60.9	92.8±0.1	93.0±1.0	**96.8**±1.4
Rhythm	87.4±1.6	66.3	60.9	90.4±1.6	88.6±0.2	**92.4**±2.4
Social Unit	70.5±5.6	42.5	35.1	92.0±0.7	**93.2**±0.1	85.5±1.4
Vowel	85.2±2.5	66.3	66.3	**91.8**±2.9	85.9±4.6	84.3±0.9

4. *Vowel type*: Given a coda recording, classify its vowel-like pattern [Beguš et al., 2025].

Table 1 shows classification accuracies for each task. As expected, non-generative embeddings specifically designed for acoustic classification tasks outperform outperform WhAM. We view those as a ceiling "sanity check" than a baseline. AVES and BirdNET perform particularly well on more specialized bio-acoustic tasks, due to the fact that they were both trained on large amounts of animal vocalizations. Notably, WhAM's representations are useful despite being trained only for generation, outperforming both naive baselines. This suggests that meaningful acoustic features emerge naturally from training for coda generation, even without explicit supervision for these tasks.

We conducted an ablation study to assess how fine-tuning affects embedding quality by evaluating different WhAM variants with specific components removed (detailed in Appendix D.2). The results reveal that fine-tuning did not significantly alter WhAM's downstream utility compared to base VampNet embeddings, despite WhAM's specialization on whale codas. However, as shown in Appendix D.3, species-specific fine-tuning was essential for enabling WhAM's core capability of translating audio into sperm whale vocalization acoustics.

5 Limitations and Future Work

The most immediate technical limitation concerns the audio codec architecture. Our current implementation only finetunes the MATM while keeping the codec fixed (see Section 3.1). This design choice, while computationally efficient, may limit the model's ability to capture nuanced acoustic features specific to sperm whale vocalizations. For instance, the recently discovered vowel-like patterns in the 3.7–5.7kHz band [Beguš et al., 2025] may be inadequately represented by a codec primarily trained on human music. Future work could explore either finetuning the entire codec or developing specialized codecs for bioacoustic signals.

Expert feedback (Section 4.2) highlighted specific limitations in click generation: unnatural onset and decay patterns, inconsistent background noise, and click properties more reminiscent of echolocation than communication codas. These limitations might be addressed through architectural modifications, such as incorporating adversarial components [Beguš et al., 2023] or introducing specialized modules that leverage domain knowledge about sperm whale click structure. Notably, the observation about echolocation-like properties led to an unexpected finding in our dataset preparation: the presence of echolocation sequences in datasets intended for communication codas. This discovery highlights a broader challenge in bioacoustics research—the difficulty of building clean, well-labeled datasets at scale. Future work should focus on developing robust methods for distinguishing between different types of vocalizations, perhaps by leveraging existing detection systems [Bermant et al., 2019].

These data quality challenges underscore the importance of thorough evaluation protocols. Expanding the expert panel would provide more robust perceptual assessments, though we acknowledge the practical challenges in recruiting specialists in sperm whale vocalizations. Additionally, developing more principled evaluation methods—and meta-evaluating these—would help establish standardized benchmarks for bioacoustic generation tasks.

While our results demonstrate impressive performance with relatively small datasets—orders of magnitude smaller than typical in modern acoustic model training—scaling up the training data could yield substantial improvements. This would require significant effort in aggregating and preprocessing additional sperm whale datasets, as our experience with DSWP+CETI highlighted the technical challenges involved in preparing bioacoustic data for machine learning pipelines.

Looking beyond technical improvements, future work could explore unsupervised learning approaches to uncover new coda features, following the success of similar approaches in bioacoustics [Beguš et al., 2025]. This could lead to discoveries about sperm whale communication that complement traditional analytical methods while providing new directions for improving generative models of animal vocalizations.

Our methodological framework—from the two-phase training approach to the expert evaluation protocol—could be adapted for studying other animal communication systems. Our experience suggests that success will require careful attention to species-specific acoustic features and close collaboration with domain experts who can identify subtle but important characteristics of vocalizations.

The gap between generating vocalizations and understanding their meaning remains vast. While WhAM represents the first attempt at acoustic translation in the context of sperm whale communication, future work should explore ways to bridge this semantic gap while maintaining minimal assumptions about the underlying communication system.

Acknowledgments and Disclosure of Funding

We thank Guy Gubnitsky, Yaly Mevorach, and Pernille Tonnesen for volunteering their time and expertise to the expert perceptual study. We thank anonymous reviewers for helpful comments, and in particular reviewer fPAV for suggesting adding BirdSet [Rauch et al., 2025] to the domain adaptation stage.

This study was funded by Project CETI via grants from Dalio Philanthropies and Ocean X; Sea Grape Foundation; Virgin Unite and Rosamund Zander/Hansjorg Wyss through The Audacious Project: a collaborative funding initiative housed at TED. This work was supported, in part, by National Science Foundation Award # 2222369.

Fieldwork for The Dominica Sperm Whale Project, which produced many of the coda recordings used in this work, was supported by a FNU fellowship for the Danish Council for Independent Research supplemented by a Sapere Aude Research Talent Award (1325-00047A), a Carlsberg Foundation expedition grant (CF14-0789), a grant from Focused on Nature, two Explorer Grants from the National Geographic Society (WW-218R-17 and NGS-64863R-19), and supplementary grants from the Arizona Center for Nature Conservation, Quarters For Conservation to SG, with supplemental funding from the Dansk Akustisks Selskab, Oticon Foundation, and the Dansk Tennis Fond to Pernille Tonnesen. Further funding was provided by Discovery and Equipment grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Hal Whitehead of Dalhousie University and a FNU large frame grant and a Villum Foundation Grant (13273) to Peter Madsen of the Marine Bioacoustics Lab at Aarhus University. We thank the Chief Fisheries Officers and the Dominica Fisheries Division officers for research permits and their collaboration in data collection; all the crews of R/V Balaena and The DSWP team for data collection, curation, and annotation of audio recordings and photoID; as well as Dive Dominica, Al Dive, W.E.T. Dominica, and Wacky Rollers for logistical support while in Dominica. We are grateful to Kristian Beedholm of the Marine Bioacoustics Lab at Aarhus University for CodaSorter; as well as Mark Johsson and Peter Tyack for in-kind contributions of Dtags and associated code.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. Musiclm: Generating music from text. *CoRR*, abs/2301.11325, 2023. doi: 10.48550/ARXIV.2301.11325. URL `https://doi.org/10.48550/arXiv.2301.11325`.

Jenny Amphaeris, Graeme Shannon, and Thora Tenbrink. Overlap not gap: Understanding the relationship between animal communication and language with prototype theory. *Lingua*, 272: 103332, 2022. ISSN 0024-3841. doi: https://doi.org/10.1016/j.lingua.2022.103332. URL `https://www.sciencedirect.com/science/article/pii/S0024384122000936`.

Jenny Amphaeris, Daniel T Blumstein, Graeme Shannon, Thora Tenbrink, and Arik Kershenbaum. A multifaceted framework to establish the presence of meaning in non-human communication. *Biological Reviews*, 98(6):1887–1909, 2023.

Jacob Andreas, Gašper Beguš, Michael M Bronstein, Roee Diamant, Denley Delaney, Shane Gero, Shafi Goldwasser, David F Gruber, Sarah de Haas, Peter Malkin, et al. Toward understanding the communication in sperm whales. *Iscience*, 25(6), 2022.

Ricardo Antunes, Tyler Schulz, Shane Gero, Hal Whitehead, Jonathan Gordon, and Luke Rendell. Individually distinctive acoustic features in sperm whale codas. *Animal Behaviour*, 81(4):723–730, 2011. ISSN 0003-3472. doi: https://doi.org/10.1016/j.anbehav.2010.12.019. URL `https://www.sciencedirect.com/science/article/pii/S0003347210005233`.

Richard H Backus and William E Schevill. Physeter clicks. *Whales, dolphins and porpoises*, 510: 527, 1966.

Gašper Beguš, Andrej Leban, and Shane Gero. Approaching an unknown communication system by latent space exploration and causal inference. *arXiv preprint arXiv:2303.10931*, 2023.

Gašper Beguš, Ronald L. Sprouse, Andrej Leban, Marcelo Silva, and Shane Gero. Vowel- and diphthong-like spectral patterns in sperm whale codas. *Open Mind*, pages 1—26, 2025. doi: 10.1162/OPMI.a.252.

Christian Bergler, Hendrik Schröter, Rachael Xi Cheng, Volker Barth, Michael Weber, Elmar Nöth, Heribert Hofer, and Andreas Maier. Orca-spot: An automatic killer whale sound detection toolkit using deep learning. *Scientific reports*, 9(1):10997, 2019.

Christian Bergler, Alexander Barnhill, Dominik Perrin, Manuel Schmitt, Andreas K. Maier, and Elmar Nöth. ORCA-WHISPER: an automatic killer whale sound type generation toolkit using deep learning. In Hanseok Ko and John H. L. Hansen, editors, *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2413–2417. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-846. URL `https://doi.org/10.21437/Interspeech.2022-846`.

Peter C Bermant, Michael M Bronstein, Robert J Wood, Shane Gero, and David F Gruber. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific reports*, 9(1):12588, 2019.

Robert C Berwick and Noam Chomsky. *Why only us: Language and evolution*. Cambridge, MA: MIT Press, 2016.

Peter B Best. Social organization in sperm whales, physeter macrocephalus. In *Behavior of marine animals: Current perspectives in research*, pages 227–289. Springer, 1979.

Rhythm Bhatia and Tomi H. Kinnunen. An initial study on birdsong re-synthesis using neural vocoders. In S. R. Mahadeva Prasanna, Alexey Karpov, K. Samudravijaya, and Shyam S. Agrawal, editors, *Speech and Computer - 24th International Conference, SPECOM 2022, Gurugram, India, November 14-16, 2022, Proceedings*, volume 13721 of *Lecture Notes in Computer Science*, pages 64–74. Springer, 2022. doi: 10.1007/978-3-031-20980-2_7. URL `https://doi.org/10.1007/978-3-031-20980-2_7`.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409. URL `https://doi.org/10.1109/TASLP.2023.3288409`.

Benjamin Britten. *The Young Person's Guide to the Orchestra, Op. 34: Variations and Fugue on a Theme of Purcell*. Boosey & Hawkes, London, 1945.

Claudio Campagna and Daniel Guevara. "save the whales" for their natural goodness. In *Marine Mammals: The Evolving Human Factor*, pages 397–424. Springer, 2022.

Maurício Cantor and Hal Whitehead. How does social behavior differ among sperm whale clans? *Marine Mammal Science*, 31(4):1275–1290, 2015.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11305–11315. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01103. URL `https://doi.org/10.1109/CVPR52688.2022.01103`.

Guobin Chen, Tony X. Han, Zhihai He, Roland Kays, and Tavis Forrester. Deep convolutional neural network based species recognition for wild animal monitoring. In *2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*, pages 858–862. IEEE, 2014. doi: 10.1109/ICIP.2014.7025172. URL `https://doi.org/10.1109/ICIP.2014.7025172`.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 15084–15097, 2021.

Jenny Christal and Hal Whitehead. Social affiliations within sperm whale (physeter macrocephalus) groups. *Ethology*, 107(4):323–340, 2001.

Jenny Christal, Hal Whitehead, and Erland Lettevall. Sperm whale social units: variation and change. *Canadian Journal of Zoology*, 76(8):1431–1440, 1998.

Kevin R Coffey, Ruby E Marx, and John F Neumaier. Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology*, 44(5):859–868, 2019.

Marie Comuzzo. Singing with whales: Exploring human and non-human connections. *SEM Student News*, 19(1), 2023. ISSN 2578-4242. URL `https://cdn.ymaws.com/ethnomusicology.site-ym.com/resource/group/dc75b7e7-47d7-4d59-a660-19c3e0f7c83e/publications/19_1musicanthropocene/comuzzo_semsn_19-1.pdf`. Special Issue: Music and the Anthropocene.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html`.

Volker B Deecke. Studying marine mammal cognition in the wild: a review of four decades of playback experiments. *Aquatic mammals*, 32(4):461–482, 2006.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023, 2023. URL `https://openreview.net/forum?id=ivCd8z8zR2`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL `https://doi.org/10.18653/v1/n19-1423`.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *CoRR*, abs/2005.00341, 2020. URL `https://arxiv.org/abs/2005.00341`.

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 34–41. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11312. URL `https://doi.org/10.1609/aaai.v32i1.11312`.

Alaina Claire Feldman. Minor listening, major influence: Revisiting songs of the humpback whale. *e-flux journal*, (118), May 2021. URL `https://www.e-flux.com/journal/118/394434/minor-listening-major-influence-revisiting-songs-of-the-humpback-whale`. Accessed: October 2025.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5):378, 1971.

Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In Alejandro Jaimes, Nicu Sebe, Nozha Boujemaa, Daniel Gatica-Perez, David A. Shamma, Marcel Worring, and Roger Zimmermann, editors, *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pages 411–412. ACM, 2013. doi: 10.1145/2502081.2502245. URL `https://doi.org/10.1145/2502081.2502245`.

Hugo Flores García, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo. Vampnet: Music generation via masked acoustic token modeling. In Augusto Sarti, Fabio Antonacci, Mark Sandler, Paolo Bestagini, Simon Dixon, Beici Liang, Gaël Richard, and Johan Pauwels, editors, *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, pages 359–366, 2023. doi: 10.5281/ZENODO.10265299. URL `https://doi.org/10.5281/zenodo.10265299`.

Jort F. Gemmeke, Daniel P. W. 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 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE, 2017a. doi: 10.1109/ICASSP.2017.7952261. URL `https://doi.org/10.1109/ICASSP.2017.7952261`.

Jort F. Gemmeke, Daniel P. W. 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 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE, 2017b. doi: 10.1109/ICASSP.2017.7952261. URL `https://doi.org/10.1109/ICASSP.2017.7952261`.

Shane Gero, Anne Bøttcher, Hal Whitehead, and Peter Teglberg Madsen. Socially segregated, sympatric sperm whale clans in the atlantic ocean. *Royal Society Open Science*, 3(6):160061, 2016a.

Shane Gero, Hal Whitehead, and Luke Rendell. Individual, unit and vocal clan level identity cues in sperm whale codas. *Royal Society Open Science*, 3(1):150372, 2016b.

Shafi Goldwasser, David F. Gruber, Adam Tauman Kalai, and Orr Paradise. A theory of unsupervised translation motivated by understanding animal communication. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Axel-Christian Guei, Sylvain Christin, Nicolas Lecomte, and Éric Hervet. Ecogen: Bird sounds generation using deep learning. *Methods in Ecology and Evolution*, 15(1):69–79, 2024.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 1331–1335. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446663. URL `https://doi.org/10.1109/ICASSP48485.2024.10446663`.

Luiz G. Hafemann, Luiz S. Oliveira, and Paulo Rodrigo Cavalin. Forest species recognition using deep convolutional neural networks. In *22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014*, pages 1103–1107. IEEE Computer Society, 2014. doi: 10.1109/ICPR.2014.199. URL `https://doi.org/10.1109/ICPR.2014.199`.

Masato Hagiwara. AVES: animal vocalization encoder based on self-supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095642. URL `https://doi.org/10.1109/ICASSP49357.2023.10095642`.

Taylor A. Hersh, Shane Gero, Luke Rendell, Maurício Cantor, Lindy Weilgart, Masao Amano, Stephen M. Dawson, Elisabeth Slooten, Christopher M. Johnson, Iain Kerr, Roger Payne, Andy Rogan, Ricardo Antunes, Olive Andrews, Elizabeth L. Ferguson, Cory Ann Hom-Weaver, Thomas F. Norris, Yvonne M. Barkley, Karlina P. Merkens, Erin M. Oleson, Thomas Doniol-Valcroze, James F. Pilkington, Jonathan Gordon, Manuel Fernandes, Marta Guerra, Leigh Hickmott, and Hal Whitehead. Evidence from sperm whale clans of symbolic marking in non-human cultures. *Proceedings of the National Academy of Sciences*, 119(37):e2201692119, 2022. doi: 10.1073/pnas.2201692119. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2201692119`.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 131–135. IEEE, 2017. doi: 10.1109/ICASSP.2017.7952132. URL `https://doi.org/10.1109/ICASSP.2017.7952132`.

Catherine Hobaiter, Kirsty E Graham, and Richard W Byrne. Are ape gestures like words? outstanding issues in detecting similarities and differences between human language and ape gesture. *Philosophical Transactions of the Royal Society B*, 377(1860):20210301, 2022.

Yang Honghui and Fang Lanhao. Simulation of marine mammal calls in deep-sea environment. In *International Conference on Autonomous Unmanned Systems*, pages 2911–2920. Springer, 2022.

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 Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291. URL `https://doi.org/10.1109/TASLP.2021.3122291`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Mark P Johnson and Peter L Tyack. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE journal of oceanic engineering*, 28(1):3–12, 2003.

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

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. Transformer-based direct speech-to-speech translation with transcoder. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pages 958–965. IEEE, 2021. doi: 10.1109/SLT48900.2021.9383496. URL `https://doi.org/10.1109/SLT48900.2021.9383496`.

Salman H. Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s):200:1–200:41, 2022. doi: 10.1145/3505244. URL `https://doi.org/10.1145/3505244`.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In Gernot Kubin and Zdravko Kacic, editors, *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 2350–2354. ISCA, 2019. doi: 10.21437/INTERSPEECH.2019-2219. URL `https://doi.org/10.21437/Interspeech.2019-2219`.

Yongcheol Kim, Seunghwan Seol, Hojun Lee, Geunho Park, and Jaehak Chung. Whistlegan for biomimetic underwater acoustic covert communication. *Electronics*, 13(5):964, 2024.

Stephanie L King and Frants H Jensen. Rise of the machines: Integrating technology with playback experiments to study cetacean social cognition in the wild. *Methods in Ecology and Evolution*, 14 (8):1873–1886, 2023.

Ekaterina Kopets, Tatiana Shpilevaya, Oleg Vasilchenko, Artur Karimov, and Denis Butusov. Generating synthetic sperm whale voice data using stylegan2-ada. *Big Data and Cognitive Computing*, 8 (4):40, 2024.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl_a_00430. URL `https://aclanthology.org/2021.tacl-1.79`.

Antonio Leitao, Maxime Lucas, Simone Poetto, Taylor A. Hersh, Shane Gero, David F. Gruber, Michael Bronstein, and Giovanni Petri. Evidence of social learning across symbolic cultural barriers in sperm whales. *eLife*, May 2024. doi: 10.7554/elife.96362.1. URL `http://dx.doi.org/10.7554/eLife.96362.1`.

Bertel Møhl, Magnus Wahlberg, Peter T Madsen, Anders Heerfordt, and Anders Lund. The monopulsed nature of sperm whale clicks. *The Journal of the Acoustical Society of America*, 114 (2):1143–1154, 2003.

Karen E Moore, William A Watkins, and Peter L Tyack. Pattern similarity in shared codas from sperm whales (physeter catodon). *Marine Mammal Science*, 9(1):1–9, 1993.

Bernard Moussad, Rahmatullah Roche, and Debswapna Bhattacharya. The transformative power of transformers in protein structure prediction. *Proceedings of the National Academy of Sciences*, 120(32):e2303499120, 2023.

Roger Payne. Songs of the humpback whale, 1970. See, in particular, "Listening Instructions".

Roger S Payne and Scott McVay. Songs of humpback whales: Humpbacks emit sounds in long, predictable patterns ranging over frequencies audible to humans. *Science*, 173(3997):585–597, 1971.

Marika Premoli, Daniele Baggi, Marco Bianchetti, Alessandro Gnutti, Marco Bondaschi, Andrea Mastinu, Pierangelo Migliorati, Alberto Signoroni, Riccardo Leonardi, Maurizio Memo, et al. Automatic classification of mice vocalizations using machine learning techniques and convolutional neural networks. *PloS one*, 16(1):e0244636, 2021.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. Birdset: A large-scale dataset for audio classification in avian bioacoustics. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=dRXxFEY8ZE`.

Johnny Reilly, John D Goodwin, Sihao Lu, and Andriy S Kozlov. Bidirectional generative adversarial representation learning for natural stimulus synthesis. *Journal of Neurophysiology*, 2023.

Luke Rendell, Sarah L Mesnick, Merel L Dalebout, Jessica Burtenshaw, and Hal Whitehead. Can genetic differences explain vocal dialect variation in sperm whales, physeter macrocephalus? *Behavior genetics*, 42:332–343, 2012.

Luke E Rendell and Hal Whitehead. Vocal clans in sperm whales (physeter macrocephalus). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512):225–231, 2003.

César Rodríguez-Garavito, David F. Gruber, Ashley Otilia Nemeth, and Gašper Beguš. What if we understood what animals are saying? the legal impact of AI-assisted studies of animal communication. *Ecology Law Quarterly*, 52(1), 2025. doi: 10.15779/Z383X83N5Q. URL `https://doi.org/10.15779/Z383X83N5Q`. Forthcoming, Fall 2025.

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.

Tyler M. Schulz, Hal Whitehead, Shane Gero, and Luke Rendell. Individual vocal production in a sperm whale (physeter macrocephalus) social unit. *Marine Mammal Science*, 27(1):149–166, 2011. doi: https://doi.org/10.1111/j.1748-7692.2010.00399.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1748-7692.2010.00399.x`.

Pratyusha Sharma, Shane Gero, Roger Payne, David F Gruber, Daniela Rus, Antonio Torralba, and Jacob Andreas. Contextual and combinatorial structure in sperm whale vocalisations. *Nature Communications*, 15(1):3617, 2024a.

Pratyusha Sharma, Shane Gero, Daniela Rus, Antonio Torralba, and Jacob Andreas. Whalelm: Finding structure and information in sperm whale vocalizations and behavior with machine learning. *bioRxiv*, 2024b. doi: 10.1101/2024.10.31.621071. URL `https://www.biorxiv.org/content/early/2024/11/11/2024.10.31.621071`.

Toshitaka N Suzuki, David Wheatcroft, and Michael Griesser. The syntax–semantics interface in animal vocal communication. *Philosophical Transactions of the Royal Society B*, 375(1789): 20180405, 2020.

Peter Tyack. Differential response of humpback whales, megaptera novaeangliae, to playback of song or social sounds. *Behavioral Ecology and Sociobiology*, 13:49–55, 1983.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In Alan W. Black, editor, *The 9th ISCA Speech Synthesis Workshop, SSW 2016, Sunnyvale, CA, USA, September 13-15, 2016*, page 125. ISCA, 2016. URL `https://www.isca-archive.org/ssw_2016/vandenoord16_ssw.html`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

William A Watkins and William E Schevill. Sperm whale codas. *The Journal of the Acoustical Society of America*, 62(6):1485–1490, 1977.

Linda Weilgart and Hal Whitehead. Coda communication by sperm whales (physeter macrocephalus) off the galapagos islands. *Canadian Journal of Zoology*, 71(4):744–752, 1993.

Linda Weilgart and Hal Whitehead. Group-specific dialects and geographical variation in coda repertoire in south pacific sperm whales. *Behavioral Ecology and Sociobiology*, 40:277–285, 1997.

Hal Whitehead. *Sperm whales: social evolution in the ocean*. University of Chicago press, 2003.

Hal Whitehead and Linda Weilgart. Patterns of visually observable behaviour and vocalizations in groups of female sperm whales. *Behaviour*, pages 275–296, 1991.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

Jiangjian Xie, Yujie Zhong, Junguo Zhang, Shuo Liu, Changqing Ding, and Andreas Triantafyllopoulos. A review of automatic recognition technology for bird vocalizations in the deep learning era. *Ecological Informatics*, 73:101927, 2023.

Lian Xu, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Deep learning for marine species recognition. *Handbook of deep learning applications*, pages 129–145, 2019.

Yossi Yovel and Oded Rechavi. Ai and the doctor dolittle challenge. *Current Biology*, 33(15): R783–R787, 2023.

Lue Zhang, Hai-Ning Huang, Li Yin, Bao-Qi Li, Di Wu, Hao-Ran Liu, Xi-Feng Li, and Yong-Le Xie. Dolphin vocal sound generation via deep wavegan. *Journal of Electronic Science and Technology*, 20(3):100171, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Fréchet Audio Distance evaluations can be found in Section 4.1. Perceptual studies with expert marine biologists can be found in Section 4.2. Experiments on downstream classification are reported in Section 4.3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: N/A.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details on model training are in Appendix E. Details on the experimental setup are in Appendix E.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the model weights, and the code used for training and evaluating the model upon publication. We will attempt to release as much of the data as possible, however we note that the data was collected by a large collaboration of marine biologists over several decades, and so we cannot commit to getting their consent to publish by publication time.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, in Appendices E and E.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All figures depict error bars or standard error figures, with the exception of Figures 3 and 12 and table 2. For Figures 3 and 12 repeating the experiments with multiple seeds would have been computationally prohibitive due to the amount of categories in each experiment. For Table 2, this is a minor, supplementary used to justify an experimental design choice, rather than a main component of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resources needed to train and finetune the model are reported in Appendix E.2. The resources needed to conduct the experiments (after training the model) are reported in Appendix E.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Appendix A

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: N/A.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We adhere to the license of all external code we use. The data used was either collected by us (the authors), or is used in accordance to the license of the dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide as much data as appropriate and possible of the new datasets we used, without de-anonymizing ourselves. The training method and model are properly documented in Appendix E.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: In Appendix E.5.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The perceptual listening study presented no risk to participants. Participants were asked to evaluate audio similar to those they regularly analyze in their professional work. No personal data was collected beyond coarse background questions and the assessment itself (see Appendix E.3), and no compensation was provided. The task falls within the participants' scope of their normal research activities and expertise.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: N/A.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Broader impacts

Our work on modeling sperm whale communication has potential implications for both scientific understanding and conservation efforts. Historically, advances in understanding cetacean communication have played crucial roles in conservation—notably, the discovery of humpback whale song by

Payne and McVay [1971] contributed significantly to public awareness and the subsequent “Save the Whales” movement [Feldman, 2021, Campagna and Guevara, 2022, Comuzzo, 2023]. While we maintain that sperm whales deserve protection regardless of our ability to understand their communication, we recognize that deeper scientific understanding often catalyzes public engagement with conservation efforts.

Our model’s capabilities might naturally suggest applications in behavioral experiments through playback studies. This is particularly tempting given that sperm whales often produce codas simultaneously—a behavior that our bidirectional model could theoretically capture by conditioning on one whale’s clicks while generating the overlapping clicks of another. However, we strongly caution against such applications at this stage. Without a deeper understanding of coda semantics and functionality, playback experiments using synthetic vocalizations could have unintended and potentially harmful consequences for these social marine mammals. Instead, we propose that this work demonstrates the potential of learning from passive acoustic observation—studying these remarkable animals through careful listening rather than active intervention. With this approach, this work could potentially play a role in assisting efforts to reinforce existing protections or create new legal protections for whales [Rodríguez-Garavito et al., 2025].

As noted in Section 5, our methodological framework could extend beyond sperm whales, potentially benefiting research on other marine mammals and, more broadly, any species that communicates acoustically. This scalability is particularly relevant as biodiversity monitoring becomes increasingly critical in the face of environmental changes. However, our experience underscores the importance of deep collaboration with domain experts—the success of this work relied on guidance from marine biologists and acousticians with decades of experience studying sperm whales. We encourage future work in this direction to similarly prioritize partnerships with species-specific domain experts, as their insights are crucial for both model development and responsible deployment.

B Preliminaries on sperm whale codas

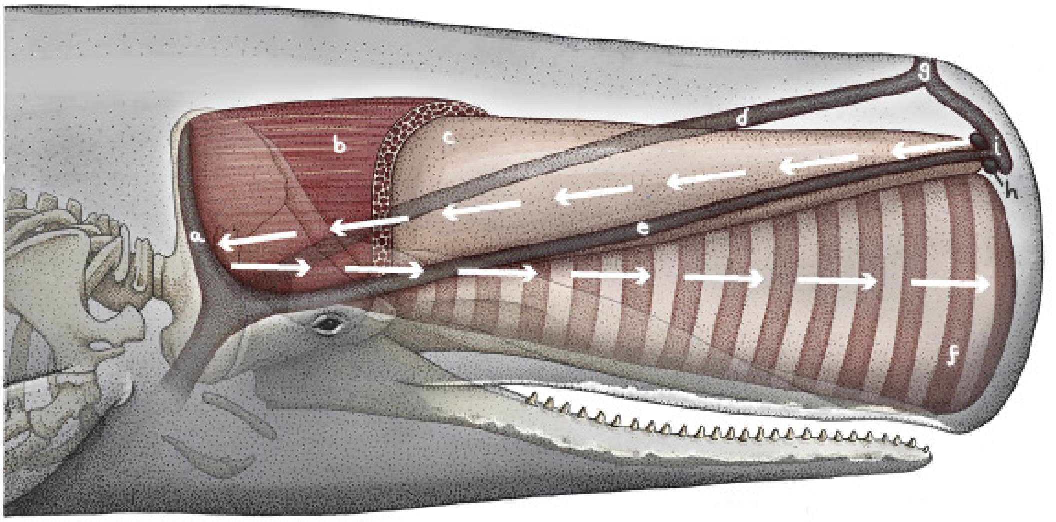

Figure 6: The sperm whale head contains the spermaceti organ (c), a cavity filled with almost 2kL of wax-like liquid, and the junk compartment (f), comprising a series of wafer-like bodies believed to act as acoustic lenses. The spermaceti organ and junk act as two connected tubes, forming a bent, conical horn of about 10m in length and 0.8m aperture in large mature males. The sound emitted by the phonic lips (i) in the front of the head is focused by traveling through the bent horn, producing a flat wavefront at the exit surface. Reproduced with permission (Andreas et al. 2022, © Alex Boersma 2021).

Sperm whales have evolved remarkable acoustic capabilities. Figure 6 illustrates the key anatomical and acoustic aspects of these capabilities, which form the basis for their complex communication system.

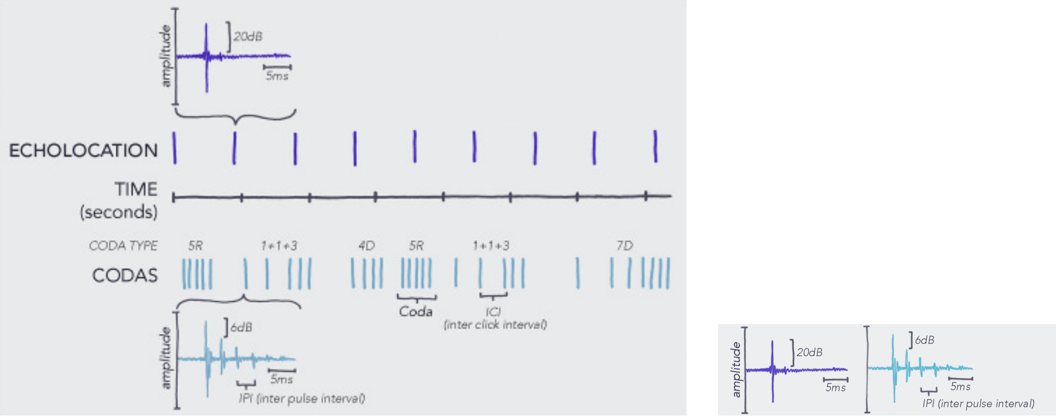

Figure 7: **Left:** Typical temporal structure of sperm whale echolocation and coda clicks. Echolocation signals are produced with consistent inter-click intervals (of approximately 0.4s) while coda clicks are arranged in stereotypical sequences called "codas" lasting less than 2s. Codas are characterized by the different number of constituent clicks and the intervals between them (called inter-click intervals). Codas are typically produced in multi-party exchanges that can last from about 10s to over half an hour. Each click, in turn, presents itself as a sequence of equally spaced pulses, with inter-pulse interval of an order of 3–4ms in an adult female, which is the result of the sound reflecting within the spermaceti organ. **Right:** Typical structure of echolocation (dark blue, left) and coda clicks (light blue, right). When observed as a waveform zoomed into a single click the type types of clicks differ observably in structure. There is far greater attenuation between the first and second pulse of an echolocation click, then the coda clicks. Further, the coda clicks resonate more in the nose of the sperm whale creating additional pulses after the first one for coda clicks. Reproduced with permission (Andreas et al. 2022, © Alex Boersma 2021).

Sperm whales live in a multileveled social structure with female lines living together in 'units' with stable membership [Whitehead, 2003]. Early acoustic research proposed that codas might serve as individual signatures [Watkins and Schevill, 1977], but subsequent studies instead suggested that different coda types may have distinct functions [Antunes et al., 2011], and that variation of coda usage among units suggested a function in unit-level social identity [Moore et al., 1993, Weilgart and Whitehead, 1993, 1997]. Even when living in the same waters, whales from different social units will only associate with units which share a similar repertoire of codas. This social segregation based on acoustic similarity was used to delineate the highest level of social organization which structures their populations, the vocal clan; and that codas function as symbolic markers of these cultural groups [Rendell and Whitehead, 2003, Gero et al., 2016a, Hersh et al., 2022]. Importantly, there is good evidence that these distinct dialects of codas, with variation in number of clicks, as well as rhythm and tempo, are the product of social learning, and not genetically inherited [Cantor and Whitehead, 2015, Rendell et al., 2012].

The clicks produced by sperm whale can be generally classified as either *ecolocation* or *coda* clicks. Echolocation clicks which function in navigation and hunting in the dark, wherein echoes of the clicks return and are interpreted by the whales in the darkness of the deep ocean, much like bats in the night sky. Conversely, coda clicks are thought to function in communication between whales and are exchanged between whales or groups of whales in social contexts at the onset of dives, during shallow dives near the surface, and during large social interactions.

Echolocation signals are produced with consistent inter-click intervals while coda clicks are arranged in stereotyped, rhythmic sequences called "codas" lasting less than 2 seconds. Codas are characterized by the different number of constituent clicks and the intervals between them (called inter-click intervals or ICIs). Rhythmic patterns and tempo of clicks define coda 'types', which are often given descriptive names. For example, a 1+1+3 coda is click-pause-click-pause-click-click-click (Figure 7).

C A listener’s guide to codas

Building on findings from our Expert Perceptual Study (Section 4.2), we present a short guide detailing perceivable similarities and differences between natural and synthetic codas. We note that, unlike the Expert Study, this guide was developed by the authors under no time constraints, and with unrestricted aid of spectrograms and familiarity with model internals. This Listener’s Guide to Codas is structured as a unifying Theme, followed by four Variations each isolating a specific cue.[4] For a broad-audience listener’s guide to whale (albeit humpback) vocalizations, see Payne [1970].

Theme. Synthetic codas generated by WhAM can be evaluated both visually and acoustically, using the same structural cues that characterize authentic sperm whale clicks. Each natural coda click typically consists of a sequence of equally spaced pulses, with an inter-pulse interval (IPI) of approximately 3–4 ms in adult females. This is a consequence of internal reflections within the spermaceti organ.

Variation A: Balance. DC offset (a shift of the waveform away from being centered at zero) does sometimes occur when recording sperm whales in the wild, particularly when using handheld recording systems which run off batteries and a constant DC voltage. It is often consistent, while synthetically generated clips will have quite a “wavy” offset. It is however interesting to note that WhAM picked up on this feature of the authentic waveforms. In addition, sperm whales do not vary the amplitude dramatically between sequential clicks within codas, while WhAM generated codas sometimes do.

Variation B: Frequency. Using a spectrogram, one can see that the frequency content of synthetic clicks is more uniform. In Figure 10, one can compare the shape of spectrograms for otherwise relatively similar clicks and note that the shape is more uniform and consistent both in time and frequency for synthetic clicks (bottom as strong orange rectangles) compared to authentic clicks which trail off both as frequency increases and across time (top, more pointed at top, with far less yellow above 10kHz, and rough along the right side). In addition, you can also observe the variation in amplitude across synthetic clicks in the waveform, but a consistent amplitude in the waveform of the authentic clicks (as described above).

Variation C: Structure. Authentic sperm whale clicks, especially coda clicks, contain the typical multiplulsed structure with a detectable inter-pulse-interval created by the head of the sperm whale and the path of the sound as it is generated (outlined above). Synthetic clicks often did not have a realistic structure either by having no pulsed structure (center of Figure 9) or an exaggerated one (right of Figure 9). While some of these effects occur in authentic clicks based on the angle of recording relative to the axis of the body of the whale making the sounds, the synthetic clicks rarely had realistic structure within clicks.

Variation D: Listening alone. Taken together, these waveform- and spectrogram-based cues are sometimes audible *even without visual aids*. A trained ear could identify synthetic codas based on subtle irregularities in amplitude, spectral consistency, and the absence of realistic multipulsed structure.

---

[4] With apologies to Britten [1945].

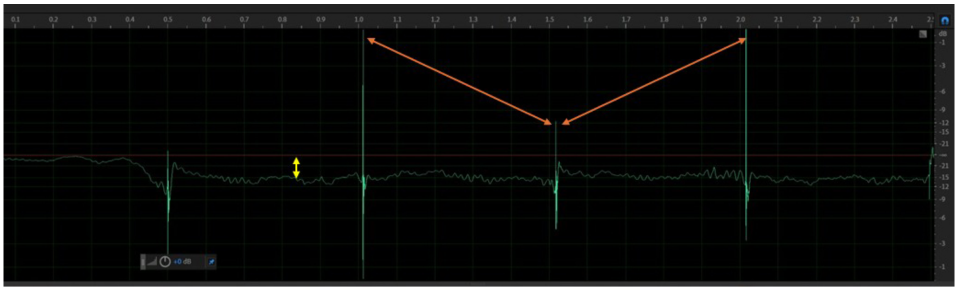

Figure 8: Sample of a synthetic coda generated by WHAM with the variable DC offset dissimilar to natural recordings (yellow arrow) and the dramatic variation in amplitude between sequential clicks (orange arrows).

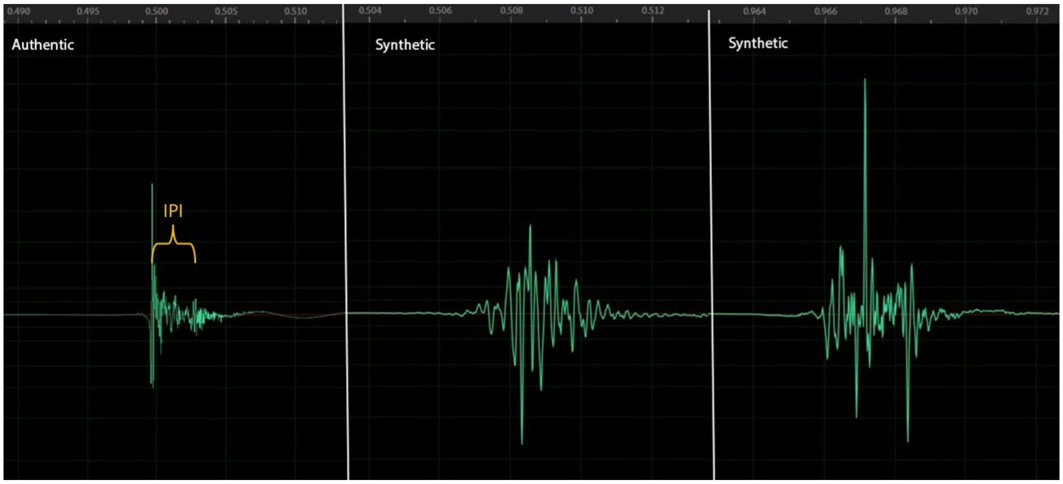

Figure 9: Pulse structure of authentic (real) and synthetic clicks.

D Additional experiments

D.1 FAD Embedding Selection

The Fr'echet Audio Distance (FAD) measures similarity between audio datasets using embeddings to map the audio into a feature space. The choice of embedding is crucial, as different embeddings capture different aspects of the signal. For analyzing sperm whale codas, we sought an embedding that prioritizes the temporal patterns critical to coda structure over background noise. This appendix describes the calibration experiment we conducted to select the most suitable embedding for our FAD analysis.

Let the coda recordings in DSWP+CETI be denoted by $\{x_1, ..., x_n\}$, we:

1. Created denoised versions $\{\hat{x}_1, ..., \hat{x}_n\}$ as detailed in Appendix E.1
2. Isolated the removed noise components $\{x_1 - \hat{x}_1, ..., x_n - \hat{x}_n\}$
3. For each candidate embedding f_i, compared:
    - $d_1^i =$ FAD score between codas and their denoised versions:
    - $d_2^i =$ FAD score between codas and their noise components:

We evaluated five audio embeddings VGGISH Gemmeke et al. [2017b], Hershey et al. [2017], Encodec-embd Défossez et al. [2023], LAION CLAP Music, LAION CLAP Audio Wu* et al. [2023], Chen et al. [2022], and BirdNET Kahl et al. [2021] using the Fréchet Audio Distance

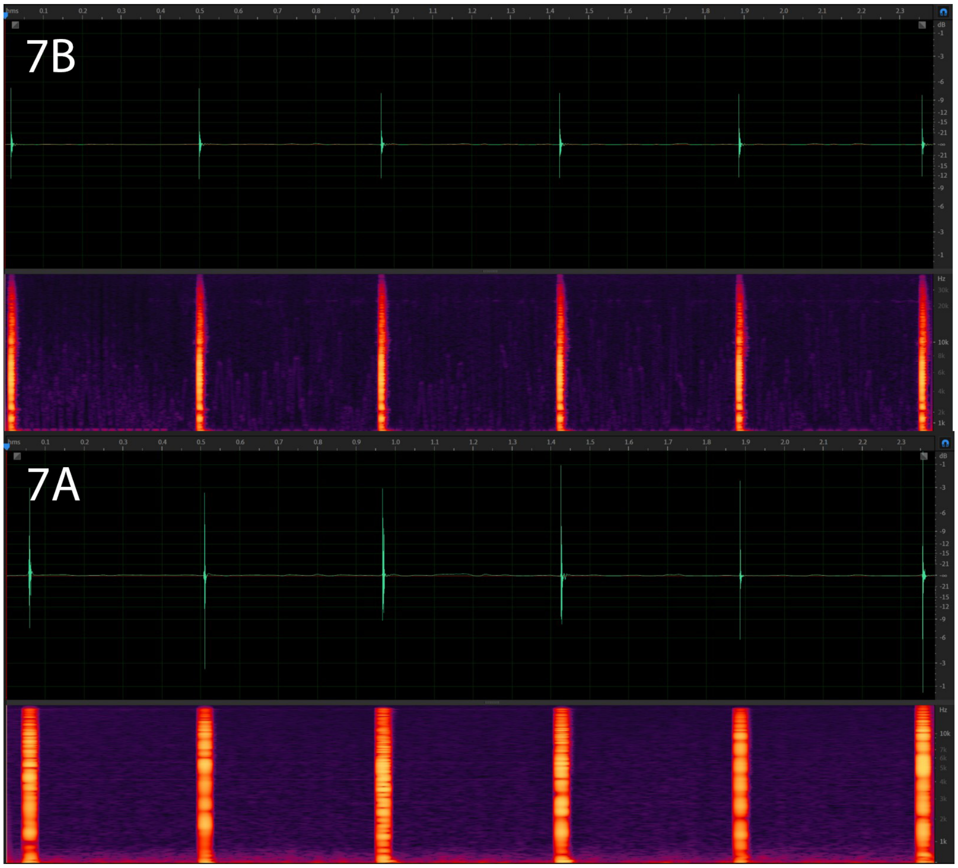

Figure 10: Waveform and spectrograms of both authentic clicks recorded from wild sperm whales from clip 7B (top) and from WhAM generated synthetic clicks from clip 7A (bottom). Here “7” is because this formed the seventh pair in the expert perceptual study (Part 1).

implementation of Gui et al. [2024]. The ratio d_2^i/d_1^i indicates how much more weight embedding i gives to background noise versus temporal structure. A larger ratio indicates stronger emphasis on temporal patterns and better suitability for the quantitative assessment of audio translation experiment. Table 2 shows these ratios for each embedding.

Table 2: Comparison of Audio Embeddings for Temporal Structure Sensitivity

EMBEDDING	d_1 (CODA VS. DENOISED)	d_2 (CODA VS. NOISE)	d_2/d_1
VGGISH	2.0844	1.5027	0.7209
ENCODEC-EMBD	25.9716	3.156	0.1215
LAION CLAP MUSIC	0.1483	0.1080	0.7282
LAION CLAP AUDIO	0.1144	0.1098	0.9597
BIRDNET	16.7817	22.4761	1.3393

Based on these results, we selected BirdNET for our main FAD experiments, as it maximized the ratio of distances between raw-to-noise over raw-to-signal.

D.2 Downstream Task Ablation Study

To evaluate the contributions of different components in WhAM, we conduct an ablation study by progressively removing elements and assessing performance across the same set of downstream tasks. The results are presented in Figure 11.

No finetuning. We test the effect of skipping domain-adaptation (step (b) in Figure 1), or skipping finetuning of VampNet altogether (steps b,c) in Figure 1). For all tasks except Social Unit classification, removing species-specific finetuning or domain adaptation does not have a significant impact on the accuracy. This indicates that the inclusion of these steps in WhAM does not significantly degrade the performance on most downstream tasks.

Tokenizer-only. We falsify the hypothesis that the neural audio codec is sufficient for capturing semantic properties in the audio by testing downstream classification directly on the acoustic tokens (Figure 2), without embedding them through the MATM. This causes a statistically significant performance drop, particularly in Social Unit classification (-10.9 points, from 70.5% ± 0.7% to 59.6% ± 2.0%)

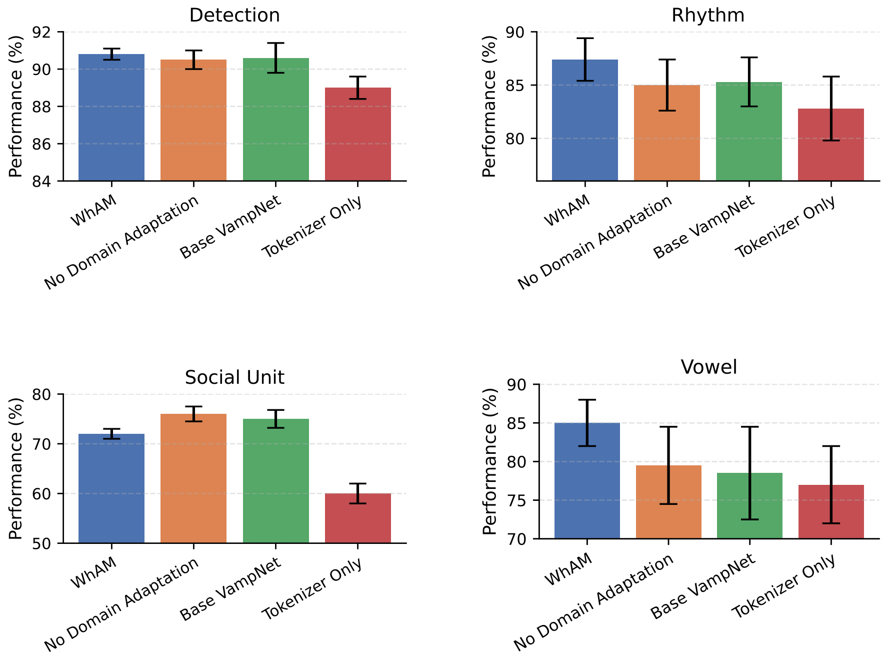

Figure 11: Classification accuracies (%) resulting from using the output of different WhAM components on downstream tasks. Each classifier head was trained using three different random seeds, with mean±stderr reported.

D.3 Fréchet Audio Distance Ablation Study

To complement the ablation study of Appendix D.2, the experiments detailed in section 4.1 were repeated four times with marine mammal sounds. First, using only the tokenizer. Second, training the model with only **Domain Adaptation** (DA, step (c) in Figure 1), skipping the **Species Specific Fine-Tuning** step (SSFA, step (c) in Figure 1). Third, training only with **Domain Adaptation**. And finally, using the full version of WhAM. These results (Figure 12) show that, as expected, fine-tuning WhAM on sperm whale data results in outputs that are more similar to sperm whale vocalizations.

D.4 Tokenizer Reconstruction Loss Study

WhAM uses the Descript Audio Codec (DAC) as its tokenizer [Kumar et al., 2023].DAC is tailored towards speech, music, and environmental sounds. To test possible degradation in encoding sperm whale coda audio, we conducted the following experiment.

Let each individual coda recording in the DSWP+CETI datasets be denoted by $x_1, x_2 ... x_n$

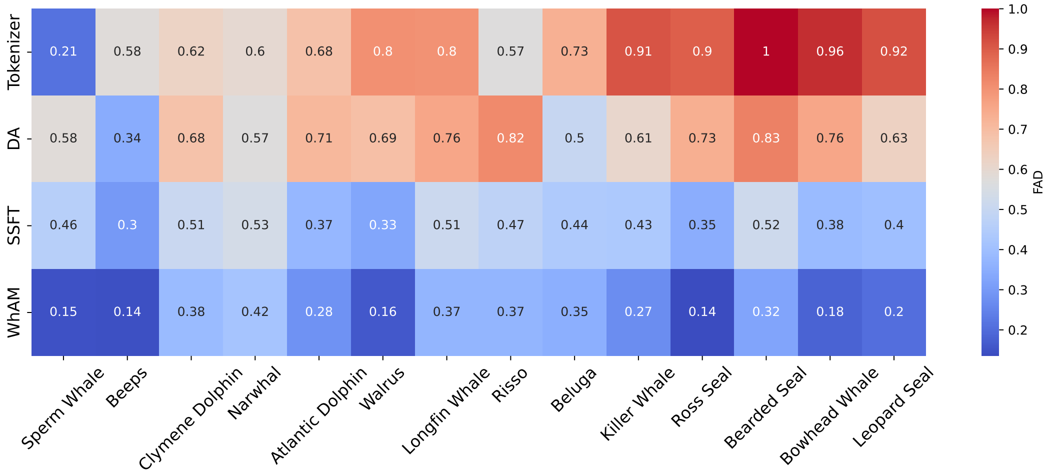

Figure 12: The effect of ablating components of the model on FAD

1. For each recording x_i, a reconstructed version, $\hat{x}_i$ was created by passing x_i into the tokenizer to generate tokens, then passing the tokens through the decoder to recover the audio recording.
2. Each x_i and $\hat{x}_i$ was then sliced into chunks of length C to calculate their respective short term fourier transforms. The transform is represented by the arrays $\{\vec{x}_{i,1}, \vec{x}_{i,2}...\vec{x}_{i,m}\}$ and $\{\vec{\hat{x}}_{i,1}, \vec{\hat{x}}_{i,2}...\vec{\hat{x}}_{i,m}\}$. Each $\vec{x}_{i,j}$ represents the magnitude all frequencies over the jth chunk of the recording x_i
3. The mean reconstruction accuracy, denoted by E is now given by taking the average normalized error between all $\vec{x}_{i,j}$ and $\vec{\hat{x}}_{i,j}$ using the formula: $\frac{1}{n \times m} \sum_{i,j} (\vec{x}_{i,j} - \vec{\hat{x}}_{i,j})^2 / (\vec{x}_{i,j})^2$

The mean error is shown in fig. 13 for a chunk size of 2.27 ms and 22.7 ms. With a smaller chunk size, E indicates which pitches the tokenizer accurately reconstructs and which pitches it does not. Using a larger chunk size, E gives an indication of what type of general noise patterns the tokenizer fails to include in its reconstruction.

As can be seen with a chunk size of 22.7 ms, the error is wave-like. Since sinusoidal waves in the frequency domain correspond to impulses in the time domain, this suggests a tendency to misrepresent impulse-like sounds in the time domain. On the other hand, using a chunk size of 2.27 ms provides an indication as to what pitches the tokenizer prioritizes. The spikes in the bands 1–6 kHz and 8–10 kHz suggest that, in general, the tokenizer tends to perform relatively poorly in those frequencies. However, this degradation is not severe enough to prevent our model from generating natural-sounding codas (Section 4.2), nor from its embeddings to “capture” vowels (Section 4.3).

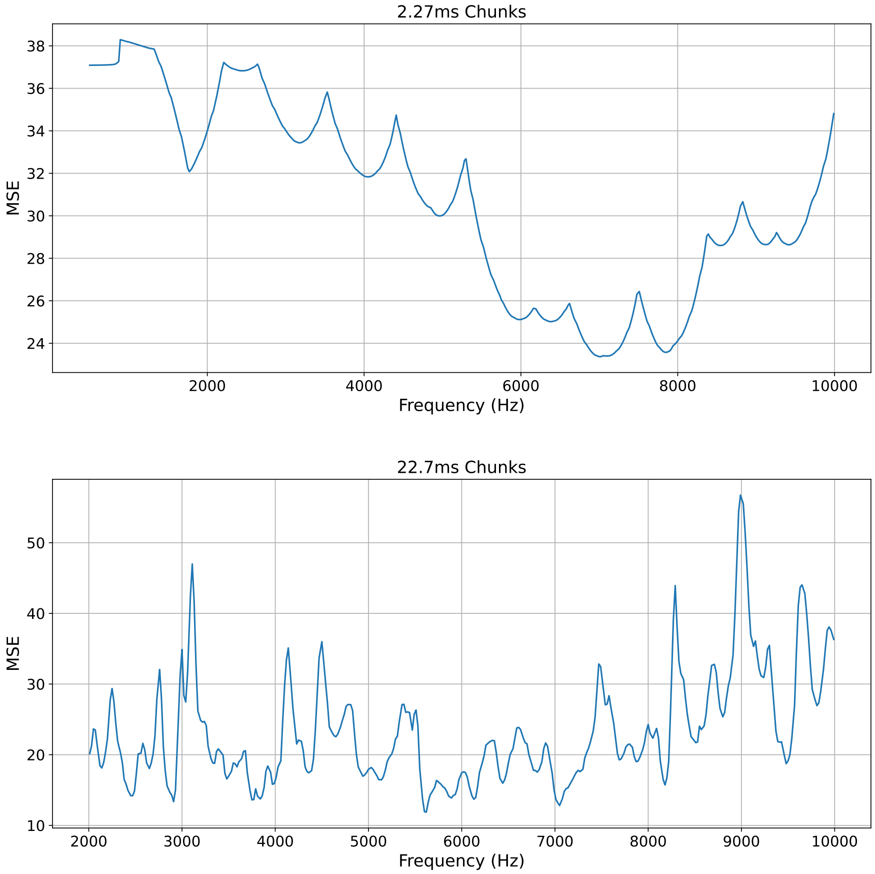

Figure 13: Tokenizer reconstruction loss study. Normalized mean squared error (y-axis) by frequency (x-axis).

E Methodology Details

E.1 Data

FSD. The FSD50k dataset includes 3,159 audio recordings labeled with the "animal" tag, amounting to a total duration of 7 hours and 45 minutes. Noisy segments were retained to preserve real-world variability in training data.

AudioSet. The AudioSet dataset was used to supplement training with additional animal vocalizations. It contains 5h8m hours of audio.

BirdSet. Consists of 6,800 total hours of recordings containing bird vocalizations [Rauch et al., 2025]. Due to space constraints and to avoid training WhAM on audio that did not contain any vocalizations, only a subset of the entire dataset was used, containing a total of 110 hours of data.

WMMS. The Watkins Marine Mammal Sound Database consists of raw, unlabeled audio recordings. The dataset contains a total of 4 hours and 8 minutes of audio. Each recording was segmented into 10-second snippets for training. No additional denoising was applied. The dataset contained vocalizations from the following mammals (names as listed on the WMMS website):

Atlantic Spotted Dolphin
Bearded Seal
Beluga (White Whale)
Bottlenose Dolphin
Boutu (Amazon River Dolphin)
Bowhead Whale

Clymene Dolphin
Commerson's Dolphin
Common Dolphin
Dall's Porpoise
Dusky Dolphin
False Killer Whale
Fin, Finback Whale
Finless Porpoise
Fraser's Dolphin
Grampus (Risso's Dolphin)
Gray Seal
Gray Whale
Harbor Porpoise
Harbour Seal
Harp Seal
Heaviside's Dolphin
Hooded Seal
Humpback Whale
Irrawaddy Dolphin
Juan Fernandez Fur Seal
Killer Whale
Leopard Seal
Long-Beaked (Pacific) Common Dolphin
Long-Finned Pilot Whale
Melon-Headed Whale
Minke Whale
Narwhal
New Zealand Fur Seal
Northern Right Whale
Pantropical Spotted Dolphin
Ribbon Seal
Ringed Seal
Ross Seal
Rough-Toothed Dolphin
Sea Otter
Short-Finned (Pacific) Pilot Whale
Southern Right Whale
Sperm Whale
Spinner Dolphin
Spotted Seal
Steller Sea Lion
Striped Dolphin
Tucuxi Dolphin
Walrus
Weddell Seal
West Indian Manatee
White-beaked Dolphin
White-sided Dolphin

DSWP. The dataset consists of codas collected between 2005–2018 in a 2000km^2 area off the coast of Dominica. Codas were recorded using various recording systems including far-field boat-based hydrophones and animal-borne tags. Recording setups were as follows:

2005: A Fostex VF-160 multitrack recorder (44.1kHz sampling rate) and a custom built towed hydrophone (Benthos AQ-4 elements, frequency response: 0.1–30kHz) with a filter box with high-pass filters up to 1 kHz resulting in a recording chain with a flat frequency response across a minimum of 2–20kHz.

2006: No recordings during this short season.

2007,2009,2011: A Zoom H4 portable field recorder (48kHz sampling rate) and a Cetacean Research Technology C55 hydrophone (frequency response: 0.02–44kHz) and no filters.

2008,2010,2012,2015: A custom-built towed hydrophone (Benthos AQ-4 elements, frequency response: 0.1–30kHz) with a filter box with high-pass filters up to 1 kHz resulting in a recording chain with a flat frequency response across a minimum of 2–20 kHz. This was connected to a computer based recording system as a part of the International Fund for Animal Welfare's (IFAW) LOGGER software package (48kHz sampling rate) or PAMGUARD (minimum 48 kHz sampling rate). In addition, recordings were also made through the deployment of animal-borne sound and movement tags (DTag generation 3, Johnson and Tyack 2003).

CETI. All systems were sampling above 96kHz with a 16bit resolution with a minimum flat (±2dB) frequency response within 1–45kHz.

The DSWP and CETI dataset contain background noise such as water sounds. To improve model performance, we denoise datasets before training on the model. A noise profile of each recording in the frequency domain was generated by sampling sections which did not contain codas. Then, we perform spectral subtraction to remove noise in the frequency domain, and transform back to the time domain of the audio signal.

All audio samples were downsampled to 16 kHz and normalized to have zero mean and unit variance when passed into VampNet.

Table 3: Quantitative Assessment Data Summary

FULL NAME	SHORTENED NAME	NUM. SAMPLES
ATLANTIC DOLPHIN	A. DOLPHIN	58
BEARDED SEAL	B. SEAL	37
BOWHEAD WHALE	B. WHALE	60
BELUGA WHALE, WHITE WHALE	BELUGA	50
WALRUS	WALRUS	38
CLYMENE DOLPHIN	C. DOLPHIN	63
NARWHAL	NARWHAL	50
LEOPARD SEAL	L. SEAL	10
LONG-FINNED WHALE	L. WHALE	10
KILLER WHALE (ORCA)	ORCA	35
ROSS SEAL	ROSS SEAL	50
RISSO'S DOLPHIN	RISSO	67

Table 4: Prompt settings for each input type.

INPUT	PERIODIC PROMPT	ONSET MASK WIDTH	NUM. OF STEPS	TYPICAL MASS	SAMPLE CUTOFF
CODAS	12	21	50	0.102	0.17
BEEPS	12	21	50	0.102	0.17
A. DOLPHIN	16	5	74	0.15	0.39
B. SEAL	7	1	70	0.15	0.44
B. WHALE	7	1	70	0.15	0.44
BELUGA	13	13	85	0.15	0.39
WALRUS	18	1	107	0.15	0.33
C.DOLPHINE	12	14	72	0.15	0.25
NARWHAL	6	4	39	0.15	0.21
L. SEAL	6	4	46	0.15	0.39
L. WHALE	15	19	57	0.15	0.42
ORCA	13	2	46	0.15	0.39
ROSS SEAL	18	3	66	0.15	0.49
RISSO	13	13	85	0.15	0.39

Note that, as with any self-supervised training setup that relies on random masking, the effective number of unique training examples far exceeds the raw audio hours. In our case: First, each 2-second audio snippet becomes a 14×120 token array. Columns correspond to time steps, and rows represent acoustic granularity During training, entire columns (i.e., time steps) are masked at random; with 120 columns, there are 2^{120} possible masking patterns per snippet. So, for example a 20 hour dataset yields 36,000 snippets, which result in $\approx 10^{40}$ possible masked training inputs.

E.1.1 Generating data for Sections 4.1 and 4.2

Three different input sources were used to generate samples for both the **Quantitative Assessment of Audio Translation** and the **Expert Perceptual Evaluation**. The prompt settings for each input type are summarized in Table 4.

Watkins Marine Mammals. Eleven species were selected from the "Best of Watkins Marine Mammals" dataset. Due to variations in vocalization characteristics and recording conditions, prompt settings were manually optimized for each species. These species and prompt settings can be found in Table 4.

Digital "beeps". Five digital beep sequences were generated. Each snippet was initialized as a zero-filled array at a 44.1 kHz sample rate. Clicks were simulated by selecting random indices and setting them to a peak amplitude of 1. To ensure realistic timing and rhythm, real coda sequences were prepended to each generated sample before synthesis. These prepended codas were then removed after generation.

E.2 Model Training

The model training procedure consisted of two phases: domain adaptation and species-specific fine-tuning.

Acoustic Tokenizer Settings. Discrete token vocabulary size (Σ) = 1024. Frequency of Input Audio N_{sam} = 16kHz. Tokenizer input length N_{sec} = 10.

Domain Adaptation. In the first phase, the model was pretrained on a mixture of general animal vocalizations, including data from FSD and AudioSet. This step aimed to establish a broad understanding of bioacoustic patterns. The model was trained for 500,000 iterations using the AdamW optimizer with a learning rate of 0.0001. A batch size of 6 was used, and gradient clipping was applied to stabilize training. The model took 123 hours to train using an AWS EC2 g5.2xlarge instance (NVIDIA A10 GPU, 8 vCPUs, 32 GB of memory).

Species-Specific Fine-Tuning. Following domain adaptation, the model was fine-tuned on whale-specific data from DSWP+CETI to adapt its representations to sperm whale vocalizations. The fine-tuning process used the same optimizer and learning rate as the pretraining phase and a batch size of 6. Training continued for another 500,000 iterations. This took 39 hours to run using an AWS EC2 g5.2xlarge.

E.3 Computational costs

All experiments were run on an AWS EC2 g5.2xlarge (NVIDIA A10 GPU, 8 vCPUs, 32 GB of memory). A full run of FAD experiments took 3 hours with a full version of Vampnet, and 1.5 hours using a Tokenizer-only model, therefore Section 4.1 and appendix D.3 took approximately 7.5 hours in total. Appendix D.1 took approximately 1.5 hours. For downstream classification, training the linear probe took at most 5.5 hours; thus, Section 4.3 and appendix D.2 took about 16.5 hours in total.

E.4 Utility of Embeddings for Downstream Tasks

Model Details. We run a forward pass through WhAM and AVES to obtain embeddings from the audio. Both WhAM and AVES output varying embeddings over time, so we average the embeddings over time to obtain 1 unified embedding for 1 audio snippet. After the embedding is obtained, we attach a two-layer feed-forward neural network as a classifier. The network consists of a fully connected layer that projects the embedding into a 128-dimensional hidden layer, followed by a ReLU activation. A second fully connected layer then generates class probabilities.

We evaluate embeddings from WhAM and AVES, comparing their performance against a random embedding baseline as well as a majority baseline classifier.

Training Data. For downstream task evaluation, we leveraged annotations in the DSWP+CETI datasets. Using human-annotated timestamps, we identified and extracted audio segments containing codas, each spanning 1–2 seconds. Each coda was labeled for one of the following classification tasks:

- **Coda Detection**: Determine whether a given audio snippet contains a whale coda.
- **Rhythm Type Classification**: Classify codas according to their rhythmic patterns. For this task, we choose to include samples whose rhythm types are among the 5 most common, because the remaining ones appear too infrequently for classifiers to be accurate.
- **Social Unit Classification**: Identify the social unit associated with each coda.
- **Vowel Classification**: Detect vowel-like patterns within whale vocalizations.

Table 5 summarizes dataset sizes for each task.

Training Process. We split the dataset into 80% training and 20% testing, using stratified sampling of labels to ensure consistent label distribution. The embedding model is frozen, and only the classifier parameters are trained. Training is performed on an NVIDIA A10G GPU for 10 epochs, using a

Table 5: Dataset sizes for downstream classification tasks.

TASK	NUMBER OF SAMPLES
CODA DETECTION	3,100
RHYTHM TYPE CLASSIFICATION	916
SOCIAL UNIT CLASSIFICATION	2,659
VOWEL CLASSIFICATION	486

learning rate of 10^{-4} and a batch size of 32. Model checkpoints are saved at each epoch, and the best-performing model is selected based on test set performance.

E.5 Expert Perceptual Evaluation

Five domain experts in sperm whale bioacoustics participated in the evaluation. Given the highly specialized nature of sperm whale vocalization analysis, the pool of qualified experts with years of direct experience analyzing and annotating these vocalizations is notably small. All participants were recruited from an established research collaboration studying cetacean communication, and each had at least three years of experience working with sperm whale codas. No compensation was given to participants.

The evaluation was conducted via Google Form. The form began with the following introduction:

Welcome

Thank you for participating in this study. Your expertise in analyzing sperm whale vocalizations is invaluable for evaluating our model.
The study consists of four parts, to be completed in order. A final section includes three short questions about your background.

Technical Setup

- Download and extract the `listener_evaluation.zip` file from a provided link
- Use headphones for all listening tasks
- Complete the experiment in a quiet environment
- You can take breaks between sections as needed

If you encounter any technical difficulties or have questions about the procedure, please contact [omitted].

Participant Identification

Name (used for tracking responses only): ______________________

E.5.1 Audio-Only Two-Alternative Forced Choice (2AFC)

Listeners were presented with 30 pairs of codas. Each pair contained an original, denoised coda and a model-generated counterpart. Participants were asked to identify which sample was the original and which was generated.

Task Instructions

In this section, you will listen to pairs of codas. For each pair, one is a natural recording and one is synthetic. Please indicate which one you believe is synthetic.
The audio files are located in the ***section1/*** folder. Each pair consists of two files:

- *1a.wav* + *1b.wav*
- *2a.wav* + *2b.wav*
- etc.

Please listen to each file **at most three times**. Base your decision only on the provided audio. Do not visualize the audio.

E.5.2 Mixed Two-Alternative Forced Choice (2AFC)

Listeners were presented with 25 individual samples: 10 real codas, 5 generated from real codas, 5 generated from walrus vocalizations, and 5 generated from digital beeps. Each listener classified each sample as either real or generated.

Task Instructions

In this section, you will listen to individual codas and classify each as either natural or synthetic.
The audio files are located in the ***section2*** folder:

- *1.wav*
- *2.wav*
- etc.

Please listen to each file at most three times. **Base your decision only on the provided audio. Do not visualize the audio.**

E.5.3 Visualized Two-Alternative Forced Choice (2AFC)

This experiment was identical to the **Audio-Only 2AFC** condition, except participants were allowed to inspect the spectrograms of each recording using their preferred software before making their decision. Marine biologists preferred Adobe Auditions, while underwater acoustics experts used Matlab.

Task Instructions

Once again, you will listen to pairs of codas (a.wav and b.wav). For each pair, one is a natural recording and one is synthetic. Please indicate which one you believe is synthetic.
The audio files are located in the ***section3/*** folder. Each pair consists of two files:

- *1a.wav* + *1b.wav*
- *2a.wav* + *2b.wav*
- etc.

Please listen to each file at most three times. **You may now visualize the audio using any software you are familiar with.**
What software will you use to visualize the audio? ______________________

E.5.4 Qualitative Assessment

Task Instructions

For this final section, please first listen to the reference synthetic codas provided in the section4 folder. These examples were chosen to represent typical outputs of our model. Then, based on these examples and your experience with all parts of the experiment, please answer the following questions
What characteristics of natural codas are well represented in the synthetic ones?
What characteristics of natural codas are missing or different in the synthetic ones?
Did you observe any patterns in the synthetic codas that do not occur in natural ones?
When **only listening** to the audio (sections 1 and 2), what helped you distinguish between natural and synthetic codas?
When **visualizing** the audio pairs (section 3), what helped you distinguish between natural and synthetic codas?

E.5.5 Background Information

Task Instructions

To help contextualize the evaluations, please tell us about your experience working with sperm whale codas.
How many years have you spent professionally analyzing sperm whale codas (e.g., in research, conservation, or educational contexts)?
What types of coda work have you performed?

- *Recording of codas in the field*
- *Development of recording methods for codas*
- *Manual detection, classification or annotation of codas*
- *Development of automatic detection, classification or annotation systems*
- *Meta-analysis (e.g. methodology development, literature review)*
- *Other...*

In what contexts have you worked with coda recordings?

- *Academic research*
- *Conservation work*
- *Industry/commercial projects*
- *Educational/training contexts*
- *Government/regulatory work*

What is your primary field of expertise?