UC Berkeley
UC Berkeley Previously Published Works

Title
Approaching an unknown communication system by latent space exploration and causal inference

Permalink
https://escholarship.org/uc/item/57p3g1kc

Journal
Royal Society Open Science, 13(8)

Authors
Beguš, Gašper
Leban, Andrej
Gero, Shane

Publication Date
2026-08-26

DOI
10.1098/rsos.250829

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/rsos

CC BY

Research

Cite this article: Beguš G, Leban A, Gero S. 2026 Approaching an unknown communication system by latent space exploration and causal inference. *R. Soc. Open Sci.* **13**: 250829.
https://doi.org/10.1098/rsos.250829

Received: 29 April 2025
Accepted: 18 June 2026

Subject Category:
Computer Science and Artificial Intelligence

Subject Areas:
artificial intelligence, behaviour

Keywords:
latent space, generative AI, generative adversarial networks, causal inference, phonetics, acoustics, linguistics

Author for correspondence:
Gašper Beguš
e-mail: begus@berkeley.edu

† These authors contributed equally to the study.

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8598324.

THE ROYAL SOCIETY PUBLISHING

Approaching an unknown communication system by latent space exploration and causal inference

Gašper Beguš[1,2,†], Andrej Leban[2,3,†] and Shane Gero[2,4,5]

[1]Department of Linguistics, University of California, Berkeley, Berkeley, CA, USA
[2]Project CETI, New York, NY, USA and Roseau, Dominica
[3]Department of Statistics, University of Michigan, Ann Arbor, MI, USA
[4]Department of Biology, Carleton University, Ottawa, Canada
[5]The Dominica Sperm Whale Project, Roseau, Dominica

GB, 0000-0002-6459-0551; AL, 0000-0003-0617-6843; SG, 0000-0001-6854-044X

We propose a methodology for discovering meaningful properties in data without ground truth by combining manipulation of the latent variables of generative models to extreme values with causal inference in an approach we call *causal disentanglement with extreme values* (CDEV). Using it, we investigate what properties the model encodes as meaningful when trained on raw audio of sperm whale (*Physeter macrocephalus*) communication. The method suggests that the model considers the number of clicks in a sequence, the regularity of their timing, as well as audio properties such as the spectral mean and the acoustic regularity of the sequences as the main components for generating believable and informative data. The first two are consistent with existing hypotheses, while the last two are proposed for the first time. We also argue that our models uncover rules that govern the structure of units in the communication system, suggesting the methodology as a viable strategy for approaching unknown data.

1. Introduction

How do we approach a communication system for which we not only do not understand what is meaningful, but also are unsure about how to go about testing for meaning? One such case is the communication system of sperm whales (*Physeter macrocephalus*), whose vocalizations likely carry meaning because they appear to be socially learned [1] and are produced in duet-like exchanges or group choruses before foraging and while socializing [2], but never while alone [3]. While evidence supports the use of codas as identity signals [4], there is little that is currently known about

© 2026 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

what individual utterances mean, or even concrete evidence about what kind of properties of the communication system could serve to carry meaning.

In this work, we propose an approach that aims to provide insight into unknown communication systems by using an expressive generative model as the learning mechanism. The network is trained with two objectives: (i) imitation of data and (ii) encoding of information (figure 1a). We then combine latent space exploration with methods borrowed from causal inference into a methodology we call *causal disentanglement with extreme values* (CDEV; figure 1a). This approach helps us test which observable properties the network has learned to encode relevant information for the synthetic vocalizations. If sperm whales encode information in their vocalizations and our model can learn to imitate those vocalizations well, the encoding in our models can likely reveal what might be meaningful in the sperm whale communication system. We argue that our technique reveals both properties that have been posited as meaningful by human researchers as well as novel properties that have, so far, not yet been hypothesized as such. The former can thus serve to bolster the credibility of the latter.

The advantage of the proposed approach is that the discovery of such properties uses as few task-specific assumptions as possible, treating the network as a black-box learning unit. The fiwGAN architecture [5] used has been repeatedly shown to learn a variety of meaningful properties about human language from audio recordings of human speech. Networks trained on raw speech data are shown to learn to associate lexical items with code values and sublexical structure with individual bits [5,6], uncovering known linguistic rules [7–9]. In other words, the method is verified on the human communication system—language—where ground truth is available. We observe the same ability to infer the hidden structure of the data in our results (§3.2). Moreover, learning is performed on raw audio without information-losing (or biasing) transformations.

To our knowledge, this is the first attempt to model sperm whale communication with deep learning on raw acoustic data. In this context, we use the network (in conjunction with CDEV) as an extraordinarily flexible and information-preserving tool for decomposing the data into meaningful, observable properties. Furthermore, the proposed approach, where individual units in the inputs of deep neural networks are manipulated to extreme values, and their effects on observable properties of the outputs are estimated via causal inference methods, could also be applied to other architectures and data with few, if any, modifications.

Sperm whales communicate in short (less than two seconds), socially learned series of *clicks* called *codas* that marine biologists have grouped into several coda *types*. Traditionally, codas were classified primarily based on the number of clicks in a coda and their relative timing [10,11]. For example, a five-click coda with the first two clicks further apart and the last three clicks closer together is classified as the *1+1+3* coda. A coda with nine clicks is classified as a *9R* type. These two codas are given as examples in figure 1b. Temporal features continue to play an important role in more recent analysis of codas that operate with more complex timing features [12]. Sperm whales exchange codas in complex dialogues and so far no understanding has been reached about the referential meaning of the codas. The goal of this paper is to provide a methodology for finding potentially meaningful properties that could serve as the building blocks of the communication system.

In the dataset used for training the models in this work, there are about 22 different coda types [13]. The actual distribution of coda types is highly asymmetric—the two most common codas comprise 65% of all coda vocalizations recorded [4]. Owing to this rarity of some types, we further limit the training set used in this work to the five most common coda types (*1+1+3*, *5R1*, *4R2*, *5R2* and *5R3*). Since we are dealing with a communication system, a significant portion of the data could not be used owing to too strong a presence of whale dialogue [2]. The residual effects of this are dealt with by our detection algorithms (§2.2). Restricting the number of coda types also allows us to test whether the network can predict unobserved coda types. In total, 2209 distinct training samples were used.

2. Methods

2.1. The architecture

The network architecture used in this work is fiwGAN [5], an InfoGAN [14] adaptation of the WaveGAN [15] model; GAN stands for generative adversarial network. In short, the input is partitioned into an *incompressible noise* X, which is sampled uniformly from $[-1, 1]$, and an additional *featural encoding* (or

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

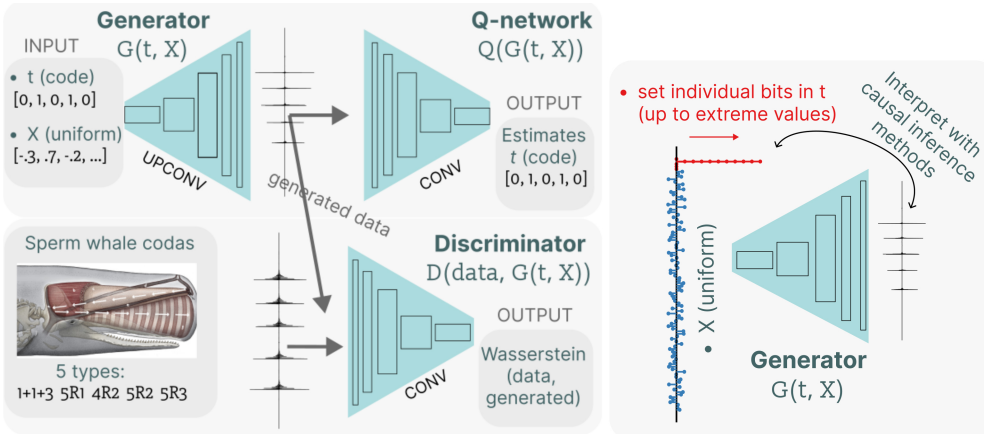


Figure 1. (a) Model and approach overview. Left: overview of the fiwGAN architecture. Right: overview of the CDEV method. (b) Examples of real codas. Left: type *1+1+3* with 5 clicks; right: type *9R* with 9 clicks (*x*-axis in seconds). The *9R* type was never part of the training data, yet our network learns to generate codas that resemble this type (§3.2). Whale drawing CC BY 4.0 ©Alex Boersma.

code) t, which is sampled as Bernoulli variables during training [5]. We thus define $[-1, 1]$ as the range of values the network encounters during training. The goal of the additional Q-network during training is to correctly determine the value of the encoding t while only having access to the generated data; a loss based on the mutual information between the two is backpropagated to the generator otherwise. This ensures consistency of output across similar values of the encoding vector (figure 1a) and encourages disentanglement in a game that mimics communicative intent—learning by imitation in a fully unsupervised manner. The loss optimized by the fiwGAN architecture [5] is thus the following:

$$\min_{G,Q} \max_{D} V_{\text{IWGAN}}(D, G, Q) = V_{\text{WGAN}}(D, G) - \lambda L_I(G, Q), \tag{2.1}$$

where V_{WGAN} refers to the loss of the WaveGAN architecture [15]. L_I is the variational lower bound of the mutual information between the encoding and the generated data $I(t; G(X, t))$ and λ is a hyperparameter (see [14] for more details).

2.2. The causal disentanglement with extreme values methodology

Beguš [9] proposes a technique where setting individual latent variables in a GAN to extreme values outside of those seen in training helps interpolate their linguistic meaning [7,8].

In this work, we conversely treat the generator as an experiment in the vein of causal inference and test for *observable* properties of the data that have (or can be) hypothesized as meaningful. When generating output samples, the incompressible noise (X) entries are sampled randomly, while the featural encoding t is set manually to a desired value. Since the consistency of output with regard to the encoding is only enforced in a loose way via the training loss equation (2.1), this relationship often only becomes readily apparent when setting the numerical values outside the bounds seen in training. There, the primary associated effect begins to dominate, which we denote as *disentanglement*. We then apply statistical

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

estimators on the candidate property samples derived from the raw generated outputs to determine whether there exists a statistically consistent relationship between the encoding and the outcomes. This procedure gives rise to a methodology we call CDEV (figure 1a). We note that we are dealing with *causality under the model* [16]: while the procedure is a proper causal experiment, it happens inside the model's world. We thus do not claim (here and throughout) ecological or behavioural causation in whales.

The space available for the featural encodings is limited; hence, finding the real-world attributes that map almost one-to-one with the encodings suggests that the generator considers them very important to generating convincing outputs, in which it is being checked by the discriminator with access to real data. We limit our featural encoding space to five bits (implying $2^5 = 32$ classes) for the five coda types present in the data to allow the model to capture compositionality. However, as we demonstrate in figure 2d, the method is robust to the number of bits chosen, as well as model training specifics. Similarly, on language data, the architecture uncovers meaningful properties even when there is a mismatch between the number of true meaningful classes and the size of the binary code [5]. Therefore, any intuition about the desired size of the encoding acts only as a rough prior. Additionally, the uniqueness of the encoding matches and the process of disentanglement is illustrated by using state-of-the-art interpretability approaches trained on unrelated models that regress the observed outcomes to the generator's inputs directly (§2.2.4).

For the first two observables considered—the number and regularity of clicks (§§3.1 and 3.2)—we use an algorithmic *click detector*. As we move towards more extreme encoding values, the output becomes progressively noisier (supplementary material, §2.1). Therefore, the detector incorporates minimal spacing thresholds obtained from real-world codas and signal filtering methods. In concert with this, we chose the upper limit of the range tested ($t = 12.5$) so that the quantities measured by algorithms remained consistent despite increasingly noisy output. The latter might, in part, be owing to the network training not having converged completely; since the goal is to discover what observable properties are encoded and not high-fidelity whale communication generation, this is not a significant impediment. Furthermore, §2.1.1 in the supplementary material demonstrates that the presented conclusions do not depend on specific detector parameters.

Furthermore, figure 3b demonstrates that the effects are apparent at an even earlier point in training than presented. Likewise, the actual numeric magnitudes of the estimated effects are less important than discovering the existence of such consistent relationships. The fact that the disentanglement might not be possible for some observables where the detection is too noise-sensitive, together with the need for sufficient sample sizes, which necessitates algorithmic measurement being possible and (potentially) incurs a considerable computational cost, thus act as the main limitations on the proposed method besides the need to specify candidate properties to test for.

The acoustic observables were derived from the raw generated audio by applying a band-pass Butterworth filter permitting frequencies between 2 and 16 kHz (based on known spectral ranges of whale clicks) and then calculating the periodogram (computed with an added Hamming window) in the case of coda-level properties (presented in supplementary material, §3) and the spectrogram in the case of click-level properties. The latter was subsequently cut up into clicks using the click detector described above, and the quantities of interest, such as the per-click spectral mean and the per-click spectral standard deviation, obtained as the weighted statistics, with the spectrogram values serving as the weights—e.g. the weighted mean of the spectrogram slice frequencies, weighted by the values of the same slice. The results presented as the click-level spectral mean and regularity were then obtained as the average of the corresponding quantities across the clicks within the coda considered.

2.2.1. Continuous-treatment causal inference

The methodology used in this work is inspired by *continuous-treatment* causal inference from the Neyman–Rubin framework point of view [17]. In short, it deals with an experimental setup where the usual treatment assignment variable Z becomes a continuous variable T, often called a *dose* owing to it being common in the pharmacological context. Given the assumptions common to the discrete case, the analysis proceeds in much the same fashion, with the treatment dose being discretized at several levels t. In the case of observational studies, the *propensity score* $e(x)$ thus becomes a function of two variables $r(t, x)$ [18]:

$$e(x) \to r(t, x) = ¶(T = t|X = x);$$

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

the most common quantity of interest is typically the difference in the conditional expectations of the outcome Y:

$$\mathbb{E}_{Y(t)}[Y(t)|r(t,X)=r]-\mathbb{E}_{Y(t')}[Y(t')|r(t',X)=r],$$

relative to some *baseline dose* t'. The choice of the latter is natural ($t'=0$) in pharmacology, for example, where no drug being administered carries a special meaning. In our context, the choice of such a baseline is not clear, as is addressed in this work by the use of the incremental causal effect (ICE) estimator (§2.2.3) and credit assignment by a surrogate model (§2.2.4). The ultimate goal is to obtain the full *outcome curve* for all t with regard to the t' chosen.

The extreme values of t allow us to isolate the effects of individual codes on generated data. Elsewhere, it has been shown that the desired effect is learned already within $[-1,1]$ [5,7,9], but extreme values isolate and amplify the effect. This justifies treating t as continuous despite being Bernoulli-distributed in training.

In terms of causal inference terminology, we have the following variables: the *incompressible noise* part of the output is treated as the 95-dimensional covariate vector X, while the *featural encoding* is considered to be the treatment t and is a five-dimensional vector. The observables considered, such as the number of clicks for a given X and t, are thus the outcome variable Y_i. Each output was generated by sampling $X \overset{\text{iid}}{\sim}_{1,\ldots,95} \sim \text{Unif}\,[-1,1]$. Such vectors can be treated as unique within the generated sample used for the estimation of a particular quantity and hence can be thought of as denoting a separate *unit* indexed by i (see supplementary material, §6, for further discussion). These were fed to the generator at each level of the treatment $t \in [-1,12.5]$ set at each of the five featural bits, meaning the covariates X were kept the same across treatment levels. The generated audio was then run through the appropriate detection algorithm, such as the click detector, giving us the outcome $Y_i(t)$. The total number of units was chosen as $N=2500$, for which we determined the estimates to have completely stabilized (see figure 3a for details). This means that we can reasonably claim to be performing a completely randomized experiment (CRE) where the outcome is observed at each treatment dose for each unit,

$$¶(T_i=t|X_i=x_i)=1,\ \forall i,$$

meaning that there is no *fundamental problem of causal inference* [19] at play here, simplifying the estimation.

2.2.2. The average treatment effect estimator

The first estimator we present is the *average treatment effect* (ATE) [17], applied separately at each treatment dose value t. Formally, we are interested in the effect relative to some *baseline dose* t':

$$\mathbb{E}[Y(t)]-\mathbb{E}[Y(t')]=\mathbb{E}_X[\mathbb{E}_Y[Y|X,T=t]]-\mathbb{E}_X[\mathbb{E}_Y[Y|X,T=t']]. \tag{2.2}$$

Since the units i are defined by their randomly drawn X, and we observe every $Y_i(t)$, this simply corresponds to the difference in sample averages for the chosen t,t':

$$\hat{\tau}(t) := \frac{1}{N}\sum_{i=1}^{N} Y_i(t)-\frac{1}{N}\sum_{i=1}^{N} Y_i(t'). \tag{2.3}$$

The treatments correspond to setting the values of single bits in the encoding while keeping the others at zero. This leaves us to consider what would be the most natural value for the baseline; the limits of the training range: −1, where the output coalesces out of pure noise (see supplementary material, figure 1) and +1, where the process of disentanglement begins, serve as logical choices and have been used in this work (e.g. in figure 2a).

2.2.3. The incremental causal effect estimator

The question of selecting the appropriate baseline dose is elegantly avoided by using the ICE proposed by Rothenhäusler & Yu [20]—the effect owing to an infinitesimal shift in dosage δ:

$$\tau_{\text{ICE}} := \mathbb{E}[Y(T+\delta)]-\mathbb{E}[Y(T)]. \tag{2.4}$$

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

Since we are dealing with a CRE (§2.2.1), the following holds:

$$\partial_t \mathbb{E}[Y|T=t, X=x] = \mathbb{E}[Y'(t)|T=t, X=x], \tag{2.5}$$

meaning that the left-hand side—the true ICE—is identifiable by the expectation of the derivative—the right-hand side. Its estimator ($\hat{\tau}_{\text{ICE}}$) is the usual sample mean of the numeric finite differences in the outcomes for single units. Given the expectation of the derivative curve, we can also define the *expected effect of an infinitesimal increase in the treatment* by averaging over the range of t:

$$\hat{\theta}_{fs} := \frac{1}{N_t}\sum_{i_t=1}^{N_t} \mathbb{E}_{Y'(t_{i_t})}[Y'(t_{i_t})|T=t_{i_t}, X=x_i] = \frac{1}{N_t}\sum_{i_t=1}^{N_t}\frac{1}{N}\sum_{i=1}^{N}\frac{\Delta Y_i(t_{i_t}, x_i)}{\Delta t_{i_t}}, \tag{2.6}$$

where N_t is the number of examined treatment levels with the corresponding t_{i_t} and N is the number of units with the corresponding x_i, each receiving each treatment level. In short, this is the *expected overall effect* if all the current treatments were infinitesimally shifted (in the positive direction) for the given sample. As a single metric, it can be interpreted as an analogue to the *average treatment effect* for binary treatments without the need for a baseline dose as a reference. We are using 22 steps in the range $[-1, 12.5]$, leading to an ICE discretization step size of 0.614.

2.2.4. Regressing the outcomes directly

While our usage of the ATE and ICE estimators bears some relation to explainability methods such as partial dependence plots [21] and local methods such as local interpretable model-agnostic explanations (LIME) [22], respectively, we can also somewhat mirror the latter in another way by attempting to estimate the effects on the outcomes with a surrogate model. In practice, this means we regress the *individual* observed outcomes $Y_i(t)$ against the full input vector for each unit in the sample at each level t, i.e. for bit 1: $x_i = [0, t, 0, 0, 0|\ X_i]$. The *outcome curve* in this setting corresponds to the *mean* of the individual inferred outcomes at each treatment level.

For this task, we have chosen boosted tree regression as implemented in the `lightgbm` package [23]. For our specific purpose, we intentionally use *no* sparsity-inducing regularization to prevent overfitting: we wish to check if the hypothesized encodings are consistently picked up by a completely unrelated non-parametric method that is (i) able to approximate any function arbitrarily closely with sufficient model complexity (ii) when unregularized, is prone to overfitting. Hence, we are looking for *consistency of explanation* across varying model complexity, *despite all odds*. In other words, we are testing whether the encodings uncovered by the other two estimators could be spurious. This consistency for the bits identified as carrying meaning is displayed in figure 11 in the supplementary material, where complex models tend to overfit when an uninformative bit is used, but, crucially, do not when using an informative bit.

Most of the hyperparameters except the main one—the number of leaves—were chosen by an early stopping of the training as determined by a separate validation set, composed of randomly sampled (in terms of X) 10% of the data; the mean squared error on this set for trees with different numbers of leaves serves as the final determinant of what we consider the optimal model. The (maximum allowed) number of leaves in the base learner thus serves as our measure of model complexity.

To explain the results and quantify the consistency of the explanations, we use feature importance values obtained from Shapley additive explanations (SHAP) [24], which uses a game-theoretic concept called the *Shapley value* to distribute attribution of the outcome to the input features. Note that while one could, in theory, use SHAP on the generative model itself, it is much more computationally efficient when used with trees [25]. As mentioned, we also wish to test our findings explicitly via an unrelated method as opposed to disassembling the generative network.

We re-emphasize that the goal here is the inference of the effect of particular parts of the input by way of letting an unrelated, expressive model 'test out the competing hypotheses' itself. Specifically, we would like to see such models of adequate but varying complexity *consistently* assign the credit for the outcome to the encodings suggested by the other methods despite not being prohibited from picking up spurious relationships. This differs from other uses of such methods in causal inference, where correctly accounting for unobserved outcomes takes precedence [26,27].

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

3. Results

We train a fiwGAN model (figure 1a), which features three deep convolutional neural networks: the generator, the discriminator and the Q-network. The generator learns to imitate training data and generate audio outputs that approximate whale codas, as it is competing with the discriminator, whose task is to disambiguate the synthetic outputs from genuine training data. In order to generate the synthetic codas, the generator takes as input a vector separated into two parts: a *code* part (denoted as t), which is designed to be informative, and an *incompressible noise* part, which we denote as X. The size of the code part used was five bits. At the same time, the Q-network attempts to infer the specific code t that had been used when generating the audio output, and thus exchanges information with the generator as the latter also gets penalized in the case of a wrong estimate. Consequently, the generator network needs to learn to both imitate a given communication system and transmit information, implying that the process can be conceptualized as informative imitation. As the codes during training are a randomly sampled Bernoulli vector (see §2.1), the process of associating informative properties that can be inferred from the raw output by the Q-network with a specific code value is a completely unsupervised one. It has been demonstrated that the fiwGAN architecture learns meaningful properties in human language where the ground truth exists [5–7,9,28,29]. It is thus reasonable to assume that the generator will learn meaningful properties in non-human communication as well.

To uncover which meaningful properties about the unknown communication system the generator has learned to use to transmit information, we use the CDEV technique: by setting latent bits in the code part of the input to extreme values and using causal inference techniques on a large sample of the generated outputs, we show a causal relationship between individual bits and properties of the generated data. This approach allows testing whether any property of the data is encoded as meaningful in the code part of the latent space. Below, we present results from the application of CDEV to the following properties of codas: the number of clicks (§3.1), click spacing and regularity (§3.2) and acoustic properties (§3.3).

The approach permits a wide array of possible estimators, of which we will present results for the ATE (see §2.2.2) [17], the ICE [20] (see §2.2.3), and results obtained from attribution techniques when an expressive model is used to regress the relationship between the code and the property of interest [24] (see §2.2.4).

3.1. The number of clicks

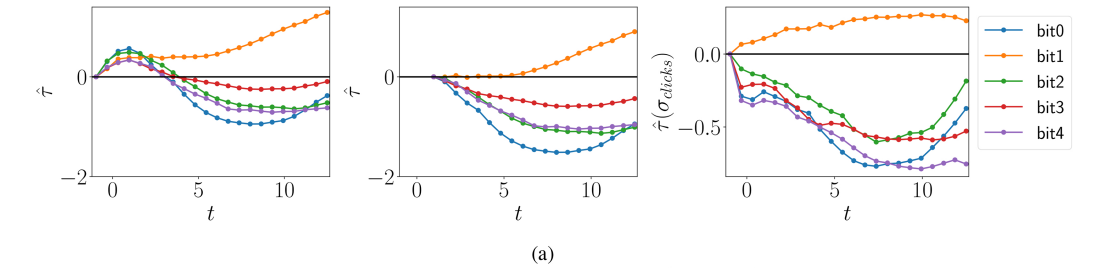
The first observable we are interested in is the *number of clicks* in the generated audio samples; more specifically, what the relationship is between the code t selected to generate the signal and the resulting number of clicks. The results for the ATE estimator are presented in figure 2a. As explained in §2.2.2, the estimator describes the net effect on the outcome of increasing the code t from some *baseline*.

In figure 2a, we observe a high degree of *entanglement* of the learned encodings within the range seen in training: $[-1, 1]$. However, when setting the treatment value above that range, all bits but *bit 1* stabilize at roughly a constant effect, which lends credence to interpreting this as a *process of disentanglement*: with higher values, the primary encoded effect in each bit starts to dominate, leading the other bits to lose their secondary effects on the number of clicks. Thus, we propose that *bit 1* is primarily associated with the number of clicks. In addition to the effect of bit 1, we also note a persistence of an effect in bit 3, which stabilizes at a different stationary value.

Examining the standard deviation in the number of clicks, as shown at the bottom of figure 2a, we again observe the same phenomenon of *disentanglement*, leading us to conclude that *bit 1* encodes the *range of clicks* output by the generator. Figure 2b presents the values of the ICE estimator for the number of clicks, corroborating the results obtained with the average treatment effect estimator as bit 1 exhibits a noticeable positive effect while the rest 'disentangle' to a minor, negative one.

Whether the effects can be attributed to the *incompressible noise* X part of the generator input can be answered by using SHAP [24], a powerful interpretability method, on boosted tree ensembles which were used to regress the outcome—the number of clicks in this case—against the whole of the input $[t; X]$ for each of the generated audio samples. In figure 2c, we present a *heatmap* plot which orders all the generated samples on the x-axis in terms of increasing treatment value t. The y-axis displays the features ordered in terms of their overall importance to the outcome as measured by SHAP. The plot above the central heatmap plot is the output of the model centred around the explanation's mean value. The

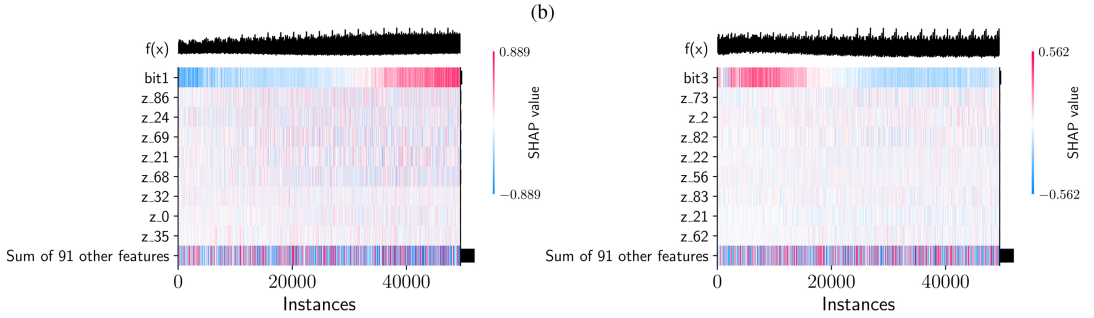
Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026



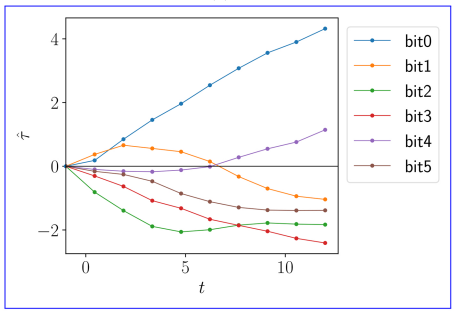
(a)

bit	0	1	2	3	4
$\hat{\theta}_{fs}$	-0.028	0.096	-0.039	-0.007	-0.046
$\hat{\theta}_{fs}\|t \geq 1$	-0.082	0.079	-0.087	-0.038	-0.083

(b)



(c)



(d)

Figure 2. The number of clicks. (a) The ATE for the number of clicks with the baseline chosen as (left) at $t' = -1$ and (middle) at $t' = 1$. Right: the ATE for the standard deviation in the number of clicks for the baseline at $t' = -1$. (b) The expected effect of an infinitesimal increase in the treatment on the number of clicks. (c) SHAP *heatmap* plots for the number of clicks for bit 1 (left) and bit 3 (right). The instances are ordered in terms of increasing t from left to right. (d) The ATE for the number of clicks with regard to the baseline at -1 obtained from *an independent* model that used an encoding size of **six** bits.

features for individual samples are coloured in terms of their contribution to the outcome. We compare the two bits that the other estimators have identified as having the largest effect (bits 1 and 3). For bits encoding the observable, we would expect to observe a *local consistency* in the assigned effects, since the units are ordered by increasing values of t. Conversely, we should observe much less consistently assigned effects for the other bits once the process of disentanglement has reached stationary values. We thus observe a greater degree of local consistency in the high-end of the treatment value (as evidenced by the direction and magnitude of the attributed effect for the individual units) for the bit encoding the property—bit 1—as opposed to bit 3, which has reached the stationary state of disentanglement. Furthermore, we observe that the incompressible noise part of the input plays a much smaller role (in terms of the magnitudes) in explaining the outcome, as well as having no local consistency with regard to the outcome.

In figure 2d, we display the ATE for the number of clicks for an independently trained (with different random seeds) model with a different latent space dimension (six versus five in figure 2a). The results show the same pattern of disentanglement with a single bit — bit 0 in this case — picking up the encoding.

This bit is a different bit from the one encoded by the model used for the results presented in §3.1, demonstrating that owing to the nature of the unsupervised learning of the encodings, the ordering of the

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

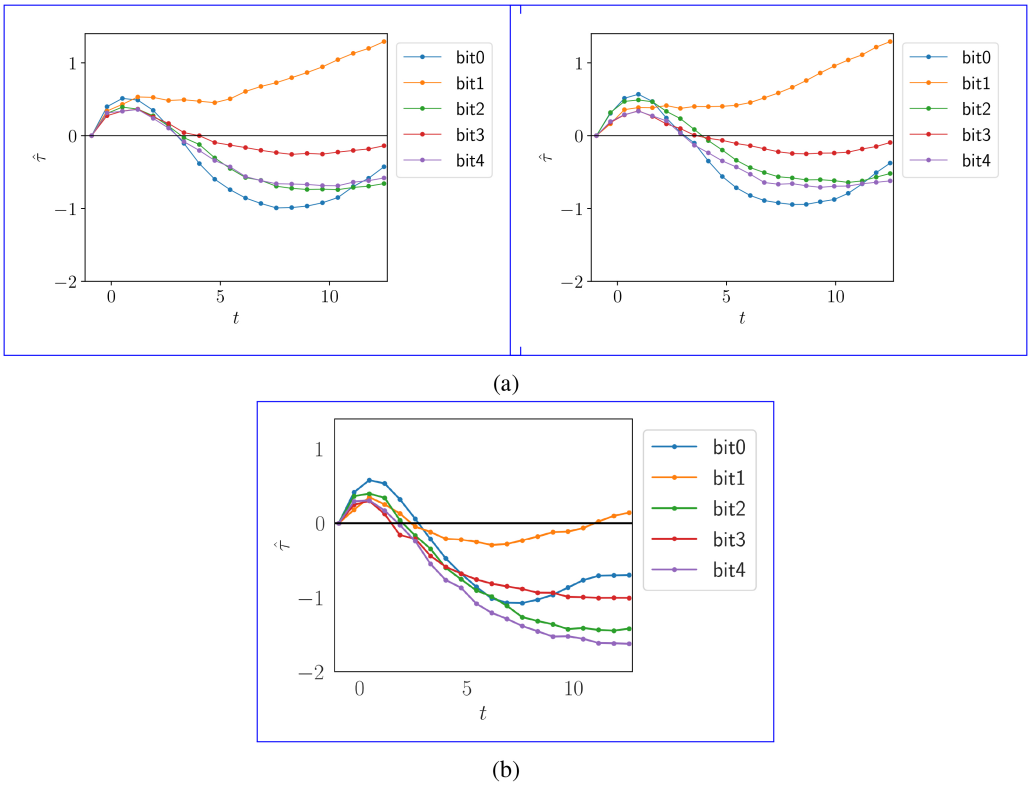


Figure 3. (a) Comparison between the ATE for the number of clicks with regard to the baseline at −1 from the data generated independently with differing random seeds, epoch 14 600 (main model). (b) Same, using epoch 12 500 of the model.

bits does not imply an ordering of the importance of properties. Interestingly, we again observe a degree of entanglement of the encoding with an additional bit, in this case, bit 4. The fact that this coordinate is bit 1 in the 5-bit model and bit 0 in the 6-bit model is consistent with the expected permutation symmetry of unsupervised codes; what matters is that a one-dimensional direction consistently captures the same property across independent runs.

In figure 3, we further present the ATEs for the number of clicks for two sets of synthetic data generated independently with differing random seeds. The variance of the response curve is negligible as the two match almost exactly, demonstrating stability of the results. Furthermore, the disentanglement happens before the epoch that we use for the majority of the results (epoch 14 600), with the same bit, which is shown at the bottom of figure 3 for a considerably earlier epoch. In other words, the results are not ephemeral or owing to under-convergence.

3.2. Click spacing and regularity

The next observables we are interested in are the *click spacing*, as measured by the mean *inter-click interval* (ICI) of a coda (i.e. the mean interval between the peaks of the click waveform), and the *coda regularity*, which we measure by the standard deviation of the ICIs within a coda. Since these quantities are not completely independent of the overall coda length, we stratify the results by the number of clicks observed.

The top row of figure 4a shows the ATE with respect to the baseline at −1 for the mean ICI for bits 1 and 2. We observe a consistent, monotonic effect in the value of bit 1, while bit 2 (and others not shown in the figure) does not display any such consistency across varying numbers of clicks. The ATEs for coda regularity are presented in the bottom row of figure 4a. We observe that bit 1 additionally encodes an increasing *coda regularity* (i.e. decreasing variance in the spacings between clicks) across all coda types (proxied here by the number of clicks).

Since bit 1 also encodes the number of clicks, this implies that the generator has learned to connect these properties: the codas with a higher number of clicks are more regular, with the clicks being more closely spaced together. This is especially poignant since the same property holds for actual whale codas.

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

Gero *et al.* [13] suggest that as coda length in clicks increases, mean ICI decreases to fit clicks within this duration limit, which appears to be the result of avoiding overlapping with the next coda within an exchange between whales. Furthermore, codas with more clicks are often more regular in their ICIs, regardless of their duration. The generator has thus inferred this connection *without the codas with a high (>5) number of clicks even being present in the dataset* owing to data quality limitations and encoded it in the limited space (five bits) it has reserved for encodings. In this, we again observe the propensity of GANs to discover hidden structures of the data and innovate in semantically meaningful ways.

As a summary of the overall ICEs, we can apply and sum the *sign* function on the expected effect of an infinitesimal increase in the treatment across codas with varying numbers of clicks, which we call the *sign score* and present it in figure 4c. This metric confirms that bit 1 is consistent in its encoding of both click spacing and regularity, regardless of the overall number of clicks in the generated codas. While bit 3 again remains slightly entangled for the latter quantity, its actual values of the expected effects were smaller than those of bit 1 (see supplementary material, §4.1).

As in the preceding section, we present in figure 4b the associated SHAP values for the mean-ICI and coda regularity for the best-fitting models for bits 1 and 2, restricted to codas with *five clicks*. In both properties, we again observe a corroboration of the effect suggested by the preceding estimators. We additionally observe greater *consistency* with regard to the expected value in local neighbourhoods for bit 1 but not for bit 2, as evidenced by the interchanging positive and negative contributions for units with the same treatment value. For the latter, a non-negligible amount of effect is also assigned to the noise part of the input.

3.3. Acoustic properties

We now apply the methodology to acoustic quantities, as captured by the spectra at the click or coda level. Prior to this work, little was known (or had been hypothesized) about the informational content of the acoustic properties of whale communication.

The first quantity we consider is the *mean spectral frequency*, where we first isolate the clicks with an algorithmic click detector (see supplementary material, §2.1), estimate the mean frequency of each periodogram and compute the average across all the per-click means in a particular generated coda. The corresponding ATE is shown on the left in figure 5a. We again observe the familiar pattern corresponding to CDEV—*bit 0* has a positive effect on the observable quantity, while all the rest converge to a smaller, stationary value. Following the outlined process, we thus hypothesize that this bit—bit 0—encodes the spectral mean.

We are also interested in *acoustic regularity*, which we measure by the standard deviation of the click spectral means *within* each generated coda. The ATE results are presented on the right in figure 5a. The results suggest an additional meaningful encoding: *bit 3* appears to encode the within-coda acoustic regularity of the output. In this case, it can be suspected that bits 1, 2 and 4 have not yet fully reached stationary values of their disentanglement.

Figure 5b presents the results for the expected effect of an infinitesimal increase on the click mean frequency and coda acoustic regularity, both for the overall range of treatments and for the range above the training range. The results corroborate the results obtained via the ATE estimator with the effect of bit 0 dominating the others in the case of mean frequency, and bit 3 in the case of spectral regularity. For the latter, bits 0 and 1, for which we have uncovered associated observable effects (hence they likely have an 'unintended' effect on the spectral mean standard deviation), are also relative outliers to the baseline of *disentanglement* evident in the remaining bits: 2 and 4.

For both properties, we present the results of the SHAP attribution for the best-fitting models in figure 5c. At the top, we again observe the expected patterns when comparing the SHAP values for the bit that picks up the encoding of the spectral mean—bit 0—and bit 4, which does not. As usual, the inferred effects move in opposite directions with regard to the expected (in terms of SHAP) outcome, with the effects becoming inconsistent in bit 4 once the process of *disentanglement* is complete. Conversely, the results for the spectral regularity shown at the bottom of figure 5c do not display the common inconsistency of the assigned effect for values with high treatment in bit 4 (i.e. the bit that does not pick up the encoding), meaning that the process of disentanglement has not reached the stationary value yet, illustrating what was suspected when discussing the ATE results above. Nonetheless, the other estimation approaches single out bit 3 as the encoding of the spectral regularity of a coda.

The potential for the spectral means of codas to be meaningful had not been hypothesized in previous research, but the hypothesis is not completely ungrounded. In a recent paper, Madsen *et al.* [30] suggest

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

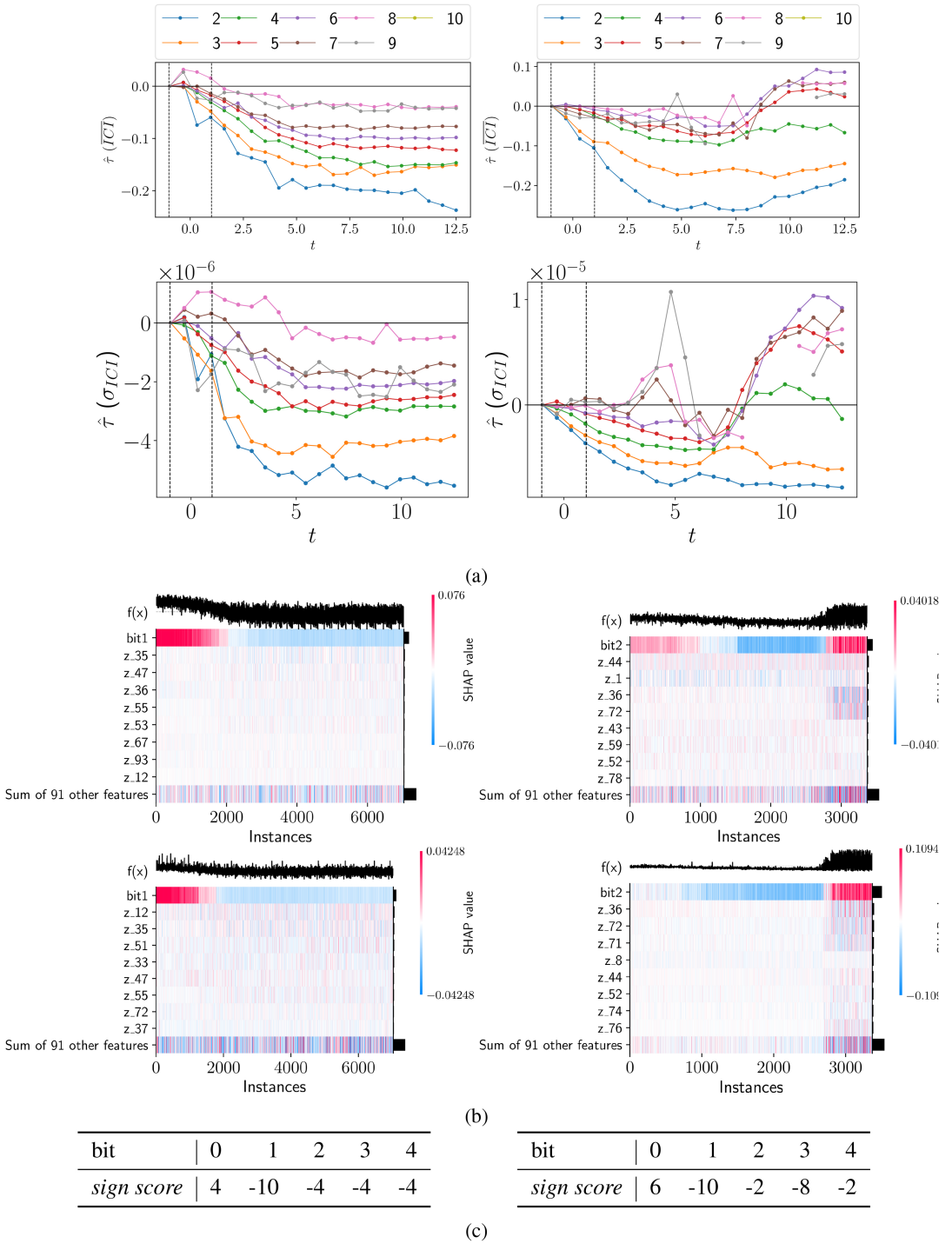
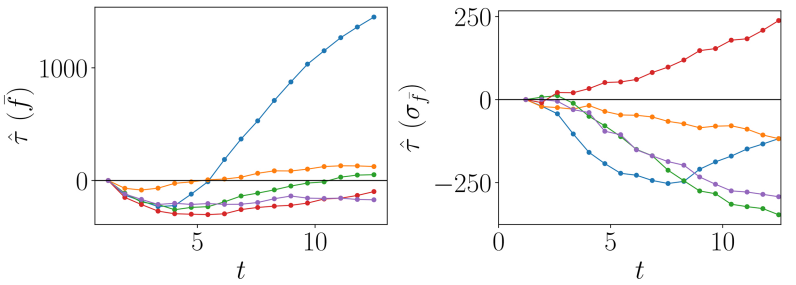


Figure 4. Click spacing and regularity. (a) The ATE results, stratified by the number of clicks. Left: bit 1; right: bit 2. Top: *mean ICI*; bottom: *coda regularity*. The training range is denoted by dashed lines, with the baseline set at $t' = -1$ for all. (b) SHAP heatmap plots for the above selection, restricted to codas with five clicks. (c) Overall *sign score* of an infinitesimal increase in the treatment. Left: the mean ICI; right: ICI standard deviation.

that odontocetes can vocalize in different registers and that their articulators are sufficiently flexible for such differences to be possible. Subsequent to the results presented here, the provided clues—*coda-level spectral mean and regularity*—were followed up on in Beguš *et al.* [31], which identified two discrete patterns in sperm whale communication—coda vowels and coda diphthongs. The first pattern corresponds to a spectral property analogous to those found in human *vowels* [32]: the codas contain examples that exclusively exhibit either a single or two clearly separated local spectral maxima at the level

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

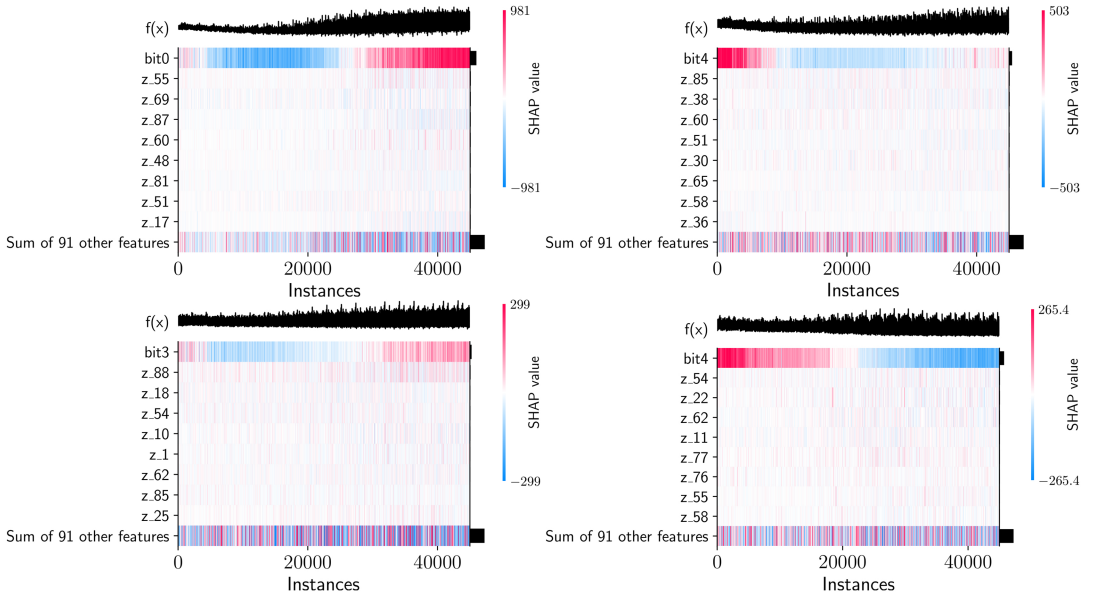


(a)

bit	$\hat{\theta}_{fs}(\bar{f})$	$\hat{\theta}_{fs}(\bar{f})\|t \geq 1$
0	62.15	128.58
1	-30.07	10.99
2	-44.55	4.56
3	-45.48	-8.71
4	-53.95	-15.18

bit	$\hat{\theta}_{fs}(\bar{\sigma}_f)$	$\hat{\theta}_{fs}(\bar{\sigma}_f)\|t \geq 1$
0	-21.17	-10.42
1	-21.99	-10.42
2	-35.85	-30.74
3	7.37	21.09
4	-34.01	-21.83

(b)



(c)

Figure 5. Acoustic properties. (a) The ATE results. Left: the ATE for the click-level *spectral mean* with the baseline at $t = +1$; right: the ATE for the click-level *spectral mean standard deviation* (acoustic regularity) with the same baseline. (b) The expected effect of an infinitesimal increase in the treatment on (left) average click mean frequency and (right) spectral mean standard deviation. (c) SHAP heatmap plots for the mean spectral frequency and acoustic regularity. Top—spectral mean: left: bit 0; right: bit 4. Bottom—spectral mean standard deviation: left: bit 3; right: bit 4.

of clicks, a pattern corresponding to *formants* found in human *a*- and *i*-vowels, respectively. This pattern corresponds to the variation in *spectral means* encoded by our model, corresponding to the encoding of the same coda type in terms of the number and regularity of clicks, but with different spectral means. These spectral properties are not crucially affected by movement of whales, and likely stem from the resonant frequencies of whales' distal air sacs [31].

The second identified pattern is a trajectory of the formant within a coda which corresponds to the pattern observed in human *diphthongs*: the formants are found to be either level, rising or falling between clicks within a coda, independently of the coda type. This corresponds to the separate encoding of the coda-level *spectral regularity* by our model, i.e. the codas with rising or falling patterns exhibit a much smaller spectral regularity. The fact that these two patterns were shown to occur in sperm whale dialogues independently of the traditionally recognized coda types (characterized by the number of clicks and their timing and encoded in *bit 1*) and each other (encoded in *bits 0* and *3*, respectively) also gives additional experimental validation of the *disentanglement* achieved with CDEV.

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

4. Discussion

The approach presented allows testing for candidate observable properties that a deep generative model encodes as meaningful as a way of gleaning information from data that are alien to us in the true sense of the word: the communication of sperm whales. In this, we leverage the power of information-theoretic GANs to encode semantically meaningful properties in a completely unsupervised fashion. Since the model is constrained in the number of such encodings it can learn, we can argue that it strongly prioritizes these factors in order to generate data.

To uncover these properties, we consider the trained model as an experiment and propose a method we call CDEV to discover the encodings. We present causal inference-inspired estimation methods that enable us to consistently pair up particular bits of the encoding with a physically observable property of the communication system. The agreement between the different methods gives further credence to the results.

Combining latent space exploration with the use of causal inference estimators on generated audio trained on data for which we have no ground truth is a novel approach, borrowing from both machine learning interpretability methods, as well as disentangled representation learning (DRL) [33].

In terms of the former, partial dependence plots [21] sample a section of the data randomly while examining the predicted outcome of a supervised model, which bears some similarities to our application of the ATE estimator (§3.1). Local surrogate models, such as LIME and SHAP [22,24], instead focus on explaining the local effect of perturbations via a surrogate model. While both methods bear some similarities to our application of the ICE estimator (§3.1), they are based on individual predictions.

In terms of the taxonomy of DRL presented in [34], our method is related to dimension-wise unsupervised DRL owing to the use of an InfoGAN [14] variant as the learning mechanism. We eschew any assumptions about the structure of the generating factors [35], as well as modifying the loss function of the architecture itself; both of these are easier to justify in applications where the ground truth exists, for example, in image generation [36,37]. Additionally, the application to audio data is a relatively unexplored case; these two factors prompt us to base our approach on an architecture [5] that has been shown to learn meaningful representations on the closest dataset where the ground truth is available—human speech—and 'perform the disentanglement *post hoc*' (i.e. from generated data) with the use of causal inference methods. Fully causal DRL methods [38] often use a structural causal model as a prior [39], which is, again, hard to justify for an unknown communication system. As an example, Bose *et al.* [40,41] use causal inference to quantify implicit causal relationships *between* latent space variables, while our use is based on effect estimation in service of uncovering encoded properties. As an example of the latter, Chockler *et al.* [42] use causal methods to interpret convolutional neural networks in a supervised classification setting when the input is occluded.

We emphasize that, in line with impossibility results for unsupervised disentanglement [43], we do not claim that individual code dimensions are identifiable in a strict sense. Different random initializations or architectural choices can rotate or permute the latent coordinates without changing the modelled data distribution as observed in figure 2d. Thus, we interpret the findings as evidence that certain factors (e.g. number of clicks) reliably concentrate along a low-dimensional direction, rather than that a specific 'bit index' is uniquely meaningful.

With our setup, we confirm that the number of clicks, which is what the existing classification developed by marine biologists is primarily based upon, indeed seems to be a fundamental property of the communication system. This can be seen as a good grounding point with regard to the credibility of the approach. While generating innovative examples of codas not seen in the data, we discover that the network correctly associates synthetic codas with a high number of clicks with their increasing regularity by encoding both properties simultaneously. Thus, it correctly infers a property of unseen real-life codas, illustrating its ability to learn the hidden structure of the data.

Using our methodology, we also uncover that two acoustic properties might be meaningful in the system: (i) the coda spectral mean and (ii) the regularity in the spectra across the clicks within a coda (coda spectral regularity). These two properties have not been considered significant by previous research. This could serve as an indicator that the communication system is not merely a Morse-like code where only the number of clicks and the intervals between them are meaningful, but that whales actively control the acoustic properties of their vocalizations and encode meaning in those properties. Following on these clues, Beguš *et al.* [31] subsequently identified concrete patterns in sperm whale communication that correspond one-to-one to the encoded acoustic properties presented here. This paper thus represents one of the first cases showing how AI interpretability techniques can facilitate scientific discovery.

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

Finally, the methodology presented can be applied to any problem for which one would like to leverage the expressiveness of deep generative models to consistently test for which properties of the data are semantically meaningful, requiring only that the latents of the generative model are controllable and that the properties under consideration are observable.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. We provide the trained models used as the learning mechanism via GitHub at https://github.com/andleb/Approaching-an-unknown-communication-system/releases and Zenodo at https://zenodo.org/records/18273374. The models include the main model used to generate the results with an encoding space size of five bits, an earlier epoch of the same, as well as an independently trained model with an encoding space size of six bits that is used to demonstrate the reproducibility of the results. The raw whale recordings are not publicly available.

We also provide intermediate generated data needed to reproduce the final results presented in the paper. The accompanying code that reproduces the results is available on GitHub at https://github.com/andleb/Approaching-an-unknown-communication-system and is archived on Zenodo at https://zenodo.org/records/15015012. It contains the source code for the model used as the base learning mechanism, both the experiment data generating code and the intermediate results, and the analysis code needed to replicate the results presented in the paper. The `README.md` document located at the root provides instructions for reproducing the results, as well as further technical details, such as the compute resources used to train the model and produce the results.

Supplementary material is available online [44].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. G.B.: conceptualization, formal analysis, funding acquisition, methodology, project administration, resources, supervision, writing—original draft; A.L.: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft; S.G.: data curation, resources, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This study was funded by Project CETI via grants from Dalio Philanthropies and Ocean X; Sea Grape Foundation; Rosamund Zander/Hansjorg Wyss through The Audacious Project: a collaborative funding initiative housed at TED. Research was funded through a FNU fellowship for the Danish Council for Independent Research supplemented by a Sapere Aude Research Talent Award, a Carlsberg Foundation expedition grant, a grant from Focused on Nature, two Explorer Grants from the National Geographic Society, and supplementary grants from the Arizona Center for Nature Conservation and Quarters For Conservation all to S.G. 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 to Peter Madsen of Aarhus University; and small grants from the Dansk Akustisk Selskab, Oticon Foundation and the Dansk Tennis Fond to Pernille Tønnesen of Aarhus University. Through 2014–2019, we were grateful for collaborative access to DTAGs and use of custom analytical tag code from Mark Johnson, Peter Tyack and Peter Madsen.

Acknowledgements. Data were collected through fieldwork for The Dominica Sperm Whale Project undertaken through permits from the Fisheries Division of the Government of Dominica. A.L. would like to thank Prof. Peng Ding (UC Berkeley) for a constructive discussion.

References

1. Rendell LE, Mesnick SL, Dalebout ML, Burtenshaw J, Whitehead H. 2012 Can genetic differences explain vocal dialect variation in sperm whales, *Physeter macrocephalus*? *Behav. Genet.* **42**, 332–343. (doi:10.1007/s10519-011-9513-y)
2. Schulz T, Whitehead H, Gero S, Rendell L. 2008 Overlapping and matching of codas in vocal interactions between sperm whales: insights into communication function. *Anim. Behav.* **76**, 1977–1988. (doi:10.1016/j.anbehav.2008.07.032)
3. Whitehead H, Weilgart L. 1991 Patterns of visually observable behavior and vocalizations in groups of female sperm whales. *Behaviour* **118**, 332–343. (doi:10.1163/156853991X00328)
4. Gero S, Whitehead H, Rendell L. 2016 Individual, unit and vocal clan level identity cues in sperm whale codas. *R. Soc. Open Sci.* **3**, 150372. (doi:10.1098/rsos.150372)
5. Beguš G. 2021 CiwGAN and fiwGAN: encoding information in acoustic data to model lexical learning with generative adversarial networks. *Neural Netw.* **139**, 305–325. (doi:10.1016/j.neunet.2021.03.017)
6. Beguš G, Zhou A. 2022 Modeling speech recognition and synthesis simultaneously: encoding and decoding lexical and sublexical semantic information into speech with no direct access to speech data. In *Proc. Interspeech 2022*, pp. 5298–5302. (doi:10.21437/Interspeech.2022-11219)
7. Beguš G. 2021 Identity-based patterns in deep convolutional networks: generative adversarial phonology and reduplication. *Trans. Assoc. Comput. Linguist.* **9**, 1180–1196. (doi:10.1162/tacl_a_00421)

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

8. Beguš G. 2022 Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks. *Comput. Speech Lang.* **71**, 101244. (doi:10.1016/j.csl.2021.101244)
9. Beguš G. 2020 Generative Adversarial Phonology: modeling unsupervised phonetic and phonological learning with neural networks. *Front. Artif. Intell.* **3**, 44. (doi:10.3389/frai.2020.00044)
10. Whitehead H. 2003 *Sperm whales: social evolution in the ocean*. Chicago, IL: University of Chicago Press.
11. Whitehead H, Rendell L. 2014 *The cultural lives of whales and dolphins*. Chicago, IL: University of Chicago Press. (doi:10.7208/chicago/9780226187426.001.0001)
12. Sharma P, Gero S, Payne R, Gruber DF, Rus D, Torralba A, Andreas J. 2024 Contextual and combinatorial structure in sperm whale vocalisations. *Nat. Commun.* **15**, 3617. (doi:10.1038/s41467-024-47221-8)
13. Gero S, Bøttcher A, Whitehead H, Madsen PT. 2016 Socially segregated, sympatric sperm whale clans in the Atlantic Ocean. *R. Soc. Open Sci.* **3**, 160061. (doi:10.1098/rsos.160061)
14. Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. 2016 InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In *Proc. 30th Int. Conf. on Neural Information Processing Systems*, pp. 2180–2188. Red Hook, NY: Curran Associates Inc.
15. Donahue C, McAuley JJ, Puckette MS. 2019 Adversarial audio synthesis. In *7th Int. Conf. on Learning Representations, New Orleans, LA, USA, 6–9 May 2019*. OpenReview.net. See https://openreview.net/forum?id=ByMVTsR5KQ.
16. Pearl J. 2009 *Causality: models, reasoning, and inference*, 2nd edn. Cambridge, UK: Cambridge University Press.
17. Rubin DB. 1974 Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701. (doi:10.1037/h0037350)
18. Imbens GW. 2000 The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706–710. (doi:10.1093/biomet/87.3.706)
19. Holland PW. 1986 Statistics and causal inference. *J. Am. Stat. Assoc.* **81**, 945–960. (doi:10.1080/01621459.1986.10478354)
20. Rothenhäusler D, Yu B. 2019 Incremental causal effects. (https://arxiv.org/abs/1907.13258)
21. Friedman JH. 2001 Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232. (doi:10.1214/aos/1013203451)
22. Ribeiro MT, Singh S, Guestrin C. 2016 'Why should I trust you?': explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1135–1144. New York, NY: ACM. (doi:10.1145/2939672.2939778)
23. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. 2017 LightGBM: a highly efficient gradient boosting decision tree. In *Proc. 31st Int. Conf. on Neural Information Processing Systems* (eds I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett), vol. 30, pp. 3146–3154. Red Hook, NY: Curran Associates, Inc.
24. Lundberg SM, Lee SI. 2017 A unified approach to interpreting model predictions. In *Proc. Int. Conf. on Neural Information Processing Systems* (eds I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett), vol. 30, pp. 4765–4774. Red Hook, NY: Curran Associates, Inc.
25. Lundberg SM, Erion GG, Lee SI. 2018 Consistent individualized feature attribution for tree ensembles. https://arxiv.org/abs/1802.03888.
26. Athey S, Imbens G. 2016 Recursive partitioning for heterogeneous causal effects. *Proc. Natl Acad. Sci. USA* **113**, 7353–7360. (doi:10.1073/pnas.1510489113)
27. Künzel SR, Sekhon JS, Bickel PJ, Yu B. 2019 Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl Acad. Sci. USA* **116**, 4156–4165. (doi:10.1073/pnas.1804597116)
28. Beguš G, Zhou A. 2022 Interpreting intermediate convolutional layers of generative CNNs trained on waveforms. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 3214–3229. (doi:10.1109/TASLP.2022.3209938)
29. Beguš G, Zhou A. 2022 Interpreting intermediate convolutional layers in unsupervised acoustic word classification. In *ICASSP IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 8207–8211. (doi:10.1109/ICASSP43922.2022.9746849)
30. Madsen PT, Siebert U, Elemans CPH. 2023 Toothed whales use distinct vocal registers for echolocation and communication. *Science* **379**, 928–933. (doi:10.1126/science.adc9570)
31. Beguš G, Sprouse RL, Leban A, Silva M, Gero S. 2025 Vowel- and diphthong-like spectral patterns in sperm whale codas. *Open Mind* **9**, 1849–1874. (doi:10.1162/OPMI.a.252)
32. Beguš G, Dąbkowski M, Sprouse RL, Gruber DF, Gero S. 2026 The phonology of sperm whale coda vowels. *Proc. R. Soc. B* **293**, 20252994. (doi:10.1098/rspb.2025.2994)
33. Bengio Y, Courville A, Vincent P. 2013 Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828. (doi:10.1109/TPAMI.2013.50)
34. Wang X, Chen H, Tang S, Wu Z, Zhu W. 2024 Disentangled representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 9677–9696. (doi:10.1109/TPAMI.2024.3420937)
35. Kim H, Mnih A. 2019 Disentangling by factorising. https://arxiv.org/abs/1802.05983.
36. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A. 2017 beta-VAE: learning basic visual concepts with a constrained variational framework. In *Int. Conf. on Learning Representations (ICLR 2017)*. See https://openreview.net/forum?id=Sy2fzU9gl.
37. Liu X, Sanchez P, Thermos S, O'Neil AQ, Tsaftaris SA. 2022 Learning disentangled representations in the imaging domain. *Med. Image Anal.* **80**, 102516. (doi:10.1016/j.media.2022.102516)

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026

38. Suter R, Miladinovic D, Schölkopf B, Bauer S. 2019 Robustly disentangled causal mechanisms: validating deep representations for interventional robustness. In *Proc. 36th Int. Conf. on Machine Learning*, pp. 6056–6065. PMLR. See https://proceedings.mlr.press/v97/suter19a.html.
39. Yang M, Liu F, Chen Z, Shen X, Hao J, Wang J. 2021 CausalVAE: disentangled representation learning via neural structural causal models. In *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 9588–9597. New York, NY: IEEE. (doi:10.1109/CVPR46437.2021.00947)
40. Bose J, Monti RP, Grover A. 2022 Controllable generative modeling via causal reasoning. *Trans. Mach. Learn. Res.* OpenReview.net. See https://openreview.net/forum?id=Z44YAcLaGw.
41. Bose J, Monti RP, Grover A. 2022 CAGE: probing causal relationships in deep generative models. ICLR 2022 conference submission. OpenReview.net. See https://openreview.net/forum?id=VCD050En7r.
42. Chockler H, Kroening D, Sun Y. 2021 Explanations for occluded images. In *2021 IEEE/CVF Int. Conf. on Computer Vision*, pp. 1214–1223. New York, NY: IEEE. (doi:10.1109/ICCV48922.2021.00127)
43. Locatello F, Bauer S, Lucic M, Rätsch G, Gelly S, Schölkopf B, Bachem O. 2019 Challenging common assumptions in the unsupervised learning of disentangled representations. https://arxiv.org/abs/1811.12359.
44. Beguš G, Leban A, Gero S. 2026 Supplementary material from: Approaching an unknown communication system by latent space exploration and causal inference. Figshare. (doi:10.6084/m9.figshare.c.8598324)

Downloaded from royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.250829/6161336/rsos.250829.pdf
by UC Berkeley (University of California) user
on 28 August 2026