# A Proposal for the Use of Backtranslation to Decode Whale Songs

Morgan Rivers
Department of Physics
Freie Universität Berlin
danielmorganrivers@gmail.com

July 1, 2024

**Abstract**

This paper suggests a novel method for decoding whale songs using backtranslation techniques. I propose a framework that involves training models to translate whale vocalizations to English and back, ensuring consistency and accuracy through reinforcement learning. The approach is to maximize the likelihood of joint probability distributions, providing a robust model for understanding animal communication.

## 1 Introduction

Decoding whale songs presents a unique challenge in the field of animal communication. I believe communication with whales will both teach us for the first time a complete language of another species, as well as enable communication with many more animals. This will be critical to both understand the desires and steward the environment for these animals, as well as raise popular awareness of how humanity is affecting wild animals on this planet.
In this paper, I propose an unsupervised translation approach motivated by understanding animal communication. This involves using backtranslation techniques to create a model that can translate whale vocalizations into English and back, ensuring that the translations are contextually and semantically accurate.

I will use the term "Whale" to refer to the language Whales are speaking. For example, "she understands both Whale and English."

## 2 Related Work

Studies on neural networks, such as AlphaFold for molecular interactions, have shown the potential for applying similar techniques to decode whale songs.

One paper specifically about whales is interesting, and partially motivated this work: "A Theory of Unsupervised Translation Motivated by Understanding Animal Communication" (https://arxiv.org/abs/2211.11081)

However, it lacks specificity and clarity for their favored approach, and tries to cover a lot of ground which is not entirely needed for simply trying to translate Whale.

# 3 Introduction

This paper proposes a novel approach to translate whale vocalizations into human language using transformer-based language models. The method involves training multiple models on whale vocalizations and human language, then using a series of translation steps and penalty functions to refine the translation process.

# 4 Data Preparation

The initial data consists of a large corpus of whale vocalizations, encoded as text with accompanying environmental context. Each vocalization is attributed to a specific whale, with information about the time and surrounding whales. This data is formatted similarly to a theatrical script, with each entry requiring less than 2048 tokens.

# 5 Model Architecture

Three decoder-only transformer language models are utilized:

- $M_w$: Trained solely on whale vocalizations
- $M_e$: Trained solely on English text
- $M_{we}$: Trained on both whale vocalizations and English text

# 6 Training Process

## 6.1 English Context Fine-tuning

$M_e$ is fine-tuned on relevant English context, including:

- English plays
- Scientific literature about whales
- Biology and science textbooks focusing on whales
- Specific context about the region and family of whales being studied

### 6.2 Whale Language Model Training

$M_w$ is trained to predict the next tokens in the whale text, excluding environmental context unless dependent on whale actions.

### 6.3 Bilingual Model Training

$M_{we}$ is trained on both whale vocalizations and English text, with a focus on maintaining proficiency in both languages.

## 7 Translation Methodology

### 7.1 Whale to English Translation

$M_{we}$ is prompted with:

```
[Environmental context]
[Paragraph of Whale]
Please translate the above paragraph to English:
```

The resulting translation is denoted as *whale_to_english*.

### 7.2 English to Whale Back-translation

The *whale_to_english* translation is then used to prompt $M_{we}$ again:

```
[Environmental context]
[Paragraph of Whale translated to English (whale_to_english)]
Please translate the above paragraph to Whale:
```

The result is denoted as *whale_to_english_back_to_whale*.

## 8 Penalty Functions

Three penalty functions are employed to refine the translation process:

- $P_{we}$: English likelihood penalty
- $P_{wew}$: Whale context probability penalty
- $P_{sd}$: Semantic distance penalty

### 8.1 English Likelihood Penalty ($P_{we}$)

$P_{we}$ is calculated using $M_e$ to evaluate the average probability per token of the *whale_to_english* paragraph. A high $P_{we}$ indicates an unlikely English paragraph.

### 8.2 Whale Context Probability Penalty ($P_{wew}$)

$P_{wew}$ is determined by inserting *whale_to_english_back_to_whale* into the overall whale text and calculating the probability using $M_w$. A high $P_{wew}$ indicates low probability of the translated whale text appearing in that context.

### 8.3 Semantic Distance Penalty ($P_{sd}$)

$P_{sd}$ measures the semantic distance between the original whale paragraph and *whale_to_english_back_to_whale*. A high $P_{sd}$ indicates significant divergence in meaning. This penalty is weighted higher than $P_{wew}$ to prioritize meaning preservation.

## 9 Putting it all together

The penalties are used for reinforcement learning applied to $M_{we}$. This process is repeated for all available paragraphs, potentially tens of thousands of times, effectively fine-tuning the model on a paragraph-by-paragraph basis. Here is a summary of the training algorithm above:

**Algorithm 1** Backtranslation Algorithm

```
Initialize parameters θ
for each iteration do
    Translate whale vocalizations to English
    Calculate probability penalty Pwe
    Translate English back to whale
    Calculate contextual and semantic penalties Pwew and Psd
    Update parameters θ based on combined penalties
end for
```

### 9.1 Datasets

The datasets provided a comprehensive set of whale vocalizations and environmental contexts, essential for training and evaluating our models.

I would primarily be interested in using two whale datasets for this work:

1) Seven months of recordings from the THEMO observatory [1], containing both ambient and anthropogenic noise but no whale clicks,

2) Recordings of 1203 tagged clicks of a single whale from the Bahamas with 120 sec of noise-only recording and recordings of over 15,000 manually tagged clicks from multiple whales recorded on the Dominica Island together with 3.6 hours of noise-only data. [2].

Also see:

- Project CETI
- Project CETI Publications
- Project CETI Workshop - Decoding Communication in Nonhuman June 12-13, 2023 @ Simons Institute for the Theory of Computing

## 10 Validation

Validation is performed using unseen whale vocalizations, comparing the accuracy of *whale_to_english_back_to_whale* to that of the training data.

## 11 Potential Challenges

1. Poor performance of $M_w$
2. Insufficient context for accurate translation
3. Tendency towards common whale utterances in back-translation
4. Development of a system to "cheat" the translation process
5. Unclear direction of penalties for efficient learning
6. Overfitting to specific training data
7. Risk of translating only sounds and grammar without capturing meaning

## 12 Miscellaneous end notes

It might be better to have an ensemble of answers for a given translation whale_to_english_back_to_whale, and then boost the model with the best of these answers. That addresses 5, as it gives the model a direction to move the prediction.

Note: this ends up looking like an auto encoder! We are expanding into English language token space, and then going back to the original space.

Another way to help it start in the right direction is to use statistical matching: most common concepts are assumed to be all the whale words in English. So then you can give it context: given the above best guess translation, improve it to a better one. And then you pick the best answer closest to the input.

## 13 Conclusion

This approach presents a novel method for translating whale vocalizations to human language using transformer-based models and reinforcement learning. While challenges exist, the use of multiple models and carefully designed penalty

functions offers a promising direction for future research in animal communication translation.

# References


[1] Roee Diamant, Anthony Knapr, Shlomo Dahan, Ilan Mardix, John Walpert, and Steve DiMarco. *Themo: The texas a&m-university of haifa-eastern mediterranean observatory*. In 2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO), pages 1–5. IEEE, 2018.

[2] Link to implementation code and database [online]. available at: https://drive.google.com/drive/folders/1HkGZcpYbrft3pGtqdt45bUZXVcrJVrdm?usp=sharing.