Spectral Analysis of Sperm Whale Vocalization

Lynn Robey

Bachelor's thesis
2024:K7

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Spectral Analysis of Sperm Whale Vocalisations

Bachelor's thesis in mathematical statistics

MASK11

Lynn Robey
Supervisor: Ted Kronvall

LUND
UNIVERSITY

Centre for Mathematical Sciences, Faculty of science
Lund University
Sweden
May 2024

Abstract

Sperm whales *(Physeter macrocephalus)* communicate with one another using an intricate language made up of sequences of clicks. Recent studies have begun to dig deeper into the structure and content of these vocalisations in an effort to detect and extract significant patterns and features. Whether or not the results of such research will eventually allow us to translate sperm whale speech into something akin to a human language is a subject of debate. Regardless of the ultimate outcome, the insight is of interest to many scientific fields, including conservation biology and bio-inspired engineering. For this project, over 1400 audio clips recorded between 1952 and 1995 were processed in order to isolate clicks from background noise. Frequency analysis on over 5000 detected clicks yielded dominant frequencies in the range of 2 kHz to 10 kHz with varying frequency distributions. It was possible to group the clicks into distinct categories based on their frequency contents, dominant frequencies, and number of strongly represented frequencies. This paper illustrates the performance of methods for automatically detecting click onsets in noisy data, as well as an effective approach for fitting an estimated signal envelope to clicks based on a Gumbel probability density function. Several patterns in the spectral features of sperm whale clicks were identified. The recurring patterns in the value and number of dominant frequencies in each click suggest the existence of multiple distinct click types, and several examples of potential click categories are described in this paper. There also appears to be a noteworthy relationship between the geographical location, dominant frequency, and spectral envelope parameters of clicks. Further studies will be necessary to validate and refine the results described here, which suggest that sperm whale vocalizations consist of a more rich and complex combinatorial structure than has previously been described and that this structure varies systematically between geographical locations.

Populärvetenskaplig sammanfattning

Vad pratar valar om? Kan man prata med valar? Några forskare tror att det är möligt och de vill översätta valarnas språk. Kaskeloter (*Physeter macrocephalus*) är stora marina däggdjur som finns i hav över hela världen. Dessa valars läte består av två typer av klick-ljud. De första är ekolokaliseringsklick för att hitta bytesdjur och navigera, men man vet lite om den andra. På 1950-talet började vetenskapen undersöka kaskeloters vokalisering och kom att tro denna andra typ av klick är som ett mänskligt språk. Kaskeloter turas om att göra komplexa serier av klick i specifika möster, ungefär som meningar. Nu, med nya teknologier inom artificiell intelligens, tänker forskarna att de kan tyda vad valar säger.

Organisationen CETI, som arbetar med att förstå hur kaskelottens språk fungerar, studerar kanskeloter utanför ön Dominica i Karibien. Med hjälp av artificiell intelligens de har börjat sammanställa en sorts ordlista och ett alfabet aspekter för kaskelot tal. Liknande metoder som används för att studera mänskligt språk och musik kan även användas på valars läten. De flesta av dessa metoder kommer från statistisk signalbehandling. Man kan beskriva ett klick som en kombination av frekvenser och tidsmässiga egenskaper. Om det finns mönster, kan de koda viktig information om innebörden av klicket. Svårigheten ligger i att hitta dessa mönster och avkoda deras betydelse. Den här rapporten beskriver några egenskaper hos klick som identifierats med hjälp av signalbehandlingsmetoder. Resultaten tyder på ett ännu mer komplext "alfabet" än vad som hittills har beskrivits för kaskeloter. Det finns fortforande mycket arbete kvar att göra, men ju mer människor lär sig desto mer komplext och intressant verkar kanskeloters språk.

Contents

Introduction

The ocean and all it contains has long been a mysterious and inaccessible place from a human perspective. We are only able to directly observe a small fraction of the marine world. Various tools have made it more accessible and unlocked some mysteries over time, but proportionally there is still nearly as much to be explored in our oceans as there is on other planets [10]. When the first recordings of sperm whales were analysed for their spectral properties in the mid 20th century, observation showed that individuals produced different patterns of clicks. It was initially assumed that the patterns were simply some version of a personal identifier, akin to a human name [21]. Changing methods of listening to and recording sound underwater, as well as more efficient methods for statistical analysis were necessary before the level of intricacy in sperm whale communication could even be guessed at. The challenge of understanding sperm whale vocalisations is of at least dual interest. One such interest comes from the ecological perspective, as having such a direct insight into the lives of these keystone marine mammals could be very useful to interpreting the well being of both whale populations and the ecosystem that they inhabit and in turn putting more effective protections in place. From a more technical and mathematical perspective, coming up with ways to interpret a completely unknown system of communication provides an opportunity for innovation and advancement of statistical methods and machine learning techniques. In this paper, methods from signal processing, including musical analysis and human formant recognition, are modified to suit the composition of whales' clicks.

Methods and tools from spectral analysis

Collections of recorded sperm whale vocalisations date back to the 1950s, pre-dating some of the most fundamental tools in modern signal processing and spectral analysis. The eponymous Joseph Fourier is generally credited with the method of describing a function using its constituent frequencies, as described in his 1807 memoir *Analytical Theory of Heat* [15]. Although traces of similar ideas are found in work done by Leonhard Euler in the 18th century as well as in work from other notable mathematicians of the past [11]. In the 21st century the Fourier transform has become ubiquitous in most, if not all, scientific disciplines, and is one of the most important concepts in signal processing and analysis. In 1965 J. W. Cooley and J. W. Tukey published their fast Fourier transform (FFT) algorithm, a more efficient way to digitally compute the discrete Fourier transform (DFT) [11].

Cooley and Tukey's FFT algorithm requires a number of operations proportional to $N \cdot log(N)$ to compute N coefficients as opposed to older methods which required N^2 operations [9]. This optimization was a significant improvement, and facilitated analysis of larger volumes of data more rapidly and with a lower computational cost. Applying the FFT to time series data such as recorded sperm whale vocalisations translates the data into the frequency domain, allowing for identification of spectral features. In human speech and music the FFT and other spectral analysis techniques have been widely used to study and describe various features and their significance, as well as to develop tools such as music transcription and speech recognition [15]. Some of the techniques that have already been developed for other applications might be useful in less familiar contexts, such as whale speech.

Sperm whale communication

Sperm whales are marine mammals that live in groups in nearly every part of the world's oceans. They are known to be highly social creatures, demonstrating strong bonds between individuals and a distinct capacity for social learning[1]. Sperm whales communicate using a complex arrangement of broadband clicks, but just how nuanced these clicks and the information they are capable of conveying is still uncertain to humans. Early observations from marine biologists and other observers identified certain patterns in the click sequences, which have come to be known as codas, and noted that the codas were often specific to certain groups of whales [1]. A "conversation"[2] between whales consists of multiple codas which whales exchange, sometimes repeating or simultaneously sounding the same codas [18]. Codas are recognizable patterns of approximately 3 to 10 clicks lasting under 2 seconds, with different groups of sperm whales around the world utilizing distinct sets of patterns. In the well-studied Eastern Caribbean Clan near the island of Dominica for example, 22 distinct coda types have been identified [3]. The clicks within a unit consist of several decaying pulses separated by around 3-4 milliseconds (figure 2), the interval between pulses is thought to depend on

[1]Young sperm whales have even been described as going through a "babbling" phase of language acquisition before they learn to use the same vocalizations as the adult whales [1].

[2]It should be noted that the use of anthropocentric terms such as "conversation" or "grammar" must be taken as convenient descriptors rather than strict definitions, as it is uncertain to what degree whale vocalizations mirror the structure of human language. It is possible that whales communicate in a way that no existing human concept can acutely describe. The use of quotations throughout this paper is meant to serve as a reminder of this distinction.

the reverberation within the whale's spermaceti organ, and thus is dependent on the size of the whale [1] (figure 1).

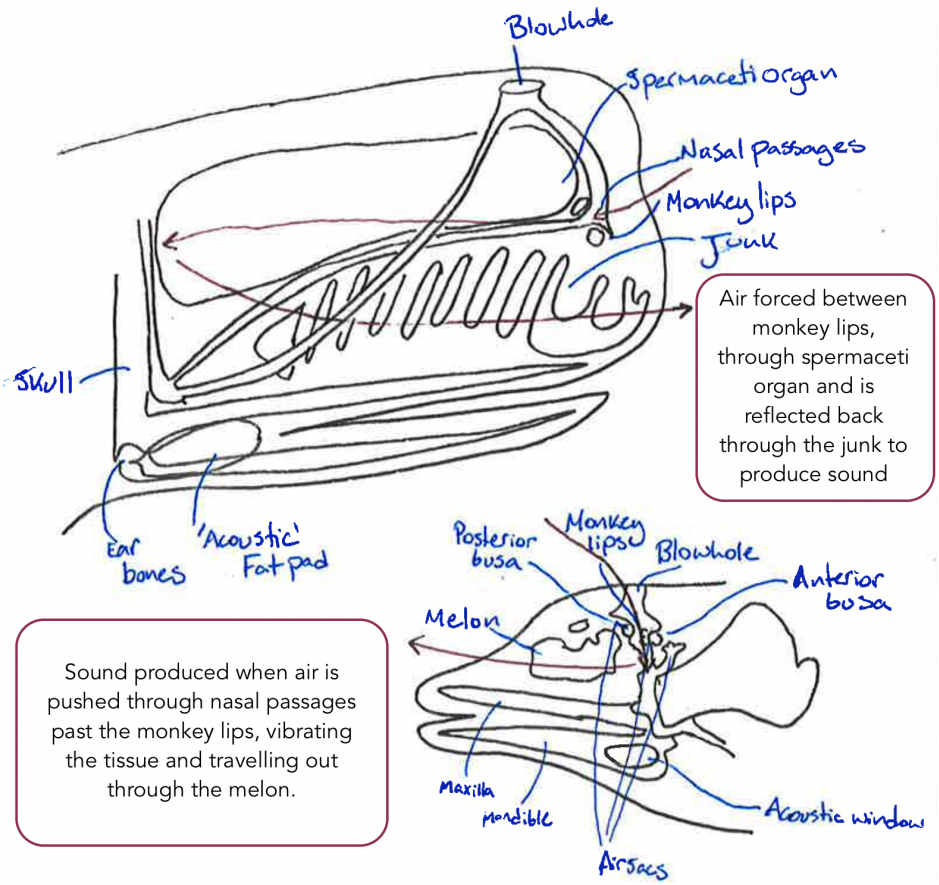


Figure 1: Diagram of sound production in sperm whales (above) and bottlenose dolphins (below). The exact mechanisms are still being studied, but sound production is believed to be controlled by air forced through these nasal complexes.

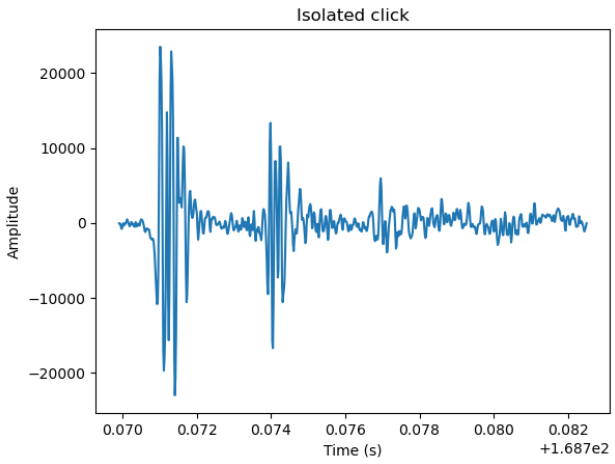


Figure 2: Example wave-form of a sperm whale click.

Coda clicks are easily distinguished from the sperm whales' higher-frequency echolocation clicks, but any potential distinctions among coda clicks themselves are more subtle. New technologies and statistical methods have improved scientists' abilities to both collect and analyze marine audio data. Consequently, the potential for successfully interpreting sperm whale vocalizations has increased greatly. For now, the purpose and meaning of sperm whale codas largely remains a mystery.

Acoustic monitoring of sperm whales

The effort to intentionally record whales whenever the opportunity presented itself began around 1957, although earlier recordings also do exist [21]. Whales in general began to become more prominent in the cultural and scientific zeitgeist with the "save the whales" movement that followed Roger Payne's discovery of singing in humpback whales at the end of the 1960s and led to the dismantlement of the whaling industry [8]. In the case of sperm whales as well as other species, recordings of their vocalisations have for the most part been taken using arrays of hydrophones hung off the sides of a boat. By 1977 temporal patterns in recordings taken in the North Atlantic were beginning to be formally studied and described [21]. However, the understanding that these clicks might make up a complex socially learned language would not gain solid footing for another couple of decades.

Background and related research

The Dominican Sperm Whale Project and Project CETI

The Dominican Sperm Whale Project (DSWP) began in 2005, principally led by Dr. Shane Gero, with the aim of conducting a long-term and in-depth study of the sperm whale population residing near the Caribbean island of Dominica [16]. The DSWP has followed the same population of whales for nearly 20 years, building an understanding of their behaviours and relationships. Project CETI (Cetacean Translation Initiative) branched from the DSWP in 2020, and was specifically focused around the project of interpreting the whales' language by leveraging new robotic technologies including more advanced machine learning tools and recording devices, coupled with computer algorithms, specifically designed for collecting large amounts of high-quality recordings of marine mammals [8]. CETI has studied the vocalizations of the Dominican sperm whale population through many different lenses. Cryptographers, marine acousticians, roboticists, statisticians, linguists and other experts have contributed their perspectives, and they have been able to propose quite a few potentially important features and qualities of conversations, codas, and individual clicks (see Appendix A). In order to get better results from unsupervised machine learning and generative models, CETI has put the infrastructure in place to vastly increase the amount of data they have at their disposal. The acoustic biologging tags, called D-tags, deployed to record vocalizations are attached directly to the whales [18], and therefore their data is much more uniform, cleaner, and can also comprise contextual information about a whale's identity and behaviour. In contrast to the traditional hydrophone recordings, the relative predictably of the tag recordings also makes more streamlined and accurate automated processing and click detection possible. While this thorough study is integral to any possibility of decoding or replicating sperm whale communication as CETI hopes to do, it is quite costly in terms of time and resources, and it is not necessarily clear if or how the new findings might compare to populations outside of the Eastern Caribbean Clan.

Published results from CETI and DSWP

One of the first publications to come from the DSWP on the topic of sperm whale codas in 2008 described observations of the temporal structure of codas and the apparent call-and-response and overlapping patterns in coda exchanges between whales [17]. Not long after, in 2011, the DSWP team

offered evidence against the idea that codas were simply individual identifiers. Using comparisons between the inter-click-intervals (ICIs) and relative shapes of the clicks in a set, they suggested that sperm whales might in fact be communicating much more information than had previously been believed [2]. In 2019, so called deep learning, inspired by contemporary analysis of human speech, was applied to the problem of identifying and classifying clicks from spectrograms. Using Convolutional Neural Networks (CNNs) the team achieved a 99.5 % accuracy rate in detecting the presence or absence of a click, with a 97.5% accuracy in classifying codas from the Dominican sperm whale data set, and 93.6% accuracy for the Eastern Tropical Pacific data set which includes a larger number of coda types [5]. In 2022 CETI garnered wide attention through a popular-science-style article outlining their project and the belief that the complexity of sperm whale communication may well be at the same or a similar level to human language [1]. An attempt to use generative adversarial network (GAN) models to determine significant features of whale codas was described by Beguš et al. in 2023, and the results supported the generally held belief that codas are a primary distinguishing feature [3]. The GAN approach, which the team labeled "causal disentanglement with extreme values", also suggested that spectral mean and regularity across all clicks within a coda may be important. While machine learning decreases the human bias in attempting to interpret vocalizations, some studies have emphasized this nonobjectivity by drawing direct parallels to human languages or even music. Sharma et al. drew direct comparisons between certain aspects of codas and the musical concepts of ornamentation, rubato, rhythm and tempo [18]. On the other hand, Beguš et al. described certain features as analogous to human vowels, diphthongs, and pitch [4]. With linguistics and music theory being well-developed fields of study, methods for describing diversity among human languages or musical principles can be useful comparisons in the investigation of whale speech. So far most studies have focused on larger-scale variations, especially at the level of codas and conversations. Comparatively little has been reported about the properties of individual clicks and the potential distinctions on a smaller-scale.

Precedent in other cetaceans

Bottlenose dolphins (*Tursiops truncatus*) have been studied far more than other cetacean species because they can be studied and kept in captivity relatively easily. Much like sperm whales, dolphins use vocalisations for both echolocation and communication. Their clicks have been described as belonging to certain categories of frequency distributions. Click types defined

by Houser et al. in 1999 formed seven groups:

1. unimodal low dominant frequency
2. unimodal low dominant frequency with a higher secondary frequency
3. bimodal equally dominant high and low frequencies
4. unimodal high dominant frequency with lower secondary frequency
5. unimodal high dominant frequency
6. wideband over frequency range without distinct peaks
7. 3 or more distinct regions in the frequency range

In dolphins, as well as false killer whales (*Pseudorca crassidens*), the categorization was done for broadband echolocation clicks with peaks in the range of 46-100 kHz [12]. A later study described four distinct general categories of dolphin click types using defining characteristics including duration, peak frequency and number of peaks in a click [7]. Other work has described a high level of finesse in dolphin echolocation clicks, and posited that they are capable of muscular regulation this precision [19]. While sperm whale coda clicks are distinct from their echolocation clicks, if they are capable of controlled frequency modulation, it is plausible that, as with dolphin echolocation, communicative clicks could similarly be categorized based on their frequency patterns. It has previously been suggested that dolphin clicks, or those of other cetaceans, might be better modeled by a Gumbel-like function than by more traditional envelope functions [6]. The Gumbel function therefore might be well-suited to modelling sperm whale clicks.

Motivating questions for this paper

Taking inspiration from prior research in dolphin vocalisations and contemporary studies on sperm whales,

- can multiple patterns be identified in the frequency content of different clicks produced by sperm whales?
- can an appropriate method of estimating the signal envelope be defined and implemented based on a Gumbel density function? Are the resulting Gumbel parameters useful descriptors of the shape of an individual click?

In summary, can distinguishing features in the spectrum and shape of individual clicks be identified and defined such that the clicks themselves can be effectively described by a simplified set of parameters?

Methods

Data

Over 1400 sperm whale audio clips recorded between 1952 and 1995 and the associated metadata are openly available on the Watkins Marine Mammal Sound Database webpage from the Woods Hole Oceanographic Institution and New Bedford Whaling Museum [13]. The quality, duration, and content of the clips varies greatly. In general, recordings were taken using a hydrophone array attached to a boat with sample rates between 10 and 166.6 kHz, and the clarity of the whales' clicks depends on the levels of background noise and distance between the whale(s) and the recording device among other factors. Some of the clips were not usable for this investigation due to the clicks being nearly indistinguishable from the general noise. After sorting out clips with especially high levels of noise, other species vocalising, or indistinguishable clicks, 804 recordings totaling 76.44 minutes were chosen for analysis and a subset of 421 were marked as especially good quality.

Onset detection

Due to the inconsistencies between datasets, automating a reliable system for detecting clicks in the data presented a substantial challenge. Multiple approaches were attempted and compared in order to determine which, if any, worked well across the majority of the data. The approaches were based

on methods for onset detection in music transcription as described by Klapuri and Davy in *Signal Processing Methods for Music Transcription* [14]. The sperm whale clicks are transient events, which are for the most part easy to pick out visually and audibly from a waveform or spectrogram representation of a recording (figure 3).

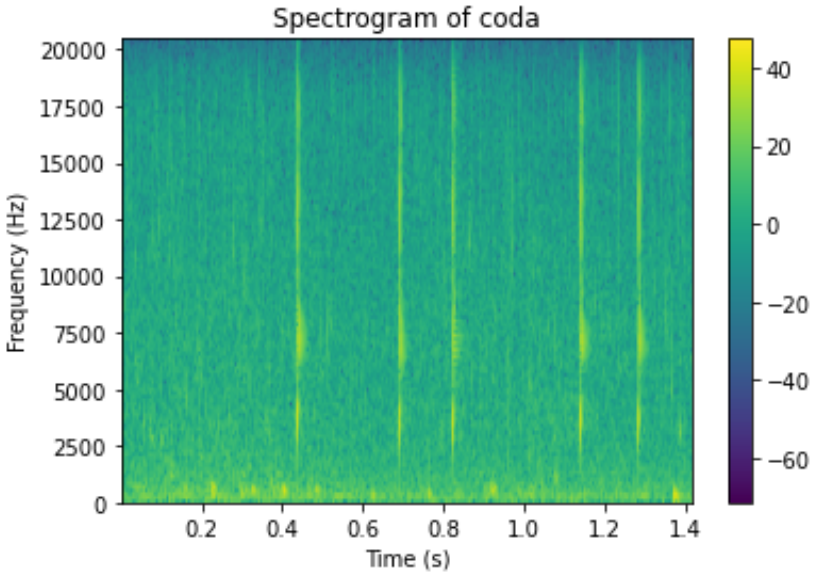


Figure 3: An example of the spectrogram representation of a 5-click coda. The 5 clicks stand out clearly in this relatively low-noise sample.

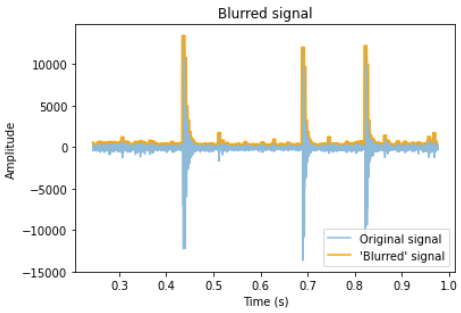
Based on the nature of the clicks, the transient event onset detection is the most applicable of the methods mentioned by Klapuri and Davy. This method as well as two modified versions were tested on random samples of the full available data set. In all, four methods for detecting clicks were compared using hand-annotated data.
Initial exploration of the collection of recordings showed that a high-pass filter in the range 2000-4000 Hz removed the periodic noise, which comes from sources including boats or the movement of water, that is present in some of the recordings without a significant change effect on the click peaks, thus a 4000 Hz Butterworth high-pass filter of order 4 was applied prior to initial click-identification and removed once clicks had been isolated. This is in keeping with other analyses where a band-pass filter ranging from 2000 Hz to 20 kHz was used [5]. The band-pass filter was also tested on this data, but it reduced the contrast between clicks and noise enough to make click detection less effective. Some data exploration also resulted in a relatively heuristic method of giving a first-glace description of a recording.

Rough cut-out

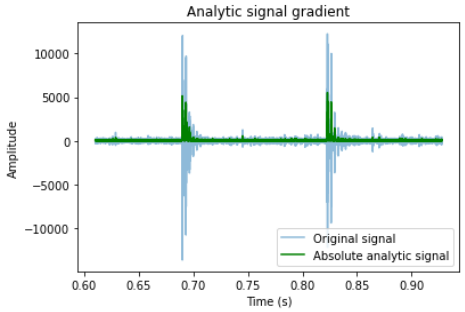
After applying the filter, a "blurred" version of the data was created by replacing each value by the maximum value in some neighborhood, resulting in a semi-smooth line over the noise, with a rough box around potential clicks (figure 4a). Labelling this "blurred" signal

$$S_{blur} = \{b_n | max(S[n - f_s\gamma : n + f_s\gamma])\} \tag{1}$$

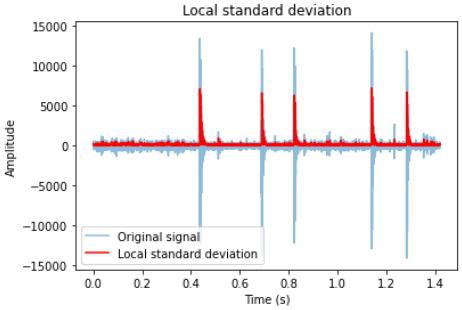
where S is the filtered signal, f_s is the sample rate, and b_n in the first and last γ seconds took the maximum value between the beginning or end of the data and either γ seconds before or after as appropriate so that S_{blur} had the same length as the original signal S. $\gamma = 0.003$ was chosen and adjusted thorugh some trial-and-error such that pulses within the same click would not be separated.



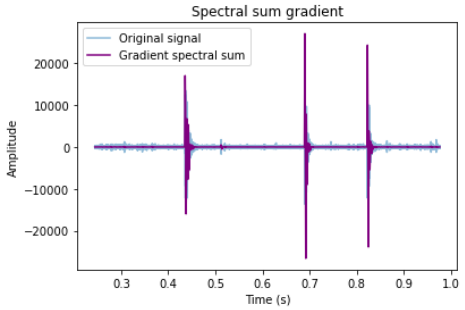
(a) "Blurred" signal $S_{blur}(1)$ with the original 3-click signal for reference. The "blurred" signal forms a rough outline around the signal peaks.



(b) Absolute value of the analytic signal D_{hil} (2) shown against the original 2-click signal.



(c) Local standard deviation D_{sd} (3) against original signal with 5 clicks.



(d) Spectral sum gradient D_{spec} (4) against original signal with 3 clicks.

Figure 4: Examples of how each onset detection method compares to the waveform of a sequence of clicks. The same click sequence is shown in all figures, zoomed in or out for better perspective of each method.

Intervals around potential clicks were defined based on where S_{blur} remained above some threshold. The height threshold began at $L_{blur} = \mu_{blur} + \sigma_{blur}$, where $\mu_{blur} = E[S_{blur}]$ is the mean over the entire clip and $\sigma_{blur} = \sqrt{E[(S_{blur} - \mu_{blur})^2]}$ is the standard deviation. The limit was increased by $\frac{max(S_{blur})}{100}$ if more than 30 intervals per second were identified. The minimum and maximum values of the "blurred" data then were used to assign a signal-to-noise ratio (SNR), typically defined $\frac{P_{signal}}{P_{noise}}$ where P is the power of the desired signal or background noise. In this case,

$$SNR = \frac{max(S_{blur})}{min(S_{blur})}.$$

This measure of SNR served as both a descriptor of the audio clips and to filter out the clips where the click onset detection would not be sufficiently reliable.

Analytic signal

Two approaches were tested to mark significant changes in the energy of the signal based on methods described by Klapuri and Davy for onset detection in musical pieces [14]. The transient event onset detection was implemented by using a Hilbert transform $\mathcal{H}$ for determining the analytic signal

$$S_A = S + i \cdot \mathcal{H},$$

where

$$\mathcal{H} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{S(t-\tau)}{\tau} d\tau$$

is the Hilbert transform of the signal. An envelope may then be created around the waveform by calculating $|S_A|$. Finally, representing the envelope's rate of change using the gradient,

$$D_{hil} = \nabla |S_A| \tag{2}$$

Sudden changes in the data should be reflected by peaks in the gradient (figure 4b). This approach was implemented numerically in Python using the signal library in the SciPy package [20]. The indices of peaks over a defined height limit and with a specified minimal distance in between peaks were extracted. Minimum height and distance limits were selected such that peaks within the same click were not included. The high variation in individual recordings combined with the differences between separate recordings meant that a flexible height limit must be used.

Local standard deviation

Inspired by the previous method (2) and observations of the data, a similar procedure was also applied to the local standard deviation of the waveform

$$D_{sd} = \{d_n | std(S[n - \zeta : n + \zeta])\} \tag{3}$$

with $\zeta = 10$. Thus, each d_n represents the local standard deviation, and peaks in D_{sd} represent a significant change in the data (figure 4c). As with S_{blur}, the standard deviation values at the ends of the audio sample were taken over a smaller range in order to maintain the same length as the original signal.

Spectral sum

Since clicks are often clearly visible in spectrogram representations of recordings, another variation on the method in (2) based on a frequency-based estimate was applied using the short-time Fourier transform (STFT),

$$D_{spec} = \{s_n | \nabla \sum_k |STFT_n|\} \tag{4}$$

In other words, each element of D_{spec} is the sum of power over the frequency range k in the corresponding frame, such that when a broad-spectrum event is present it is reflected by a larger value for s_n (figure 4d). The spectral-based onset detection (4) was designed to identify peaks in the gradient of the transformed vector, D_{spec}, with an initial height threshold of $L_{spec} = \mu_{spec} + \sigma_{spec}$, where μ_{spec} is the mean over the entire clip and σ_{spec} is the standard deviation defined as they were for the previous method (1). If more than 30 clicks per second were identified, the height threshold would be raised by $\frac{max(D_{spec})}{100}$ until fewer than 30 clicks per second were included. For the onsets based on the envelope-gradient D_{hil}, a "blurred" outline was created, similar to what was done for S_{blur} (1). The minimum value of the result min_{hil} was taken and used to set the initial threshold for identifying peaks $L_{hil} = min_{hil} + \sigma_{hil}$ with the standard deviation σ_{hil} of $|D_{hil}|$ being defined as before. An equivalent procedure was taken for the local standard deviation vector D_{sd} (3).
The minimum height min_{sd} based on the "blurred" vector was added to twice the standard deviation, i.e. $L_{sd} = min_{sd} + 2 \cdot \sigma_{sd}$, to set the initial height limit for peak-finding. The SciPy was again used, with the corresponding height limit and a minimal distance of 0.015 seconds, to identify the peaks in D_{hil} (2) and D_{sd} (3). For both of the two previous methods, if the number

of detected peaks was greater than 30 per second, the height limit would be increased by a fraction of the maximum value until the number fell below the threshold. The minimal distance of 0.015 seconds was based on the fact that click length varies but is made up of several pulses around 3-4ms [1], however, many of the recordings have a large amount of reverberation or echoes. Because of the presence of reverberation as well as the characteristic pulses of the clicks, peaks within 20ms after the previous one were removed post-detection.

Click analysis

To isolate clicks for further analysis, the chunks cut out based from S_{blur} (1) were narrowed down, first by checking whether any onsets based on the envelope-gradient method, D_{hil} (2), were present in that interval. If no onsets were detected, the section would be passed over. If more than one was found, the interval was checked to determine if it needed to be broken into multiple clicks or shortened. The remaining sections were refined so that the length was in the range of approximately 0.015-0.03 seconds with a significant shift in energy present within the first few milliseconds and the starting and ending points falling at the time where D_{sd} (3) was minimal within the appropriate time span and section isolated by (1).
The spectral properties of each click were then examined in the time-frequency domain by means of a periodogram. The periodogram is traditionally expressed as

$$R(f) = \frac{1}{n}\left|\Sigma_{t=0}^{n-1} w_t x_t e^{-i2\pi ft}\right|^2$$

with window function w_t. Various windows were tried against a selection of test clicks to compare performance, attempts to use a Gumbel-style windowing function gave wide main lobes and a smooth representation (figure 5), however a Hann window was ultimately chosen for analysis. The Hann window is often used for broadband random signals, and has minimal effect on frequency resolution and reduces spectral leakage. The window function is defined

$$w_t = 0.5 - 0.5cos(\frac{2\pi t}{M-1}) \quad 0 \leq t \leq M-1.$$

Spectral peaks from the periodogram were collected for the clicks isolated in each audio clip with a SNR value above five. The clicks were categorised based on the clearly prominent frequency peaks and labelled as belonging to one of six general categories; figure 5 is a good example of two fairly clear peaks.

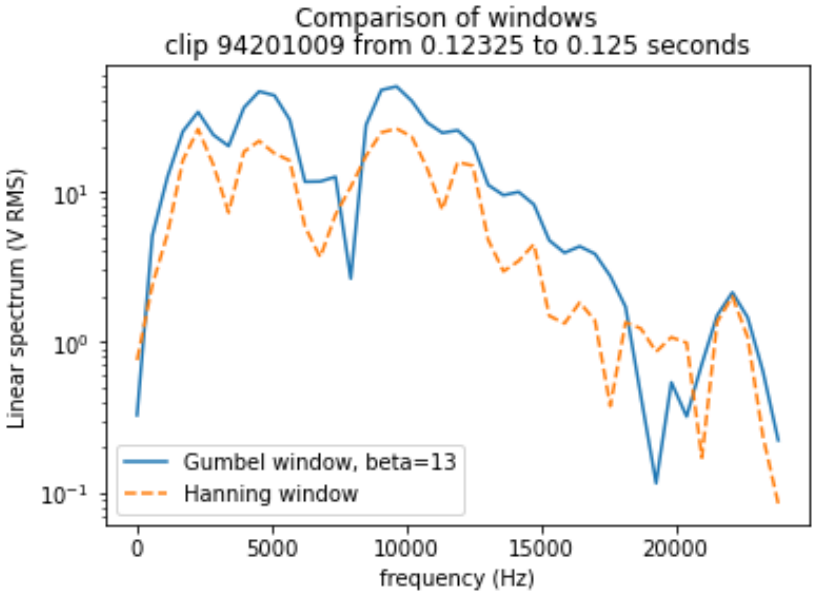


Figure 5: Comparison of Gumbel and Hanning windows. A customized window based on the Gumbel distribution gives more readibility on the log-scale, but reduces peak accuracy.

The frequency information, could then itself be examined and compared. In many clips there was a constant and strong noise just above 30000 Hz, if and when this was picked out as a peak in the periodogram, it was excluded as it was clearly not part of the click. The peaks in the periodogram were found by scaling the periodogram such that the values were between 1 and 0 and a uniform limit could be applied. This limit was set to a minimum height of 0.5 to considered a peak. The categories were defined based on categories used in previous studies on dolphin clicks, adjusted based on observations of common patterns in the whale click data. The categorisation was done algorithmically, marking each click as having either

1. one lower dominant frequency
2. one higher dominant frequency
3. two equally dominant frequencies (height difference of <0.1 on the scaled periodogram)
4. a dominant frequency followed by a secondary frequency
5. a secondary frequency followed by a dominant frequency

6. more than two frequency peaks

Another measurement, the click spectral centroid, was also considered. It is calculated as the sum of the magnitudes a_n from the Fourier transform of the signal, weighted by the corresponding frequency x_n divided by the sum of the magnitudes [14],

$$SC = \frac{\Sigma x_n a_n}{\Sigma a_n},$$

and was calculated for each detected click. The centroid does not provide the same level of detail as the periodogram peaks, but gives an average measure of where power is concentrated in the data.

Envelope fitting

The shape of a click in standard wave form is characterised by a sudden and sharp initial increase in amplitude, which tapers off more slowly as a series of pulses of decreasing amplitude. Due to these characteristics, a Gumbel density function (figure 6) was deemed suitable to fit a smooth estimated signal envelope to the click waves. Here let S denote the original click signal, and s samples from S. The maximum Gumbel density function takes two arguments, μ defining the centre on the x-axis and β defining the spread.

$$g(s) = \frac{1}{\beta} e^{-\frac{s-\mu}{\beta}} e^{-e^{-\frac{s-\mu}{\beta}}}$$

.
The envelope was fit by first transforming the click data using a Hilbert transform (figure 13a) and the analytic signal as defined in the calculation of D_{hil} (2). In the style of an empirical distribution function, the scaled cumulative sum of S_A is found by

$$F(t) = \frac{\sum_{j=1}^{t} s_{Ai}}{\sum_{j=0}^{T} s_{Aj}}, \quad t = 1, ..., T$$

(figure 13b). The Gumbel cumulative distribution function (CDF) can be expressed as

$$G(s) = e^{-e^{\frac{-(s-\mu)}{\beta}}}.$$

If $G(s)$ is double log-transformed it becomes a linear function,

$$\widetilde{G}(s) = -log(-log(G(s))) = -log(-log(e^{-e^{\frac{-(s-\mu)}{\beta}}})) = \frac{(s-\mu)}{\beta}.$$

$F(s)$ is transformed in the same way, giving $\widetilde{F}(s)$. Fitting by least-squares a line $y(s) = as + b$ to the log-transformed $\widetilde{F}(s)$, and taking $a = \frac{1}{\beta}$ and $b = \frac{-\mu}{\beta} = -a\mu$, gives the parameter values for the envelope as $\beta = \frac{1}{a}$ and $\mu = \frac{-b}{a}$.
Using the resulting μ and β, let $z = \frac{S-\mu}{\beta}$, the function representing the desired envelope is

$$\frac{1}{\beta} e^z e^{-e^z}$$

.

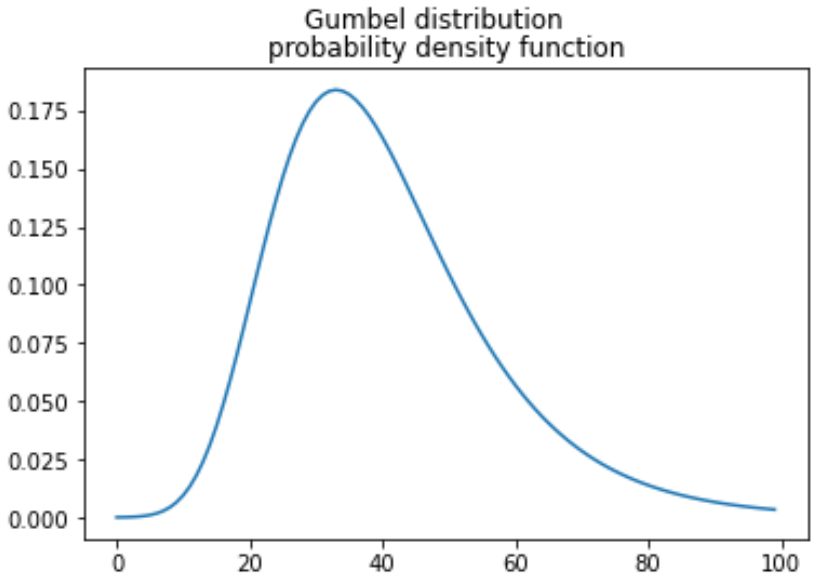


Figure 6: Shape of a Gumbel PDF with μ=0 and β=2. The shape of the curve is reminiscent of the general shape of a single sperm whale click.

Results

Onset detection

Potential onsets could be detected with reasonable reliability by applying a 4000 Hz high-pass filter to the signal to reduce noise, and using a combination of the "cookie-cutter" or "blurred" method (1) and the envelope gradient method (2). Testing against an annotated subset of the data to assess the performance of this and the three other onset detection methods showed that the accuracy of number of clicks found was similar for all methods. Figure 7 shows visually how the onset detection processes perform on a fairly clear

example clip.

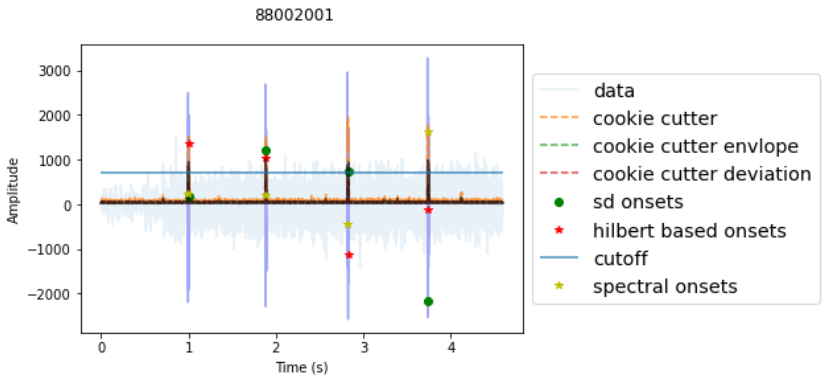


Figure 7: Example of onset detection applied to clip 88002001 from the dataset. All four methods correctly identified the four clicks in the clip.

The detection process was carried out on the set of 804 audio clips, 424 of which were annotated and could be used to check detection accuracy (figure 8). All methods performed very well on the least noisy data, however for the less clear data the gradient-based methods, especially D_{sd} (3) and less dramatically D_{hil} (2), tended to result in more false negatives, while D_{spec} (4) is quite sensitive in noisy data occasionally returned a large number of false positives.
The clicks used in the spectral analysis were isolated using a combination of the two most effective detection methods. this combined onset detection process returned the correct number of clicks at a rate of 63.208% with a mean error of 0.877. The accuracy is slightly lower than it was using only the cut-outs, but included fewer false positives.

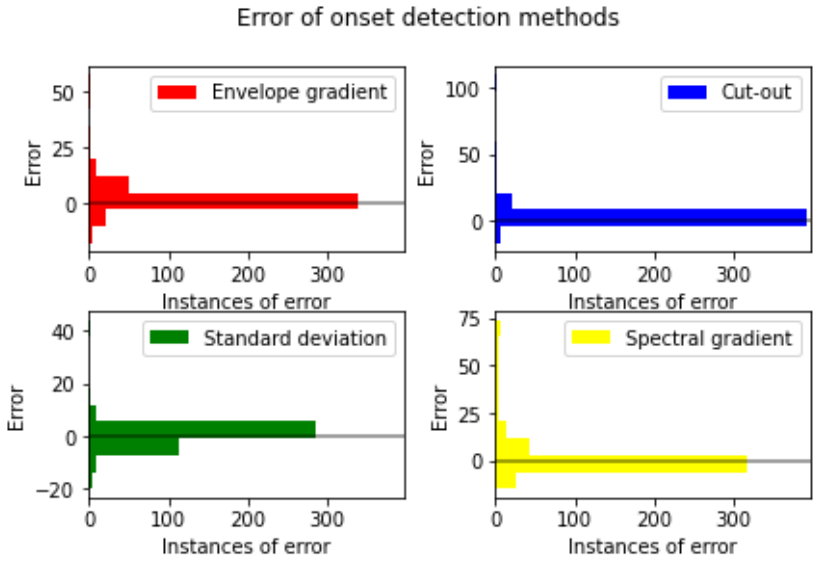


Figure 8: Histograms of error for each onset detection method. The error is based on an annotated subset of the data, such that the number of detected onsets can be compared to the true number of audible clicks. Envelope gradient (equation 2) returned a correct count rate of 43.396% with a mean error of 1.670, for the cut-outs (equation 1) this increased to 63.443% and 1.722, the standard deviation (equation 3) gave 56.604% and -0.47406, and spectral gradient (equation 4) 58.726% and 3.0731.

Frequency analysis

Sperm whale clicks contain frequencies primarily within the range 2000 Hz and 20,000 Hz. Individual clicks last approximately 20 milliseconds, but this length varies, presumably depending on the whale's size. For the clicks extracted, the two common dominant frequencies could immediately be observed in the spectral centroid distribution (figure 9), where the primary peak falls around 3500-5000 Hz and a secondary peak appears above 7500 Hz.

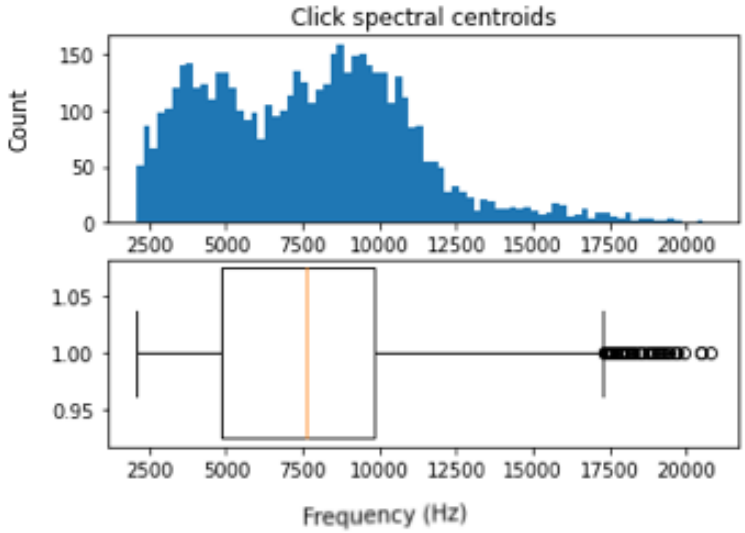


Figure 9: Histogram of spectral centroids of detected clicks and first quantile, median, and third quantile shown in the box plot. The centroids are not exact measurements of dominant frequencies, but reflect where energy is concentrated on the frequency spectrum. In this case frequencies are clearly concentrated in two places, just below 5000 Hz and just below 10000 Hz.

In the periodograms of individual clicks, similar patterns could be seen in the distributions of frequency as those previously described for bottlenose dolphin clicks. Implementing an automatic sorting method based on simple rules, it was possible to assign clicks to each category (figure 10). For clicks with a single lower frequency, the mean peak frequency was 2535 Hz, and the mean in higher single-peak clicks was 8345 Hz. It may even be possible to split the lower dominant frequencies into more specific frequency groups (figure 11). The remaining categories all had means between 4000 and 5000 Hz accounting for all peaks in the frequency distribution (figure 12). The exact frequencies that were dominant in the multi-peak clicks were less clustered around certain values, but fell within the same range as the single-frequency clicks and had a similar shape.

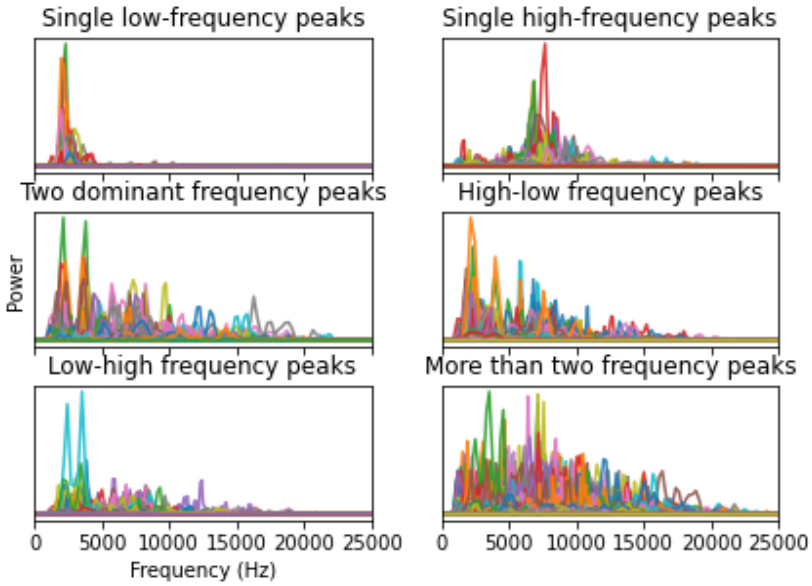


Figure 10: Overlay of categorized zero-padded periodograms of clicks. These plots show the general shape of each category using the examples from the available data, and not intended to serve as especially detailed descriptions. The number of clicks in each category, read left-to-right top-to-bottom, were 1867, 444, 153, 619, 485, 1339.

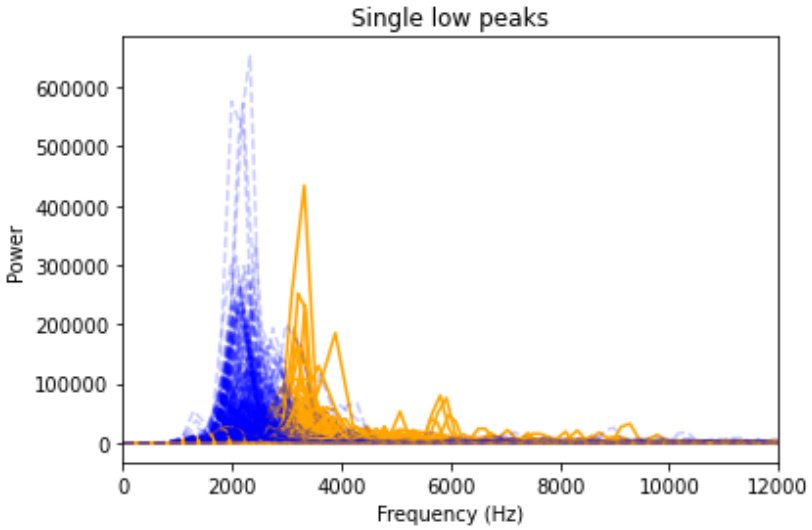


Figure 11: Periodograms of single low-frequency category with <3000 Hz in blue and 3000-6000 Hz in orange. Visually, the separation into two specific low-frequency groups results in two clear peaks.

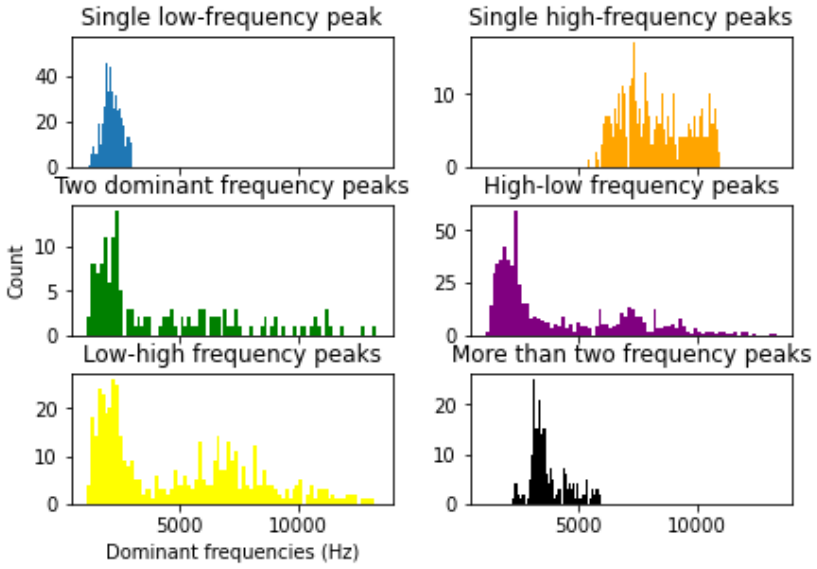
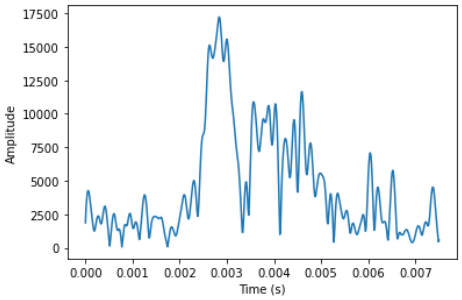


Figure 12: Histogram of dominant frequencies for clicks in each category.

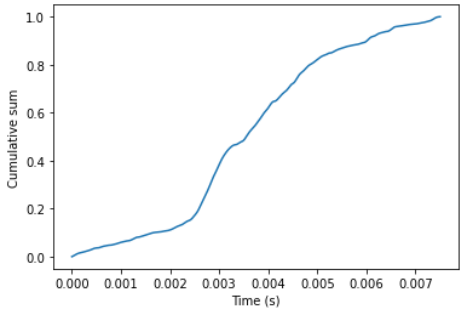
Fitted envelope

The envelope function was well suited to clicks that were extracted cleanly by the detection algorithm. The example in figures 13a-13d illustrates the steps of fitting the envelope and the final result. The wave and the envelope plotted together in figure 13d have been scaled to 1. As a measure of how well the envelope fit the data, the squared residuals were taken at the linear stage shown in 13c. In order to account for divergent behaviour at the ends of the transformed cumulative sum, caused by taking a double log of values approaching 1 as $log(log(1)) = log(0) \rightarrow -\infty$, the envelope fitting process was repeated with an additional step of truncating the cumulative sum at the beginning and end before fitting the linear line. Envelope functions were fit to a total of 5133 detected clicks. Truncating the values used to fit the line reduced the mean of the squared residuals from 99.166 to 22.836.
A comparison of the clicks was done based on their β parameter, most dominant frequency, and recording location. There may be latent variables associated with geographical location where the whale recordings were made, since it is related to the year and possibly other unknown circumstantial details. It does however appear that the click frequency and spread, represented by the β, is correlated with geographic location (figures 14 and 15). The clearest distinct groups are those from Dominica (figure15a), which cover a larger range of dominant frequencies but have generally lower β values, and Malta (figure 15j), which have mostly lower dominant frequencies but a wider range

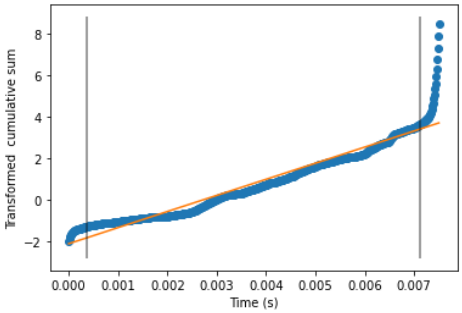
of βs.



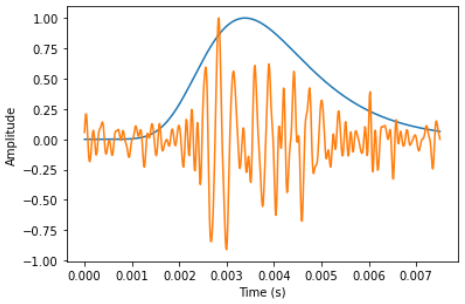
(a) Absolute value of the analytic signal S_A of click



(b) Scaled cumulative sum of S_A



(c) Fitted line of transformed cumulative sum with vertical lines marking the 5% of points at each end that were truncated to improve the estimation of the signal envelope.



(d) Click wave and Gumbel envelope with parameters determined by the fitted line

Figure 13: Process of fitting a Gumbel-style envelope to a single click.

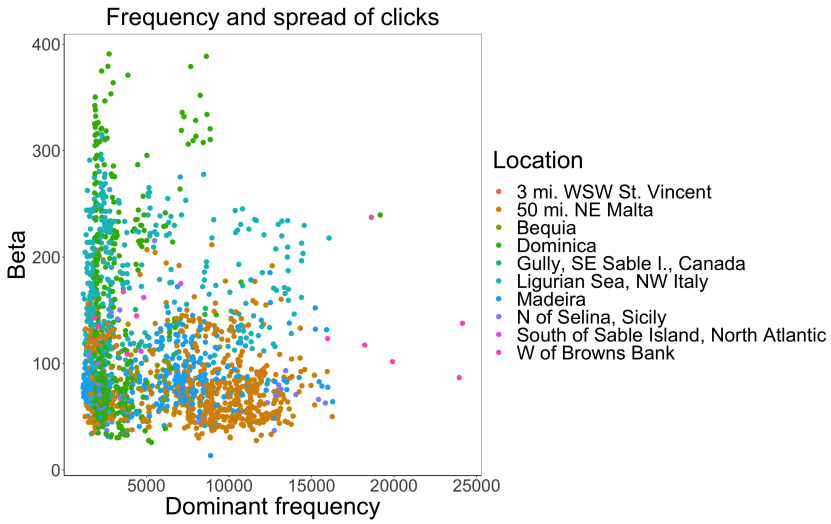
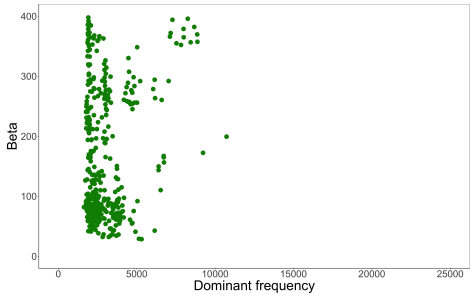
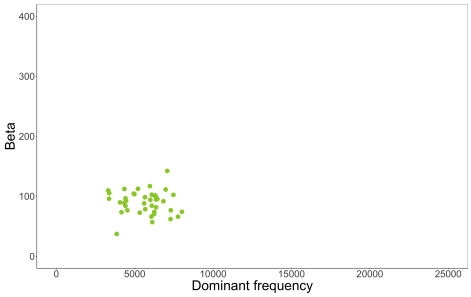


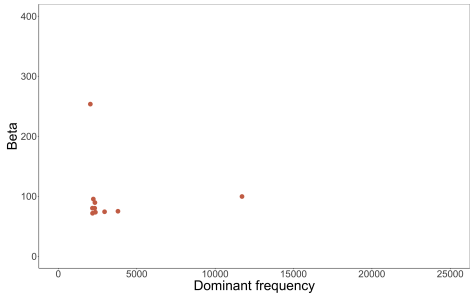
Figure 14: Click parameters and location. The values are largely concentrated in one area, but the variation seems to differ based on location.



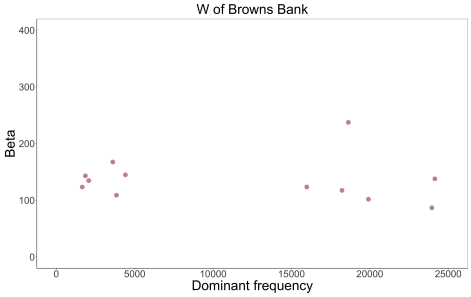
(a) Dominica



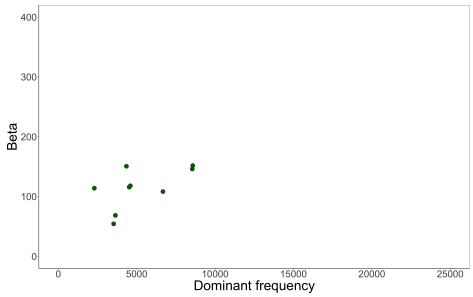
(b) Bequia



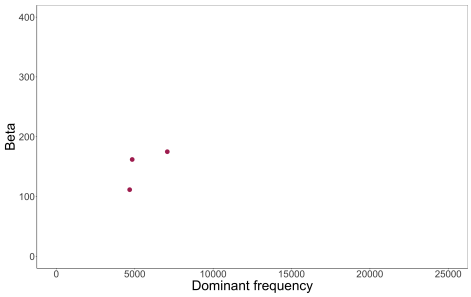
(c) St. Vincent



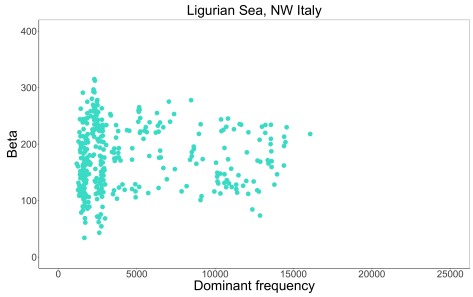
(d) Browns Bank



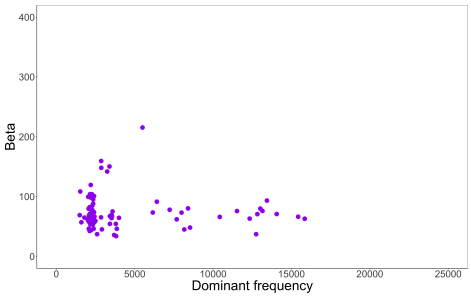
(e) Gully, SE Sable Island



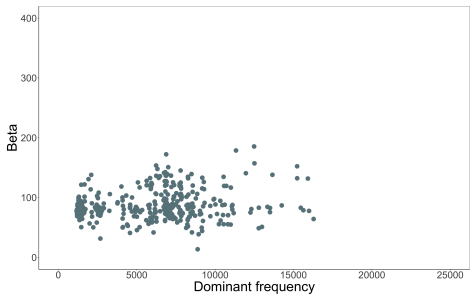
(f) South of Sable Island



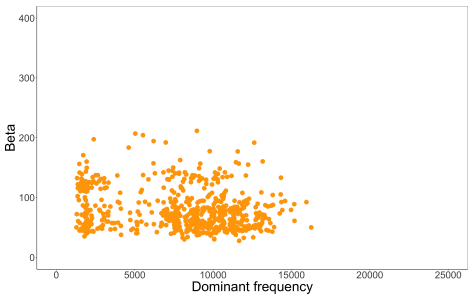
(g) Ligurian Sea



(h) N of Selina, Sicily



(i) Madeira



(j) Malta

Figure 15: Click parameters for individual locations.

Discussion

The structure of sperm whale clicks is much more nuanced than it immediately appears from the perspective of a human observer. Even using older and noisy data, certain spectral characteristics could be extracted from the data. The results are in general agreement with contemporary reports from other parties, but might suggest more variation than has been previously reported in terms of the frequency distributions. If these categories can be confirmed by additional testing, they would add even more complexity to proposed "phonetic alphabet" of sperm whales. The distributions of dominant frequencies may fall into categories similar to those previously defined for dolphin clicks rather than the binary categorisation that has been suggested. The categories described in this report could be supported by a large number of examples from the dataset and detected clicks, but even more specificity may be possible. It would be beneficial to validate or negate all categories tested, here and elsewhere, by alternative methods such as machine learning. While coda patterns have been the primary means of identification and investigation since sperm whale codas first became a subject of interest, the potential of finding defining features on a smaller scale could be useful as a complement, or when entire codas cannot be taken from a recording. In figures 14 and 15 the range of β values and strongest dominant frequencies appeared to vary based on the ocean region. In some regions, such as near Madeira, the frequencies fell within a wider range of values while βs remained within a narrower range. In other areas however, such as near Dominica, the opposite was true. Studies on different sperm whale acoustic data sets could give more insight into what distinctions are significant in the frequency content of clicks. The variation in dolphin clicks was linked to different contexts or environments. Accordingly, click features, as with codas types, may vary geographically or by population. Machine learning may also be a useful tool in order to validate differences in frequency patterns. Considering the apparent clustering in click shape and frequency by location, more detailed investigation into the cause of the grouping would be appropriate. Further studies might take into account more contextual variables, and data could be deliberately collected for the purpose of examining the connection between click frequencies and context.

There are several potential sources of error that may have affected the results of this study. The frequency and location information may be related to or affected by the varying methods and sources of the recordings. Different arrangements of hydrophones and their position relative to the vocalising whale might affect the frequency recorded. The inconsistencies in sampling rate may also influence the resulting audio. Similarities in the recording

conditions for each location, such as the date or equipment, could result in clearer clustering than would be present had those factors been controlled for. These factors should be taken into consideration in any future studies. Initial click onset detection in noisy and variable data could be performed reasonably well with minimal assumptions about the nature of the transient event. Even when employing conservative limits to avoid false positives, a large number of clicks could be extracted with good accuracy. Depending on the purpose of this type of detection, limits could be adjusted or one method in particular could be used in isolation. Multiple validation could also be used under the assumption that if all four methods agree to some level of precision, a transient event is present.

Revisiting motivating questions

The motivating questions described earlier in this paper were all addressed in this study. The results can be summarised as follows:

- patterns in the frequency content of clicks could be heuristically determined and clicks sorted into categories depending on the number and values of dominant frequencies. The categories proposed and tested here were initially based on the categories that have been defined in past studies on dolphin clicks, and adjusted according to observed frequency characteristics of clicks. Further work will be necessary in order to confirm or improve the suitability of the categories applied to sperm whale clicks.

- the Gumbel-style signal envelope estimation fit clicks fairly well, especially those clicks which were cleanly extracted from high quality recordings. Truncating the transformed cumulative sum before fitting a line further improved the estimation. Considering the β parameter from the estimated signal envelope in conjunction with the most dominant frequency and recording location for each click, showed some apparent clustering, suggesting that they represent significant features of the clicks.

Appendix A: Vocalisation terminology

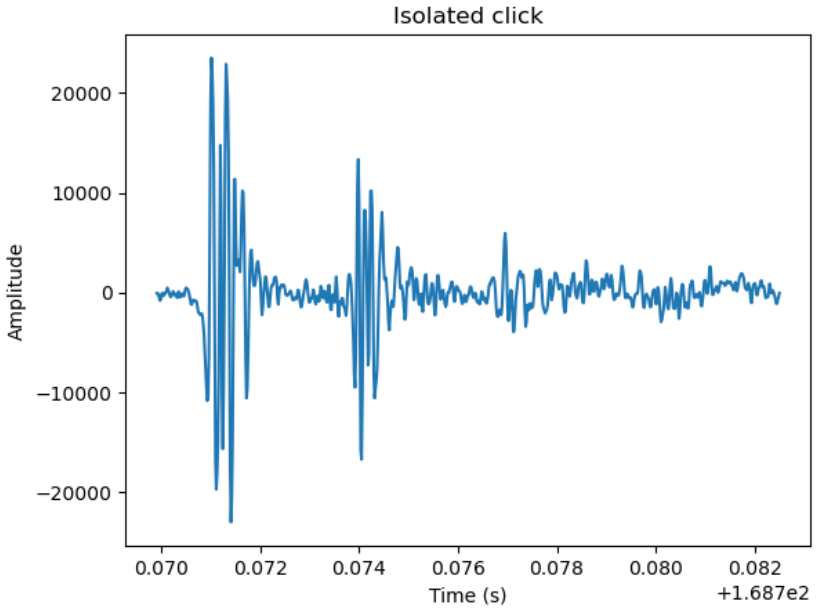


Figure 16: A single click

- **Click**: A single vocalization within a coda made up of several rapid pulses [1].
- **Inter-pulse-interval (IPI)**: Time between the pulses in a click.
- **Coda**: Short burst of clicks (<2s) that can be categorized into discrete patterns based on number of clicks and time between them [18].
- **Inter-click-interval (ICI)**: Time elapsed between individual clicks within a coda.
- **Duration**: Sum of ICIs in a coda [18].
- **Rhythm**: Normalised ICI category (ratios in ICIs rather than measured time) [18].
- **Spectral frequency**: [4]
    - **Coda level**: Mean frequency on a coda-level spectrogram.
    - **Click level**: clicks isolated prior finding the mean frequency across all clicks in a coda.

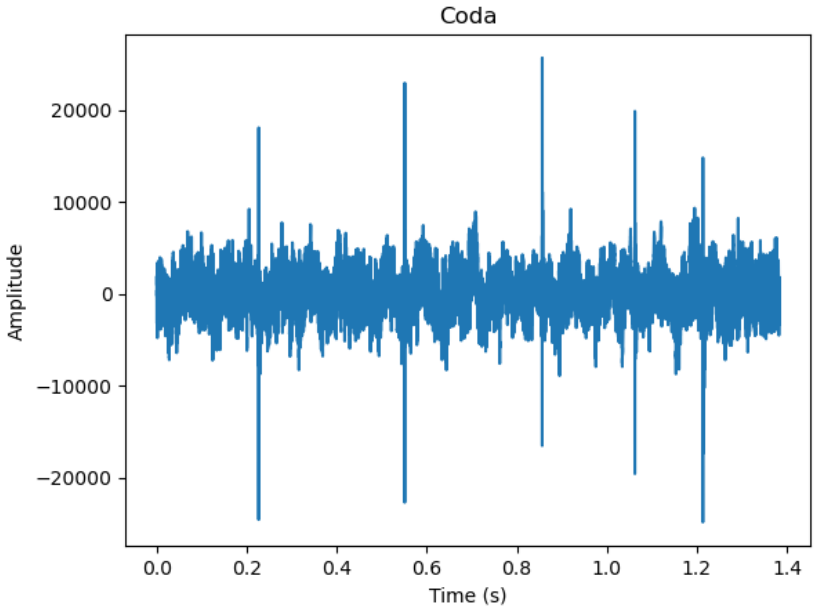


Figure 17: A coda with 5 clicks and ship noise in the background

- **Acoustic regularity**: Standard deviation of click spectral means in a coda [4].
- **Exchange**: Coda "conversation" between two or more whales, with matching, overlapping and repeating of codas.
- **Rubato**: Structure in variation in coda duration over an exchange (One whale might repeat a coda over the course of an exchange, but slightly alter the duration from one repetition to another) [18].
- **Ornamentation**: "Extra" clicks at end of a coda that otherwise matches the others in an exchange [18].
- **Tempo drift**: Difference in two codas' durations from the same speaker [18].
- **Tempo**: Duration, independent of rubato [18].
- **Chorusing**: Interactive exchange between two or more whales.
- **Conversational context**: Features (e.g. ornamentation) depend on the larger-scale context or patterns in an exchange [18].

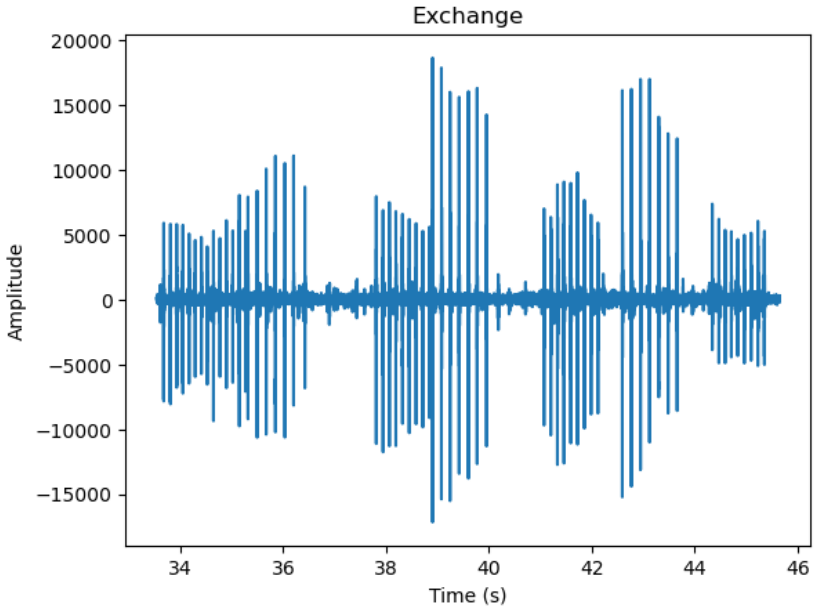


Figure 18: Part of an exchange between two sperm whales

- **Biological context (cues)**: Activity or position of an individual during or immediately post/prior to vocalization (e.g. diving, encountering another individual, etc.).

- **Trajectories**: Directional formant frequencies [4].

- **Coda dipthongs**: Rising, falling, rising-falling and falling-rising formant patterns are observed on individual codas, defined by formant frequency trajectories [4].

- **Coda vowels**: Recurrent spectral properties, likened to formant frequencies in human language [4].

    - A-type: Single pronounced spectral peak below 10kHz (around 5800Hz).
    - I-type: Two spectral peaks below 10kHz (around 3700Hz and 6200 Hz).

Acknowledgments

This work would not have been possible without the use of the audio recordings made available online at the Watkins Marine Mammal Sound Database from the New Bedford Whaling Museum and Woods Hole Oceanographic Institution. The computational analysis of the audio data was implemented using the NumPy and SciPy packages in python.

References

[1] Jacob Andreas et al. "Toward understanding the communication in sperm whales". In: *iScience* 25.6 (2022), pp. 104–393. DOI: https://doi.org/10.1016/j.isci.2022.104393.

[2] Ricardo Antunes et al. "Individually distinctive acoustic features in sperm whale codas". In: *Animal behaviour* 81 (2011), pp. 723–730. DOI: 10.1016/j.anbehav.2010.12.019.

[3] Gašper Beguš, Andrej Leban, and Shane Gero. "Approaching an Unknown Communication System by Latenet Space Exploration and Causal Inference". In: (2023). arXiv: 2303.10931 [stat.ML].

[4] Gašper Beguš et al. "Vowels and Diphthongs in Sperm Whales". 2024. DOI: https://doi.org/10.31219/osf.io/285cs.

[5] Peter C. Bermant et al. "Deep Machine Learning Techniques for the Detection and Classification of Sperm Whale Bioacoustics". In: *Scientific reports* 9 (2019), p. 12588. DOI: https://doi.org/10.1038/s41598-019-48909-4.

[6] Johan Brynolfsson, Isabella Reinhold, and Maria Sandsten. "A time-frequency-shift invariant parameter estimator for oscillating transient functions using the matched window reassignment". In: *Signal Processing* 183 (2021), p. 107913. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2020.107913. URL: https://www.sciencedirect.com/science/article/pii/S0165168420304576.

[7] Guiseppa Buscaino et al. "Pulsed signal properties of free-ranging bottlenose dolphins (Tursiops truncatus) in the central Mediterranean Sea". In: *Mar. Mam. Sci.* 31 (2015), pp. 891–901. DOI: https://doi-org.ludwig.lub.lu.se/10.1111/mms.12194.

[8] Project CETI. *Project CETI*. 2024. URL: https://www.projectceti.org/about (visited on 04/15/2024).

[9] James W. Cooley and John W. Tukey. "An algorithm for the machine calculation of complex Fourier series". In: *Math. Comp.* 19 (1965), pp. 297–301. DOI: https://doi.org/10.1090/S0025-5718-1965-0178586-1.

[10] Isabelle Gerretsen. *Why Nasa is exploring the deepest oceans on Earth.* Jan. 13, 2022. URL: https://www.bbc.com/future/article/20220111-why-nasa-is-exploring-the-deepest-oceans-on-earth (visited on 04/20/2024).

[11] Michael T. Heideman, Don H. Johnson, and C. Sidney Burrus. "Gauss and the History of the Fast Fourier Transform". In: *Archive for History of Exact Sciences* 34.3 (1985), pp. 265–277. DOI: http://www.jstor.org/stable/41133773.

[12] D. S. Houser, D. A. Helweg, and P. W. Moore. "Classification of dolphin echolocation clicks by energy and frequency distributions". In: *J. Acoust. Soc. Am.* 106 (1999), pp. 1579–1585. DOI: https://doi.org/10.1121/1.427153.

[13] Woods Hole Oceanographic Institute and New Bedford Whaling Museum. *Watkins Marine Mammal Sound Database.* 2024. URL: https://whoicf2.whoi.edu/science/B/whalesounds/fullCuts.cfm?SP=BA2A&YR=-1 (visited on 04/17/2024).

[14] Anssi Klapuri and Manuel Davy. *Signal Processing Methods for Music Transcription.* Springer New York, NY, 2006. ISBN: 978-0-387-30667-4. DOI: https://doi.org/10.1007/0-387-32845-9.

[15] Elena Prestini. *The Evolution of Applied Harmonic Analysis.* Applied and Numerical Harmonic Analysis. Springer New York, NY, 2016. DOI: https://doi-org.ludwig.lub.lu.se/10.1007/978-1-4899-7989-6.

[16] Dominica Sperm Whale Project. *The Dominica Sperm Whale Project.* 2024. URL: http://www.thespermwhaleproject.org (visited on 04/20/2024).

[17] Tyler M. Schulz et al. "Overlapping and matching of codas in vocal interactions between sperm whales: insights into communication function". In: *Animal behaviour* 76 (2008), pp. 1977–1988. DOI: 10.1016/j.anbehav.2008.07.032.

[18] Pratyusha Sharma et al. "Contextual and Combinatorial Structure in Sperm Whale Vocalisations". Dec. 8, 2023. DOI: https://doi.org/10.1101/2023.12.06.570484.

[19] Josefin Starkhammar et al. “Detailed analysis of two detected overlaying transient components within the echolocation beam of a bottlenose dolphin (Tursiops truncatus)”. In: *J. Acoust. Soc. Am.* 145 (2019), pp. 2138–2148. ISSN: 4. DOI: https://doi.org/10.1121/1.5096640.

[20] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[21] William A. Watkins and William E. Schevill. “Sperm whale codas”. In: *J. Acoust. Soc. Am.* 62.6 (Dec. 1977), pp. 1485–1490. DOI: https://doi.org/10.1121/1.381678.

Bachelor's Theses in Mathematical Sciences 2024:K7
ISSN 1654-6229

LUNFMS-4073-2024

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/