Revealing structure in vocalisations of parrots and social whales

Dissertation
for the award of the degree
“Doctor rerum naturalium”
Division Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

within the doctoral program Physics of Biological and Complex Systems
of the Georg-August University School of Science (GAUSS)

submitted by
María Florencia Noriega Romero Vargas
from
Ensenada, México

Göttingen, 2017

Thesis committee

Prof. Dr. Marc Timme, Max Planck Institute for Dynamics and Self-Organization

Prof. Dr. Florentin Wörgötter, Third Institute of Physics, Georg August University Göttingen

Dr. Kurt Hammerschmidt, German Primate Center

Members of examination board

1st Referee: Prof. Dr. Florentin Wörgötter, Third Institute of Physics, Georg August University Göttingen

2nd Referee: Prof. Dr. Marc Timme, Max Planck Institute for Dynamics and Self-Organization

Dr. Kurt Hammerschmidt, German Primate Center

Dr. Viola Priesemann, Max Planck Institute for Dynamics and Self-Organization

Prof. Dr. Stefan Luther, Max Planck Institute for Dynamics and Self-Organization

Prof. Dr. Reiner Kree, Institute for Theoretical Physics, Georg August University Göttingen

Date of oral examination

August 7th 2017

Contents

Chapter 1

Preface

Darwin's evolutionary theory [1] relocated humans from their divine place to the animal kingdom. Nevertheless, amongst all animals humans still occupy a privileged place, with traits believed to be exclusive and often superior. Jane Goodall was one of the first to challenge this belief when she reported the usage of tools by wild chimpanzees in the 60s [2]. Since then multiple studies have joined reporting animal behaviours that were believed to be exclusive to humans. Nowadays it is known that animals experience emotions and communicate them with individuals of their species [1, 3], they suffer from depression, have different personalities [4, 5] and cultures [6, 7, 8, 9], can be creative [10] and innovative [11, 12]. Still, there is one trait reserved for humans: language. Humans' faculty of language, is what still pays the bills for the privileged place humans have among animals [13]. But is language an exclusive human attribute that sets us apart from all other animals? This provocative question has not only inspired tales and fables through history but is central to the controversial discussion about the origins of language [1, 14, 15].

Many animals communicate by exchanging vocal signals with diverse purposes such as attracting mating partners [16, 17], defending territory [18, 19], maintaining group cohesion [20, 21] or alerting danger to other group members [22, 23]. Animal communication can be complex and even display parallels with human language such as semantics, syntax and vocal learning [24]. Many monkeys use semantic calls to refer to different predators. Campbell's monkeys, for instance, have one call for eagles and another call for leopards [25]. But the story does not end here, these calls are also used in syntactic combinations. Similarly to the way adding the suffix -hood at the end of the word brother changes its meaning to brotherhood, Campbell's monkeys use the suffix -oo to soften the function of these highly specific alarm calls. So that, adding -oo turns the leopard call into a call used for general disturbances within the canopy and the eagle call into a call used for less serious areal threats, such as the presence of eagles in the neighbouring groups or a branch falling [26]. Critical for speech, however, is the ability to learn and produce novel sounds, an ability that falls short in primates — due to the limited control they have over their vocal organs — but that is widely developed in many birds and marine mammals. Parrots and whales are vocally very flexible; well known for their abilities for mimicking human and other artificial sounds. Beyond human entertainment, these animals also use mimicking in their natural communication. These animals learn sounds by copying group members, a capacity known as vocal learning [27]. Certainly, animal communication systems are not as complex nor developed as human language. Yet, the parallels between language and animal communication invite us to consider that language may not be an isolated human attribute but may lay in a continuum with other animal capabilities. Studying animal communication may shade light on this open question.

Like language, animal vocal communication exhibits diverse structures that reveal aspects about the signaller. To illustrate this point, let us do a thought experiment. Consider for a moment we are aliens studying humans trough their vocal signals — speech. We do not understand the meaning of these vocal signals, let alone their minds. Yet, similar to the way astronomers know about stars they never visited, can tell their age and distance from structures in the light, we may learn about humans through structures present in their vocal signals. One of these structures is how these signals cluster geographically. These geographical patterns are, of course, a consequence of the different languages and tell that humans' vocal capacity is not genetically coded but learned. Looking in more detail, we would see that in addition to the learned aspects, there are universal characteristics that do not depend on the geographic location. For example, syntax will show in the way speech is made out of vocal units that are combined in non random ways. Another characteristic of speech, is how its tempo variations often correlate with the emotional state of the speaker. High arouse levels correlate with a higher tempo than low arouse levels. Tempo is a prosodic cue important to paralinguistic communication. The existence of languages, their syntax and prosody is essential to the way humans use sounds to communicate. Notice that these characteristics do not depend on understanding the meaning of words, yet they reveal important aspects of human communication. Likewise, we can learn about animals by studying their vocalisations.

Animal vocalisations are acoustic signals that fall within the scope of **bioacoustics**, a cross-disciplinary field that studies life sounds, mostly[1] of animals. Besides studying animal vocalisations, bioacoustics finds applications in areas such as monitoring ecosystems [29, 30, 31, 32], environmental conservation [33, 34], mitigation [35, 36] and even the search for extraterrestrial intelligence [37]. Bioacoustics studies sound production, sound detection and sound propagation. While the first two are involved biophysical processes, sound propagation is a purely physical phenomenon. Understanding the physical properties of sound can be insightful in trying to adopt an animal's perspective (Appendix A) and so, their communication needs.

Because of its physical properties, sound is an effective means of communication. Sound is a mechanical wave that propagates transporting information from its source at a speed determined by the medium (Appendix A). Sound needs a medium to propagate and it can travel large distances in dense environments like water [38, 39]. Electromagnetic waves on the other hand, get absorbed underwater and reflected by obstacles like tree leaves. Sound therefore, is more effective than light to communicate in such bulky environments.

The sound we hear typically comes from a vast range of sources, e.g. people speaking, birds singing, cars passing. Animal brains are very good at separating sound into its sources [40]. After all, natural selection favoured those individuals capable of differentiating the growl of a leopard from the yawn of a mate within complex soundscapes. While being an easy task for us, separating sound sources is challenging for a machine.

A common early step in most bioacoustic studies is that of annotating audio recordings with relevant information, such as animal vocalisations. This step is often addressed *manually* by listening to the sounds or by looking at spectrograms (graphical representations of sound, see Appendix B). However, observer-based analyses are susceptible to errors and slow compared to machines, thereby limiting in the amount of data that can be processed. It is desirable to automatise the extraction of information from audio files, a step also referred as audio annotation.

[1] yet, not exclusively. Plant bioacoustics is also a matter of study [28].

This thesis proposes methods to investigate structure in bioacoustic signals. For this two frameworks are proposed. The first concerns the automatic annotation of audio recordings by using supervised machine learning methods. The second concerns a quantitative analysis of temporal and combinatorial patterns in vocal sequences of animals by using non-parametric statistics. These methods are used to investigate vocalisations of two wild living animals — known very little — in their natural ecosystems: lilac crowned parrots (Chapter 7) and pilot whales (Chapter 6). All definitions and methods particular of each framework are explained in detail in the introductory sections of Part I and Part II of this thesis.

Structure of the thesis

Figure 1.1 illustrates the structure of the thesis. Part I describes methods for automatically extracting information out of the recordings. Chapter 2 revises applications of machine learning for bioacoustics and presents the core methods developed in this thesis for automatically annotating recordings. These methods are used later for detecting whale calls (Chapter 3) and classifying them into call types (Chapter 4). Part II delves into structures of animal vocal sequences. Chapter 5 overviews the study of animal vocal sequences and presents the methods used in the later chapters to quantify patterns in vocal sequences of pilot whales (Chapter 6) and of parrots (Chapter 7).

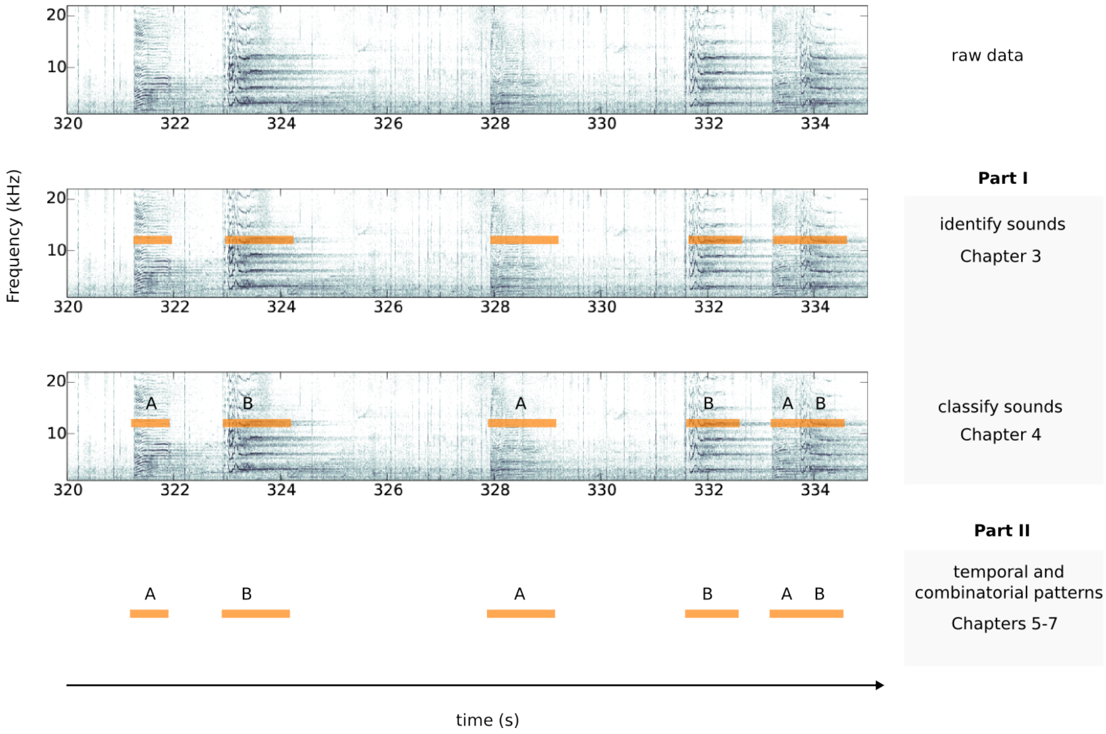


Figure 1.1: Illustration of the challenges addressed in this thesis. The graphs show the spectrogram (a three dimensional representation of sound in terms of its short time power spectral density) of a recording with whale vocalisations. Raw data are recordings of animal sounds for which detection and classification methods are presented in the first part of the thesis. Temporal and combinatorial patterns of the sounds are analysed in the second part of the thesis.

1. PREFACE

Part I

Automatic annotation of bioacoustic recordings

Chapter 2

Background

Bioacoustic studies often start by extracting information out of recordings. This information can expressed in terms of text annotations –a standard and flexible format, handleable by many audio processing platforms. This chapter presents a framework for automatically annotating audio recordings with a supervised machine learning approach. The framework is applied in the next two chapters to address two bioacustic tasks: detecting whale calls (Chapter 3) and classifying them into call types (Chapter 4).

2.1 Why annotating bioacoustic recordings?

Bioacoustic data is a valuable resource in the study of animal communication, ecology, environmental conservation, among other applications (see Chapter 1). The number of projects constantly monitoring ecosystems has increased substantially over the last years skyrocketing the amount of bioacoustic data. Some of these projects include monitoring stations at the sea such as Darewin, orcalab, palaoa observatory, MobySound [41]; crowd sourcing projects such as xeno-canto, bird biodiversity [42], and other institutions like the Alberta biodiversity monitoring Institute from the Cornell lab of ornithology [43]. Not only the monitoring of ecosystems has benefited from the data collection technologies, but also studies of individual species are often performed with automatic recorders such as the Dtags [44] used for marine mammals.

Technology has boosted the ease of data collection but processing methods for extracting information out of these datasets has not paired up. The complex soundscapes of wild bioacoustic scenarios produce datasets that are challenging to process. Humans are good at detecting patterns but this way of processing data imposes temporal and observer bias limitations. Automating the annotation of acoustic recordings would enable to investigating large datasets in a reproducible fashion.

2.2 Machine learning in bioacoustic tasks

The need for processing large volumes of bioacoustic data has driven many proposals for automating this step. A variety of bioacoustic tasks have been addressed automatically, such as: species classification [45], individual identification [46, 47], sound type sorting [48, 49, 50, 51, 52] and sound clustering [53, 54]. Despite these proposals for automating annotation, observer-based analysis are a common approach. Thus this step forms an important bottleneck in bioacoscutic studies. Ravenpro and avisoft are powerful softwares for analysing bioacoustic data, but both require paid licenses, their source code is not available and their graphical user interfaces makes them user friendly but not flexible to large scale applications. Other alternatives like ARBIMON [55] and the orchive [56] are cloud hosted meaning that one needs to export the recordings to a server were these are processed, losing control over the data.

This chapter presents a framework for annotating bioacoustic signals using the Python programming language, deemed appropriate given the popularity of it across the scientific community. The rest of the chapter describes the information flow for annotating recordings using supervised machine learning methods and the architecture of the code that I have developed for this project.

2.3 Framework for the annotation of bioacoustic signals

Annotation files are a convenient way of summarising the information in audio recordings (Fig. 2.1). These are plain text files, which can be loaded and exported by most audio processing software. The aim of the proposed framework is to generate such files from unknown audio recordings (Fig. 2.2). This problem can be approached using supervised machine learning, where a classifier is trained with labelled data —in this case annotated recordings— to distinguish patterns from annotations.

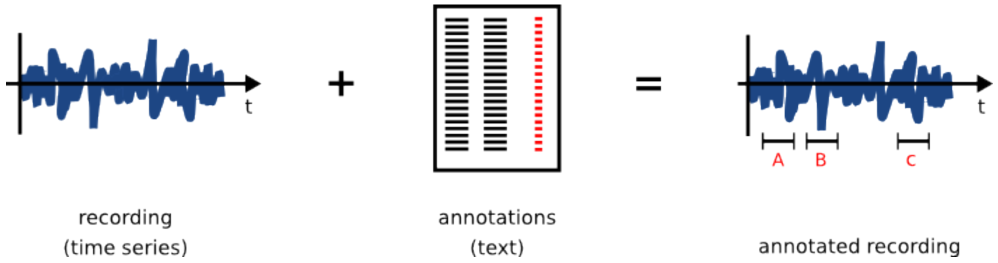


Figure 2.1: Recording annotations. Raw audio files are one dimensional (for mono channel) time series of pressure levels, whose segments can be annotated with relevant information, e.g. animal sounds, represented in the diagram with the letters A, B and C. Annotations are simple text files with the temporal coordinates and labels of the annotated segment. Each row of the text file contains one segment with typically three columns two for the temporal coordinates and one for the label. The temporal coordinates are often the starting and ending time of the segment.

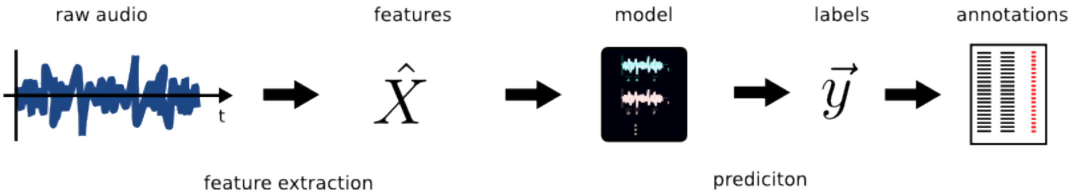


Figure 2.2: Information flow for the generation of audio annotations. Diagram of a system for automatically annotating recordings. Features are extracted from raw audio, represented in the diagram with a matrix $\hat{X}$, and fed to a model (black box) that predicts labels, that are transformed into annotations. Black box represents a model trained with previously annotated data (Fig. 2.3).

Classification tasks are often separated into two steps: feature extraction, where data is put into a suitable representation for the desired task; and training, where an algorithm minimises a chosen evaluation metric to find the best parameter values for the model given the data at hand.

Below I go through each of these steps and describe how they are handled by the code developed for this thesis. Descriptions of the spectral features can be found in Appendix B.

2.3.1 Feature extraction

In the feature extraction step, raw audio is transformed into classification instances. Raw data is handed as collection of audio and annotation files (Fig. 2.4) from which

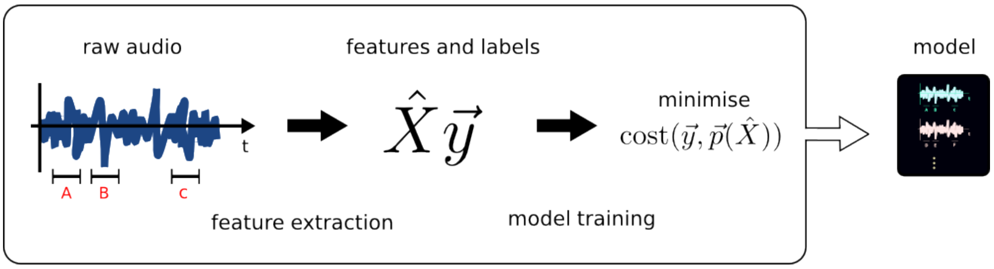
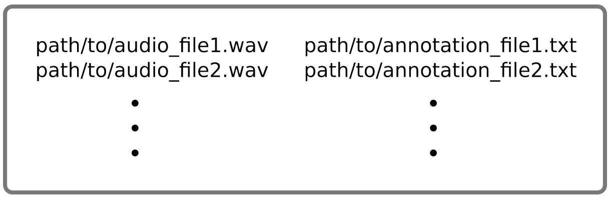


Figure 2.3: Model training. A classifier (black box) is trained with labelled data. Features $\hat{X}$ are extracted from audio and labels $\vec{y}$ from annotations and together are used for training a model that tries to minimise a cost function of the error between the prediction $\vec{p}$ (a function of $\hat{X}$) and the ground truth $\vec{y}$.



```
path/to/audio_file1.wav    path/to/annotation_file1.txt
path/to/audio_file2.wav    path/to/annotation_file2.txt
          •                            •
          •                            •
          •                            •
```

Figure 2.4: Collection of annotated audio files. Text file with two columns, first column has the path to an audio file and second column has the path to its corresponding annotation file.

features are extracted into a data structure of the form $\hat{X}\ \vec{y}$, where $\hat{X}$ is a matrix of m instances (rows) and n features (columns); and $\vec{y}$ a vector of size m with the instance labels. Each row of matrix $\hat{X}$ quantifies one classification instance with a column for each feature. All instances (rows) must have the same number of features (columns).

From the raw audio to the $\hat{X}\ \vec{y}$ form there is room for multiple preprocessing and feature extraction steps. In this thesis, preprocessing steps denote transformations that maintain the data dimensionality, $\mathbb{R}^n \to \mathbb{R}^n$, whereas feature extraction steps refer to transformations that change the waveform representation to a (Euclidean) space of possibly different dimension than the raw signal space $\mathbb{R}^n \to \mathbb{R}^m$, e.g. a spectral representation.

In the context of my code, preprocessing and feature extraction steps are instances of the class `Transformation` which can be stacked using class `transformationsPipeline` (details in section 2.4).

2.3.2 Training

The sklearn python module [57] contains a variety of machine learning functionalities that I combine in my workflow. Sklearn classifiers take data in the form $\hat{X}\ \vec{y}$, out of which a model can be trained calling the fit method. Classifier hyperparameters are tuned with a cross validation `GridSearchCV`. In this thesis, I use support vector machines (SVMs) as classifier of choice. It is well known to be one of the best off-the-shelf classifiers, a detailed description can be found Appendix C.

2.3.3 Evaluation of the classifier performance

Classes are said to be unbalanced when the number of samples between the classes differs. In such cases the metric chosen to evaluate the classifier is crucial, since this is the mean through which we tell the algorithm what our aim is. Classifier evaluation is important, first for tuning the classifiers hyperparameters, typically done with a cross validation set; and second for assessing the final score of a classifier over a test dataset. Some common evaluation metrics are described below.

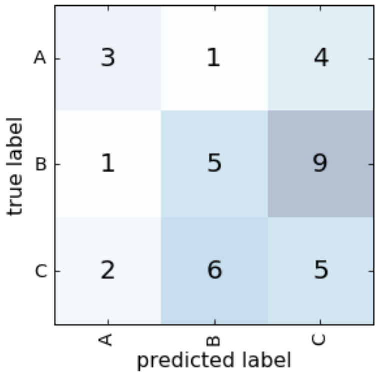


Figure 2.5: Confusion matrix. Example of a confusion matrix of two arrays with three classes.

Confusion Matrix

The confusion matrix is a table that compares the labels of two vectors. In a classification problem, the vector of true labels $\vec{y}$ and the vector of predictions $\vec{p}$. The diagonal elements of the matrix count correctly classified instances whereas off diagonal elements the misclassified instances.

The confusion matrix carries all information about how $\vec{p}$ compares to $\vec{y}$, the problem is however, that being a matrix there is not a single number denoting the quality of fit so we need to define a score which is a scalar value based on this matrix that we aim to maximise. Some common ways of summarising the confusion matrix are presented below.

Accuracy

The accuracy is based on the diagonal elements of the confusion matrix divided by the total number of classified samples. Given the true labels $\vec{y}$ and the predicted labels $\vec{p}$, the accuracy is given by

$$ACC(\vec{y}, \vec{p}) = \frac{1}{m} \sum_{i=1}^{m} \delta_{y_i, p_i}, \tag{2.1}$$

where m is the number of instances and δ is the Kronecker delta that is 1 if both entries are equal and 0 otherwise.

The problem with the accuracy is that it does not distinguishes classes and when dealing with unbalanced datasets can be misleading since classes with fewer instances will be underrepresented in the final score.

The metrics described below are defined for binary labels, typically called positive (p) and negative (n). Multi-label vectors can be mapped into binary vectors by focusing on one class. So that the class of interest is p and all other labels are n.

Precision

The precision measures the relevance of the prediction and is given by the fraction of true positives, T_p, with respect to the number of predicted positives,

$$P = \frac{T_p}{\text{all positive predicted}} = \frac{T_p}{T_p + F_p}, \tag{2.2}$$

where F_p is the number of false positives. High precision indicates low false positive rate and low precision indicates high false positive rate.

Recall

The recall measures the sensitivity to detect the class of interest and is given by the fraction of true positives with respect to the total number of positives

$$R = \frac{T_p}{\text{all the positives data}} = \frac{T_p}{T_p + F_n}, \tag{2.3}$$

where F_n is the number of false negatives. High recall indicates low false negative rates and low recall indicates high false negative rates.

F_1 score

The F_1 score is the harmonic mean of precision and recall

$$F_1 = 2\frac{P \times R}{P + R}. \tag{2.4}$$

In this thesis scores are reported in percentage scale. So instead of ranging from zero to one, the equations above are multiplied by 100, thereby ranging the scores from zero to 100, with 100 the highest score.

2.4 Code architecture and core programming tools

The developed code integrates audio signal processing tools of the python module librosa [58], with machine learning tools from the python module sklearn [57], in order to effectuate machine learning tasks on audio data. The modular architecture of the code facilitates the execution of experiments (Fig. 2.6) by allowing to easily exchange the datasets, feature extraction methods and machine learning estimators. Below I describe the core programming classes used for the classification experiments, in a top down order.

WARNING: The following section contains considerable amounts of Python slang.

Experiment class

The class `experiment` bounds the settings for the machine learning task (experiment): (1) input data, through a collection of annotated audio files, (2) feature extraction settings through a `TransformationsPipeline` (3) the classification settings through sklearn's `pipeline` and `GridSearchCV` and (4) the path to an output file where the classification performance scores are printed. An experiment can be iterated to scan different combinations of parameters, e.g. feature extraction settings and classification parameters.

Transformation pipeline

The class `TransformationsPipeline` contains instructions —processing steps— for extracting features. For instance, three processing steps can be: normalise the waveform, apply a band pass filter and compute the spectrogram. Some attributes of this class are: a list of the names of processing steps, a callable that can be used to extract the features, a string with the feature extraction settings, among others. The class can be initialised with the function `makeTransformationPipeline` which takes a list with the processing steps. The processing step are handled as tuples with two entries. The first entry is a string used to identify the processing step, e.g. "normalise_waveform" and the second entry contains an instance of the class `Transformation` used to apply the processing steps.

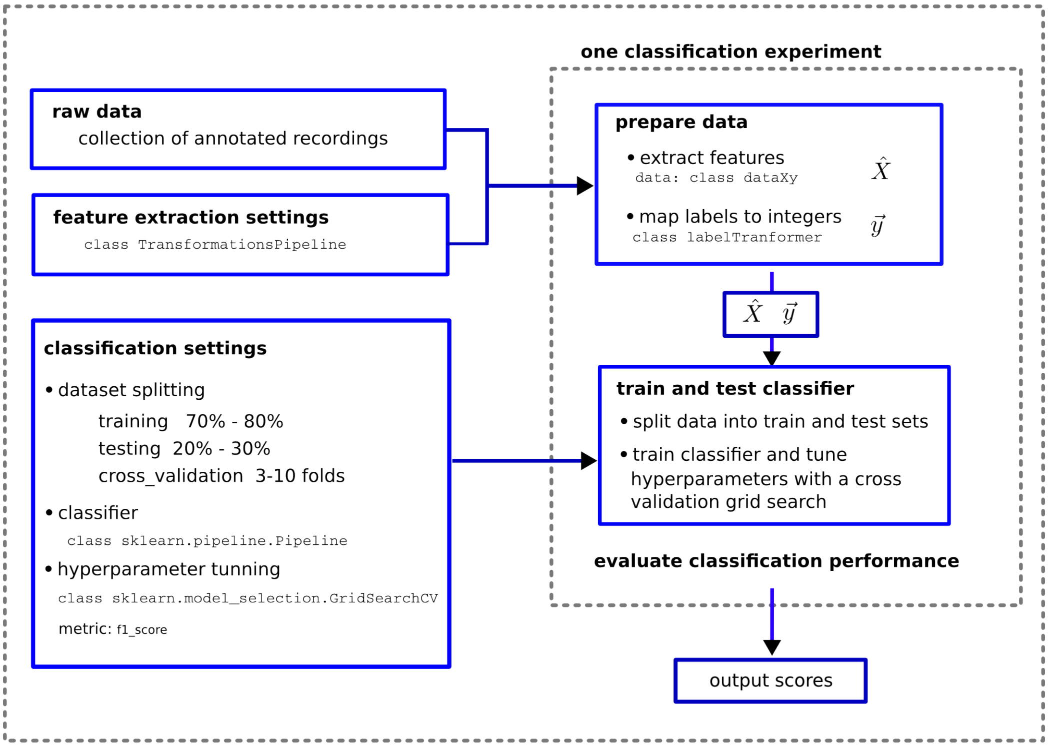

one classification experiment
raw data
collection of annotated recordings
feature extraction settings
class TransformationsPipeline
prepare data
• extract features
data: class dataXy
$\hat{X}$
• map labels to integers
class labelTranformer
$\vec{y}$
$\hat{X}$ $\vec{y}$
train and test classifier
• split data into train and test sets
• train classifier and tune hyperparameters with a cross validation grid search
evaluate classification performance
output scores
classification settings
• dataset splitting
training 70% - 80%
testing 20% - 30%
cross_validation 3-10 folds
• classifier
class sklearn.pipeline.Pipeline
• hyperparameter tunning
class sklearn.model_selection.GridSearchCV
metric: f1_score

Figure 2.6: Diagram illustrating the information flow for a classification experiment with audio data. A classification experiment is regarded as the process of training and testing a classifier, given a dataset (raw data), and the settings for feature extraction and classification. An experiment can be initialised and carried using the class `experiment` that bounds the experiment settings with an output file meant to keep records of the classification scores. Settings can be easily modified to evaluate the performance of a classier under different conditions, e.g. using different feature extraction parameters.

Transformation

The class `Transformation` defines processing steps used for extracting features. For instance, one can use a `Transformation` to set-up a band pass filter. A `Transformation` can be initialised with a tuple of two entries. The first entry should have a string with the name of the processing step and in the second entry its settings as a dictionary of key word arguments (`kwargs`). Going back to our example with the band-pass filter, the first entry would be "band_filter" and the second entry a dictionary with the filtering bands.

Chapter 3

Automatic detection of whale calls with spectral features

3.1 Introduction

Many algorithms for processing bioacoustic recordings focus on the classification of sounds from presegmented data. However, a real speed-up in the processing of large scale datasets can only be achieved by eliminating manual steps. The aim of this chapter is to automate this step by training a model to segment whale calls. In terms of the annotations, this means generating the temporal coordinates of the calls.

In this chapter, I adjust the framework from Chapter 2 to train support vector machine (SVMs) classifiers (Appendix C) to segment pilot whale calls from four recordings collected in the wild. Classification performance is compared between two spectral features (Appendix B) mel-spectrum and MFCC. For each of these feature representations, a range of parameters are scanned in order to assess their influence on the classification performance.

3.2 Dataset

The dataset consists of four audio files (tapes named 111, 113, 114 and 115) with pilot whale sounds recorded in the wild. Raw data is in *wav* format and has a sampling rate of 48 kHz. Environmental sounds cannot be controlled in the wild and made our recordings highly heterogeneous in terms of the sources of background noise, signal to noise ratio and the proportion of call segments in the sample. These factors affect the

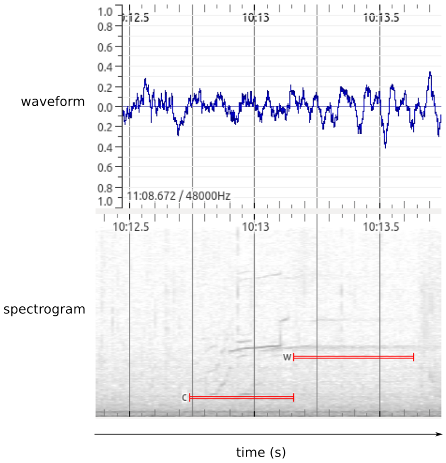


Figure 3.1: Waveform and spectrogram. Low signal to noise ratio impedes extracting the calls by simply thresholding the waveform (upper panel). Spectrogram shows a whale call labelled as c and the echo of the call labelled as w.

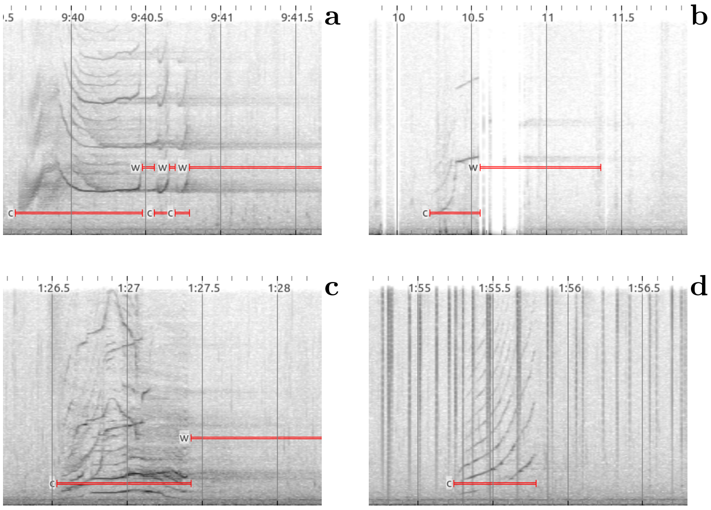


Figure 3.2: Challenges in the dataset. Examples of recordings with (**a**) echoes labelled as w, (**b**) missing signal, (**c**) overlapping calls low signal to noise ration and (**d**) presence of other whale sounds such as clicks and buzzes.

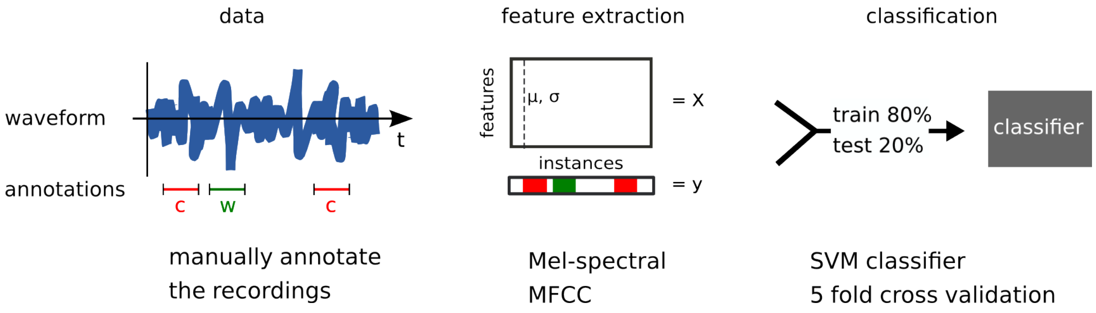


Figure 3.3: Information flow for training a classifier to detect whale calls. Raw audio is annotated with segments of calls (c) and weak sounds (w). Classification instances are frames of the recording for which spectral features are extracted and summarised for each spectral band with the mean and the standard deviation. A support vector machine (SVM) classifier is trained with 70% of the data and tested over the rest. Classifier hyperparameters are tuned with a 5-fold cross validation.

quality of the recordings in different ways. To assess their effect on the classification performance, each recording was treated as an independent dataset.

Signal to noise ratio was generally low impeding to extract the whale sounds by simply thresholding the spectrogram (Fig. 3.1) as it is often done in controlled environments like labs or aquariums. Background noise sources included engines, vessels and sounds from other animals. Different physical constraints also affected the quality of the recordings such as the acoustics due to the rugged relief of the fjords; and the distance, deepness and direction of the whale with respect to the hydrophone.

The dataset was manually annotated using audacity [59]. Segments with whale calls were labelled with a *c* (Fig. 3.2). Because the aim of the classier is to extract the whale calls, other whale sounds like clicks and buzzes were regarded as background noise. Weak tonal sounds such as low intensity calls and call echoes occurred frequently in our dataset. Their acoustic properties are similar to those of calls, so they were labelled as another class with the letter *w* (Fig. 3.2). Unannotated sections were regarded as background noise and were automatically tagged with the label *b*.

3.3 Design of the machine learning task

The aim of the classifier is to compare the performance of two spectral feature representations: mel-spectrum and MFCC; in the task of extracting pilot whale calls from a recording (Fig. 3.3). This is done training a classifier with frames from recordings labelled with either of the three classes: *c* for calls, *w* for weak sounds and echoes, and *b* for the rest of the recording.

Feature extraction

The mel-spectrum and the MFCC features are both spectral representations since they are based on a Fourier transform of the raw signal. The spectral resolution of these representations is mediated by the number of mel-filters and the number of MFCCs, here referred to as **frequency bands**. Experiments varying the number of frequency bands are carried out to investigate their effect on the classification performance.

Three steps were involved in the feature extraction. First, waveforms were normalised by the maximum amplitude. Then spectral features were extracted (Appendix B) using an FFT window of 512 samples with 0% overlap. Finally, features were temporally summarised computing the mean and the standard deviation for each frequency band over a number of **summarisation frames** (Appendix B). In addition to the number of frequency bands, the number of summarisation frames are varied in the experiment.

The proportion of *b* samples exceeded the other two classes, sometimes in more than one order of magnitude. Unbalances in the dataset can yield bad results. Due to the aim to detect whale calls, the number of samples of classes *b* and *w* was balanced to number of samples of class c. This was done by randomly disregarding samples from classes *b* and *w* so that their numbers match the ones of *c* samples.

Classification

A support vector machine classifier (SVM) (see Appendix C) with linear kernel was trained with 80% of the data. Classifier hyperparameters were tuned with a 5 fold cross validation. Classifier performance was assessed with the F_1 score for class c.

Experimental parameters

Temporal and spectral resolution was varied to assess its effect on the classification performance. The number of frames ranged per instance from 2 to 40 and the number of frequency bands form 1 to 20.

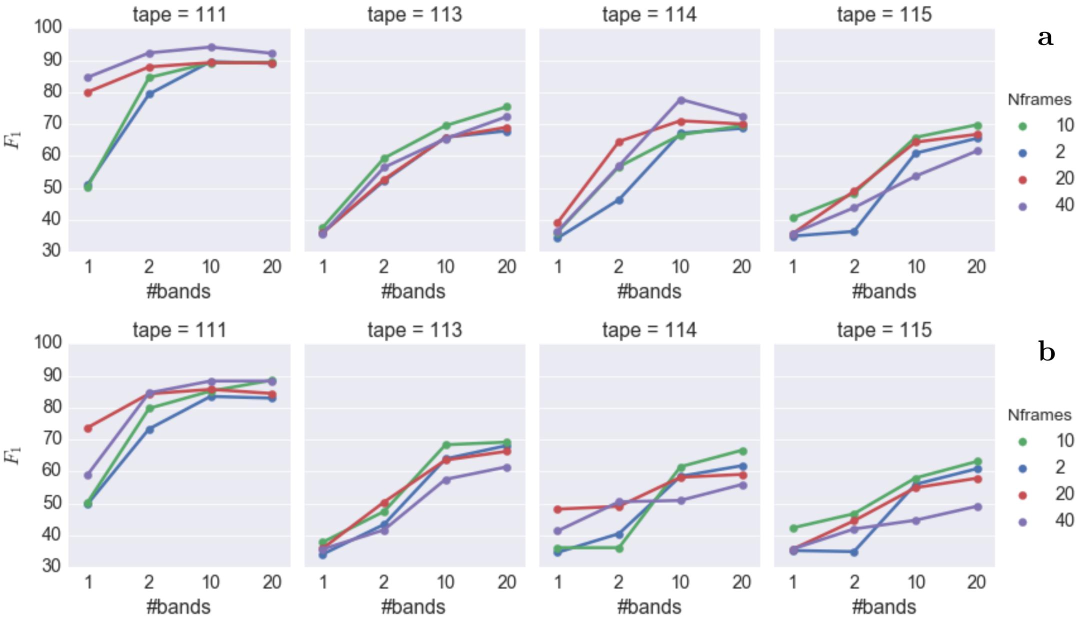


Figure 3.4: Classification performance with spectral features. F_1 score of class c (calls) as a function of the spectral bands for the four tapes with (**a**) mel-spectral and (**b**) MFCC features. Score expressed in percentage scale. Colours indicate the different number of frames per classification instance.

3.4 Segmentation of whale calls with spectral features

I found that the classification performance with both feature representation increases with the spectral resolution, yielding highest scores between 10 to 20 frequency bands (Fig. 3.4). Contrary to this parameter, the number of frames did not influence the classifier performance with a clear trend. However, care should be taken when comparing the scores for the different number of frames since the number of samples decreases with the number of frames.

Mel-spectral features yielded better scores than MFCC features (Fig. 3.5). This was observed consistently for the four tapes and different parameter combinations. The performance of the fitted model depended strongly on the dataset. The scores of both feature representations varied more than 20% between the different tapes. Regardless of the feature representation and the parameter combination, tape 111 always yielded the highest score among the tapes. This was due to the small proportion of w samples this tape has (Fig. 3.6).

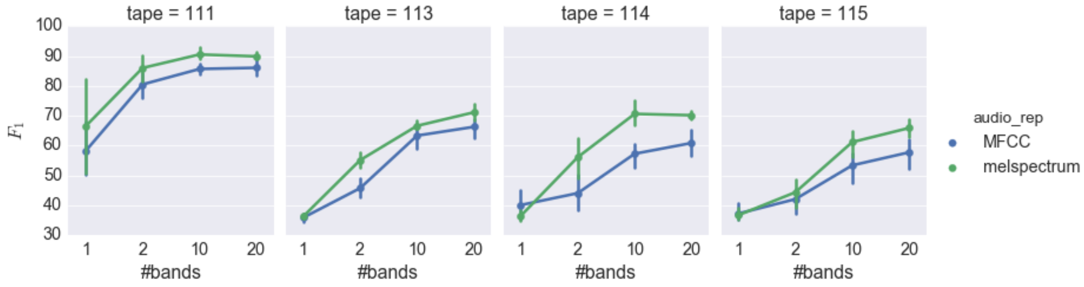


Figure 3.5: Comparison of the classification performance with mel-spectral and MFCC for each tape. Classification performance, measured as the F_1 score of class c (calls) as a function of the number of spectral bands, being the number of mel-filters for the mel-spectral features and the number of MFCCs for the MFCC features. Score expressed in percentage scale.

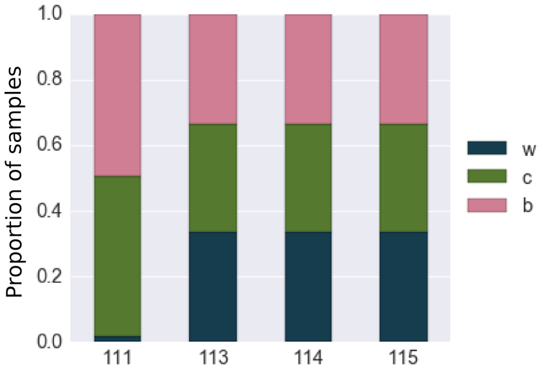


Figure 3.6: Sample composition of the datasets. Proportion of the samples of each class weak sounds (w), calls (c) and background noise (b) in each tape.

3.5 Summary and discussion

Classifier performance was found to be susceptible to the chosen features, their combination of parameters and the dataset. Differentiating calls from background noise yielded better scores with mel-spectral features than with MFCC features. Because differentiating calls from background noise is easily done from the power spectral density, higher order structures like the periodicity in the harmonics —well captured by the MFCC features— are not relevant for this task.

The fitted model depended highly on the dataset and the composition of samples of each class. By training models for each tape independently we were able to identify the proportion of weak sounds in the sample to be the major challenge for detecting whale calls successfully. The acoustic properties of class w lay between the two other classes and distinguishing these samples is challenging even for the human eye, thus, it is not surprising that higher proportions of these samples have yielded worse scores. This stresses how the scores of a classifier depend on the characteristics of the dataset.

Besides MFCC and mel-spectral features other spectral features could have been tried, like pure spectrogram or a cepstrogram. Pure spectral features are high dimensional which imposes two difficulties over lower dimensional features: (1) they are more vulnerable to overfitting and (2) training models takes more time. As for the cepstrogram, this representation is similar to the MFCC features in that both compress the periodicities of the power spectral structure with a second power spectral transformation. Given that the mel-spectrum outperformed the MFCCs, it is unlikely that the cepstrum would outperform the mel-spectrum.

Chapter 4

Automatic classification of whale calls with spectral features

4.1 Introduction

Many toothed whales such as orcas and pilot whales produce sounds, named calls, with distinctive spectro-temporal characteristics. Calls can be sorted into types according to their acoustic characteristics. These types have been found to reflect the social structure of many marine mammals [60], and are a frequently studied object of these animals. It would be desirable for naturalists to automate the sorting of call types, speeding up the process and preventing human errors. In terms of the annotations (Chapter 2), this means generating call type tags for the segments identified previously in Chapter 3.

In this chapter, I train support vector machine (SVM) classifiers (Appendix C) to distinguish 71 call types from pilot whales using spectral features (Appendix B). We compare the classifier performance using three spectral features —cepstrum, mel-spectrum and MFCC— based on the analysis defined within the framework proposed in Chapter 2. For each of these feature representations, a range of parameters is scanned in other to assess their influence on the classification performance. A second batch of experiments is carried out over a benchmark dataset of killer whale calls to test the robustness of the features in a different dataset.

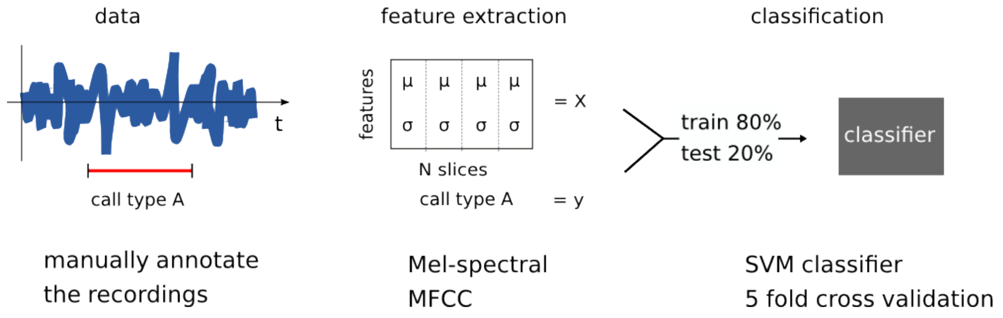


Figure 4.1: Information flow for training a classifier of call types. Audio files with annotated call types are transformed into classification instances. Spectral features are extracted from each call and sliced into N segments of equal length, figure shows $N = 4$. Two kinds of spectral features are tried here: mel-spectral and mel Frequency Cepstral Coefficients (MFCC). Features are summarised with the mean (μ) and the standard deviation (σ) for each spectral band. A support vector machine (SVM) classifier is trained with 80% of the data and tested over the rest. Classifier hyperparameters are tuned with a 5-fold cross validation.

4.2 Design of the machine learning task

The aim of the task is to compare three spectral feature representations —cepstral, mel-spectral and MFCC— in terms of their performance in classifying whale calls with an SVM. Spectral representations depend on a series of parameters that control their temporal and spectral resolution, e.g. the window size of the fast Fourier transform (FFT). Thus we carry out experiments to scan combinations of these parameters to assess their influence on the classification task. Details on the feature extraction and classification settings are explained below.

Feature extraction

Classification instances were prepared through a three step feature extraction procedure: (1) waveforms are normalised by the maximum absolute value, then (2) spectral features are extracted (details of the features in Appendix B) and (3) different instance lengths —due to differences in the duration of the audio files— are normalised. The last step is important since the classifier can only compare vectors of the same size. Length normalisation is done by slicing each instance into equally spaced segments and computing the mean and the standard deviation of each frequency band in each

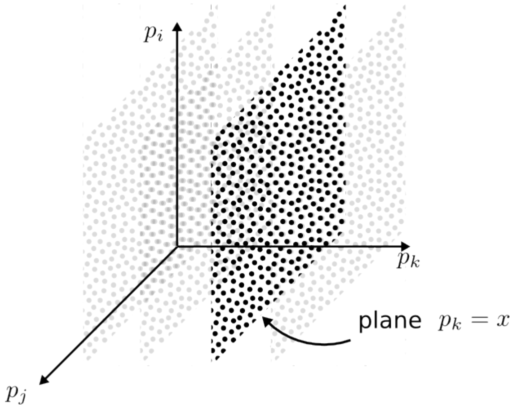


Figure 4.2: Phase space of scanning parameters. Each dimension represents a parameter, e.g. the size of the FFT, the number of slices, or the number of spectral bands and each dot a combination of parameters. The planes indicate sets of points with parameter p_k fix to value x.

segment (Fig. 4.1). The **number of slices** is one of the parameters scanned in the experiments. Parameters such as the number of coefficients, or number of quefrencies, of the cepstrum; the number of mel filters of the mel spectrum; and the number of MFCCs of the MFCC features; tune the spectral resolution. Here these parameters are referred to as **spectral bands** and combination of them are scanned in the experiments.

Classification and evaluation

Features are used to train a support vector machine classifier with radial basis function Gaussian kernel. Classifier hyperparameters such as the penalty C, and the kernel coefficient γ were tuned with a 5-fold cross validation grid search algorithm from sklearn [57] (Appendix C). The dataset was split using 80% for training and the rest for testing.

Because the classes are unbalanced the accuracy is not a suitable metric for evaluating the classifier's performance. Instead we use the macro average of the F_1 score, which averages the score using the same weight for all the classes

$$\langle F_1 \rangle_c = \frac{1}{N} \sum_{i=1}^{N} F_1(c_i), \tag{4.1}$$

where $F_1(c_i)$ is the F_1 of the i-th call class c_i and N the the number of classes.

For each spectral representation several combination of parameters were tried (Fig. 4.2). Given a metric S (Chapter 2), we define $\Delta_{p_i}(x)$ as the mean of S in the subspace of

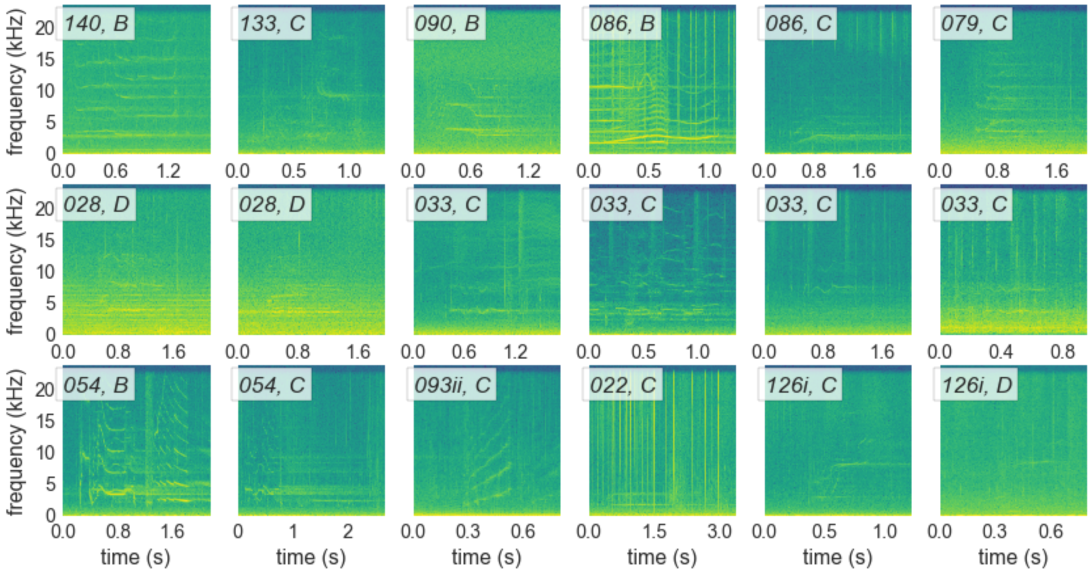


Figure 4.3: Pilot whale calls. Randomly selected samples from the pilot whale catalogue. Labels in the top left indicate the call type, with an alphanumeric tag, and the quality for the recording, with alphabetical ranking from A (best) to D (worst).

scanned parameters with p_k fixed to x,

$$\Delta_{p_i}(x) = \langle S(\vec{p}|_{p_i=x})\rangle. \tag{4.2}$$

Δ_{p_i} is a function of the value x so I use the range of Δ_{p_i} to assess the **influence of a parameter** p_i along its scanned values.

4.3 Classification of pilot whale calls

The experimental set-up described above was used to classify calls from pilot whales. This section presents results of the classifier performance for each of the spectral features cepstral, mel-spectral and MFCC. We start describing the dataset and then move on to the results.

4.3.1 Dataset

The dataset consists of 3885 audio files of ca. 1 s with pilot whale calls extracted manually from longer recordings with a sampling rate of 48 kHz. The identified calls were inspected in terms of their spectro-temporal features —frequency modulation,

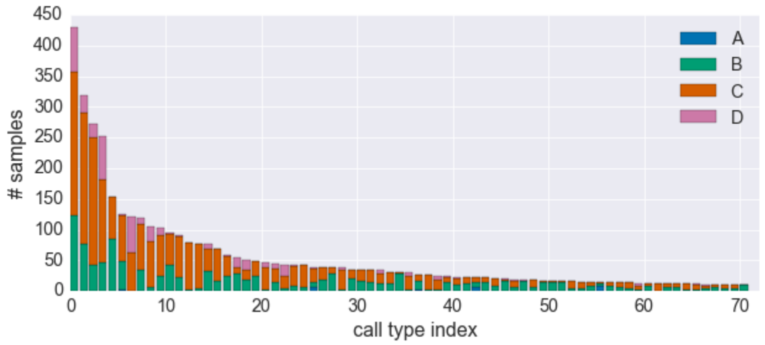


Figure 4.4: Distribution of call samples. Number of samples of each call type. Colours indicate quality of the recording ranked alphabetically from A (best) to D (worst).

parameter, p_i	scanned parameters, x	range of $\Delta_{p_i}(x)$
FFT window size	256, 512, 1024	1.8
# quefrencies	1 - 39	41.9
# slices	1 - 15	8.8

Table 4.1: Cepstral feature parameters and their influence on the call classification task. Parameter influence measured as the range of Δ_{p_i} (Eq. 4.2) with the macro average of F_1 as metric.

distance between harmonics and presence of tonal and noisy elements— and placed into 71 call categories by Dr. Vester. Details of the class definition are reported here [61]. Quality of the recordings was assessed “manually” and ranked alphabetically in either of the four qualities from A (best) to D (worst). Figure 4.3 shows the spectrogram of 18 samples randomly drawn from the dataset. Each class has at least 10 samples, yet the number of samples for each call type is highly unbalanced. The most frequent call class has more than 400 samples while one third of the call classes has less than 20 samples (Fig. 4.4). Additionally, most samples are poor quality recordings with almost 80% belonging to the lowest two qualities (Fig. 4.4).

4.3.2 Cepstral features

Classifying calls with cepstral features yielded a maximum F1-score of 53.6±0.1 (5-fold cross validation), with 35 cepstral bands, a Fourier transform window of 512 samples with 50% overlap and 5 slices. Below, I describe the performance of the classifier under

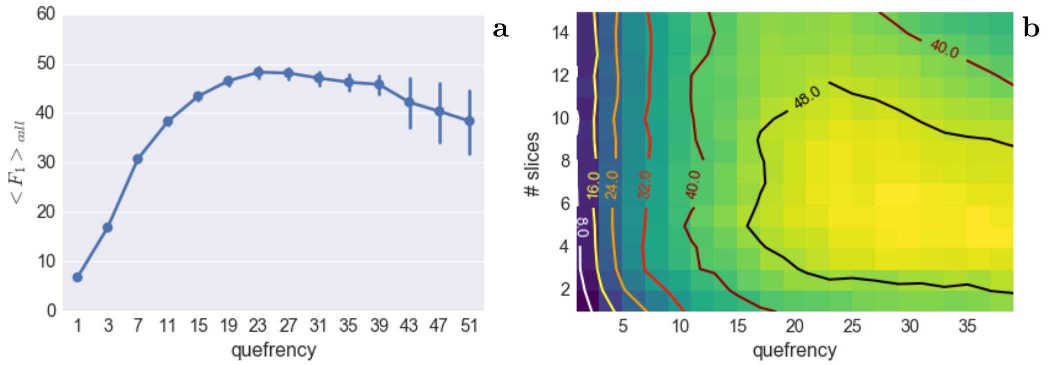


Figure 4.5: Classification performance with cepstral features. **a**, Classification score as a function of the number of cepstral coefficients. **b**, Classification score as a function of the number of cepstral coefficients and the number of slices. Using the macro average of the F_1 score and a 5-fold cross validation.

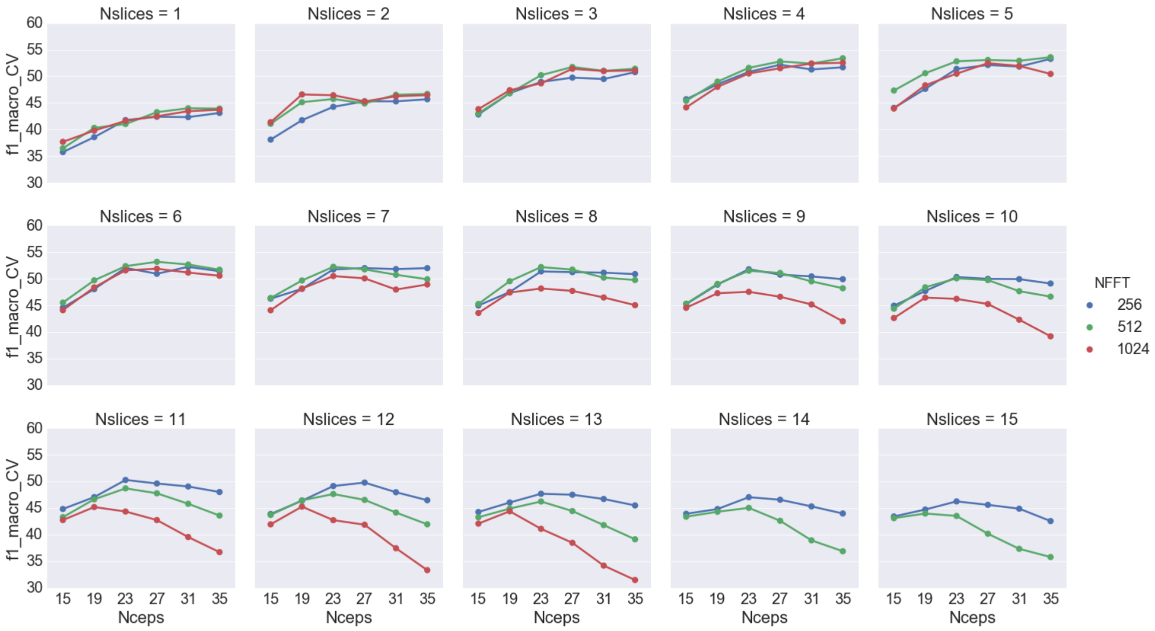


Figure 4.6: Influence of the number of cepstral bands. Classification performance as a function of the number of cepstral coefficients (> 15) for different number of slices and the three FFT window sizes. Classification performance measured with the macro average of the F_1 score and a 5-fold cross validation.

parameter, p_i	scanned parameters, x	range of $\Delta_{p_i}(x)$
FFT window size	256, 512, 1024	4.4
# mel filters	1-96	48.6
# slices	1-15	14.2

Table 4.2: Mel-spectral parameters and their influence on the call classification task. Parameter influence is measured as the range of Δ_{p_i} (Eq. 4.2) with the macro average of F_1 as metric.

different parameter combinations.

Table 4.1 shows the scanned parameters and their effect on the classification performance. Among the scanned parameters, the number of quefrencies influenced most, then the number of slices and least the size of the FFT.

The classification performance increases with the number of cepstral bands, stagnating around 15 bands (4.5 (A)). Above this point, the influence of other parameters, the size of the FFT window and the number of slices become more important. Figure 4.5 (B) shows the relation between the number of quefrencies and the number of slices, in which, the region with 4 to 7 slices and more than 20 cepstral coefficients shows highest performance.

The size of the Fourier transform has a small effect in the classification performance, being only relevant when the number of slices is: greater than 7 for a FFT window of 1024 samples; and greater than 10 for a FFT window of 512 samples (Fig. 4.6). Large FFT windows compromise the temporal resolution, yielding few temporal samples and an inefficient temporal summarisation. However, in the optimal region of 4-5 slices the size of the Fourier window has no important effects.

4.3.3 Mel-spectral features

Classifying calls with mel spectral features yielded a maximum F1-score of 65 ± 1 (5-fold cross validation) with an FFT window of 1024, 76 mel-filters and 3 slices. Table 4.2 shows the scanned parameters, their variation range and their effect on the call classification task with mel-spectral features. The number of mel-filters is the most influential of the scanned parameters.

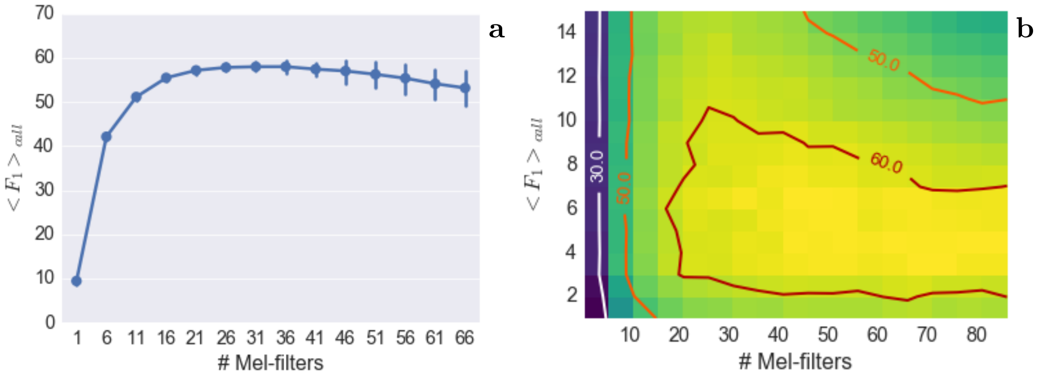


Figure 4.7: Classification performance using mel-spectral features. **a**, Classification score as a function of the number of mel-filters. **b**, Classification score as a function of the number of mel-filters and the number of slices, with a Fourier window of 512 samples, 50% of overlap. Classification performance measured with the macro average of the F_1 score and a 5-fold cross validation.

The classifier performance improves with the number of mel-filters; beyond the stagnation at ca. 16 filters the other parameters become important (Fig. 4.7a). The number of slices displays an optimal region between 3 and 7 slices (Fig. 4.7b). The FFT window is the least influential parameter. Its effect in the optimal region for the number of mel-filters (> 16) and the number of slices (between 2 and 9) is shown in Fig. 4.8. The smallest FFT window (256 samples) does not capture enough frequency resolution, having a maximum score almost 5% below the maximum score obtained with the larger FFT windows. The FFT windows 512 and 1024 yielded best scores with 4 and 5 slices. Beyond this point larger FFT windows are less effective capturing meaningful information.

4.3.4 MFCC features

Classifying calls with MFCC features yielded a maximum F1-score of 73.28±1.4 (5-fold cross validation), with 128 MFCC, 37 mel-filters, a Fourier window of 512 samples a 50% overlap and 4 slices. Table 4.3 shows the varied parameters and their influence in the classification task. The number of MFCC has the strongest effect in the classification performance, with an influence of 58% over the classification score, followed by the number of slices, then the size of the FFT window and at the end the number of mel-spectral filters.

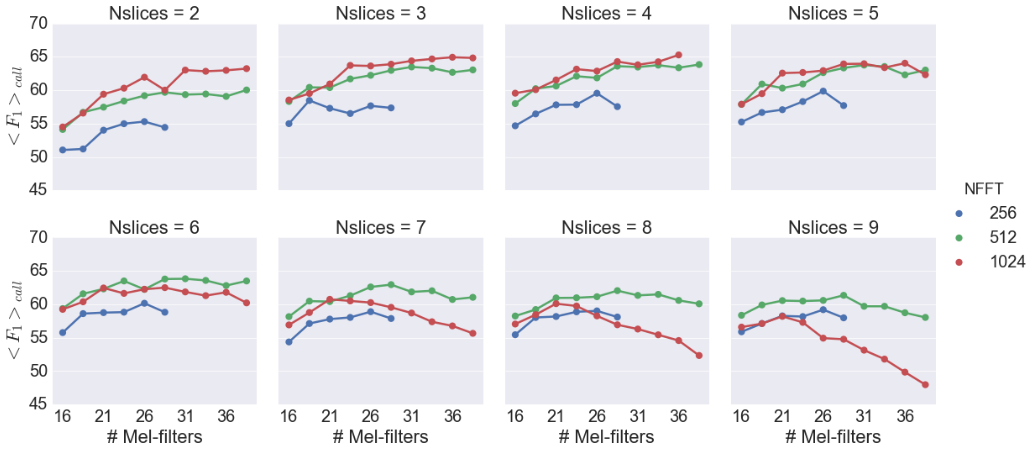


Figure 4.8: Influence of the number of mel-filters. Classification performance using mel-spectral features as a function of the number of mel-filters (> 16) for different number of slices and the three FFT window sizes scanned.

parameter, p_i	scanned parameters, x	range of $\Delta_{p_i}(x)$
FFT window size	256, 512, 1024	5.6
# mel filters	32, 64, 128, 256	3.5
# slices	1-15	9.5
# MFCC	1-39	58.4

Table 4.3: MFCC parameters and their influence on the call classification task. Parameter influence is measured as the range of Δ_{p_i} (Eq. 4.2) with the macro average of F_1 as metric.

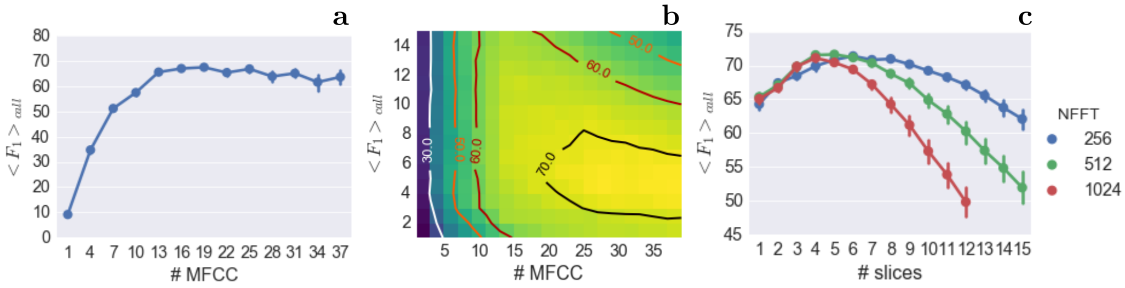


Figure 4.9: Classification performance using MFCC features. a, Classification score as a function of the number of MFCC. **b**, Classification score as a function of the number of MFCC and the number of slices, with a Fourier window of 512. **c**, Classification score as a function of the number of slices for the three FFT window sizes with more than 15 MFCCs. Classification score is given by the macro average of the F_1 score with a 5-fold cross validation.

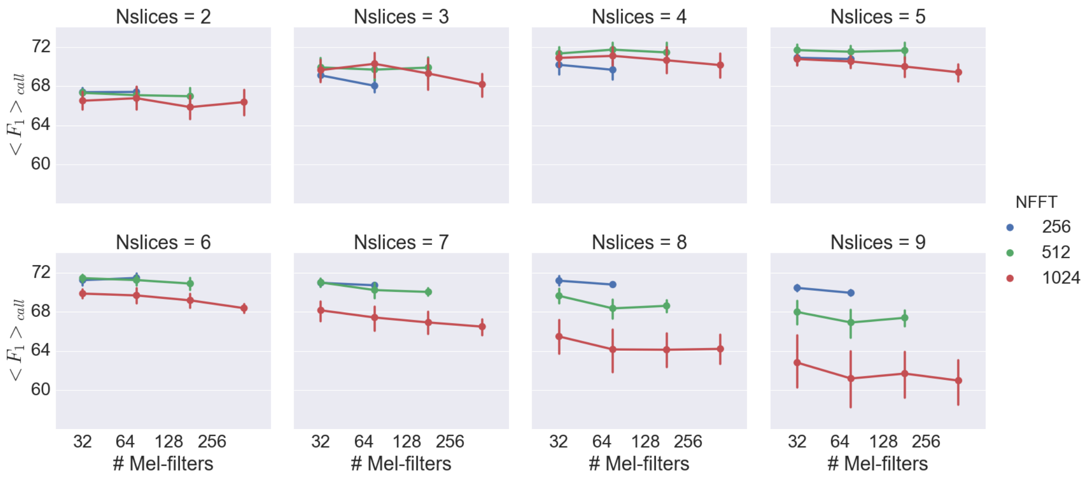


Figure 4.10: Influence of the number of mel filters. Classification performance using MFCC features as a function of the number of mel-filters, for different number of slices and the three FFT window sizes. Macro average of the F_1 score with a 5-fold cross validation was used to estimate classification performance.

Classification score improves with the number of MFCCs up to 20 where it stagnates and the other parameters become more important (Fig. 4.9a). Beyond 20 MFCCs, the number of features starts playing an important role and the number of MFCCs must be traded off with the number of slices. This can be appreciated in Fig. 4.9b where the classification score, as a function of the number of MFCCs and the number of slices, displays a region of optimal performance between 3 and 7 slices and more than 20 MFCCs.

All FFT window sizes yielded scores above 80% (Fig. 4.9c). The number of mel-filters has no important effects on the classification performance.

4.4 Classification of killer whale calls; benchmark dataset

In the previous section was obtained that MFCC features outperformed mel-spectral, and cepstral features. Now we test the robustness of this result comparing the performance of the two top representations —MFCC and mel-spectral— on the classification of killer whales calls from a benchmark dataset.

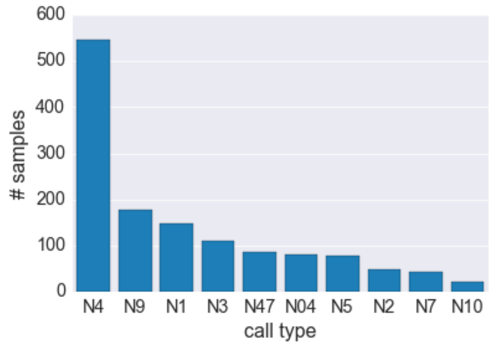


Figure 4.11: Counts of the call samples in the orchive call catalogue. Call classes are tagged with an N and a number. Call types are sorted according to their frequency.

The orchive dataset

The orchive [62] is an open dataset (available at http://data.orchive.net/) with recordings of sounds from northern resident killer whales from the western coast of Canada. Sounds include calls, whistles buzzes among other whale sounds. Releasing the dataset was a collective effort of OrcaLab, who collected the recordings and Steven Ness who, at the time at University of Victoria, prepared the dataset as part of his PhD thesis [56]. The catalogue consists of individual audio files of ca. 1 s with sampling rate of 44kHz. Calls are annotated according to John Ford's call catalogue [63], where call types are labelled with a capital N and a number, indicated in the file name.

The dataset was prepared parsing call types from file names and keeping the subset of calls with at least 10 samples per call type. This yielded 1340 samples with 10 categories (4.11) which I used for the classification task. Figure 4.12 shows some samples of the dataset.

Machine learning task

The tasks were carried out scanning combinations of parameters in the optimal performance regions identified in section 4.3. For both features we use FFT window of 512 samples with 50% overlap and tried 2, 4 and 5 slices. For the mel spectral features we use 32 and 54 mel-filters and for the MFCC we use 31 and 36 MFCC with 64 mel-filters which in section 4.3 were observed not to influence the classification performance significantly. I use 80% of the calls for training the classifier and the rest for testing it.

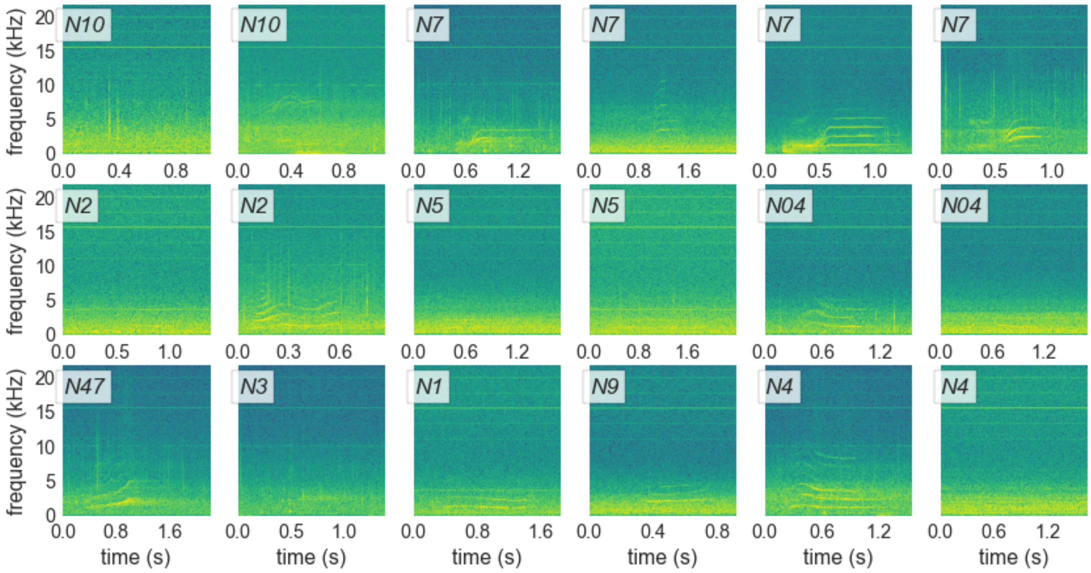


Figure 4.12: Orchive calls. Randomly selected samples from the orchive call catalogue. Labels in the top left indicate the call type.

Results

The classification performance obtained with the orchive dataset agreed with our results from the previous section that MFCC outperform mel-spectral features. The scores obtained for the pilot whales are higher than for the orchive, yet the superiority of the MFCC over the mel-spectrum was confirmed by the cross validation F_1 score and the 4 metrics over the test set —accuracy, precision, recall and F_1— which consistently scored higher for the MFCC than for the mel-spectrum (Fig. 4.13).

4.5 Summary and discussion

Out of the three spectral representations we tried here, MFCC features performed best classifying whale calls with a support vector machine. For the pilot whale classification, MFCC outperformed mel-spectral features by almost 10% and mel-spectral features outperformed cepstral features by 10%. The superiority of the MFCC features was confirmed classifying killer whale calls from the orchive dataset.

Feature extraction parameters influenced the performance of the classification. Among them, the most influential parameters were the number of frequency bands and the

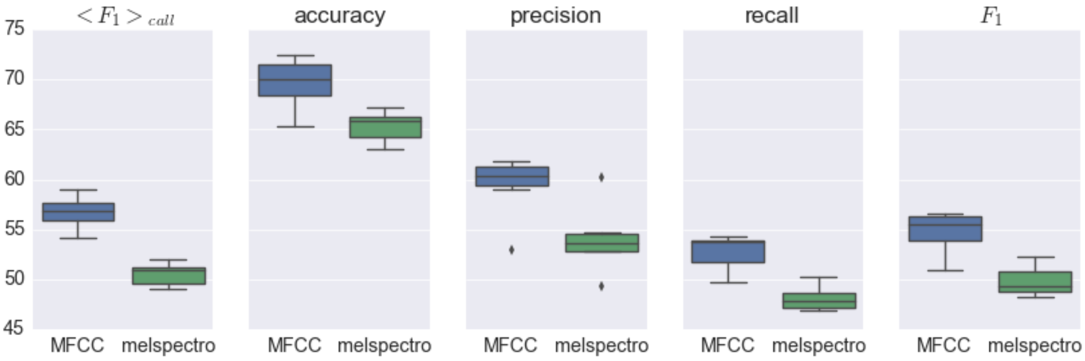


Figure 4.13: Classification scores obtained with the orchive dataset. All scores shown are macro-averages (Section 2) of the call types, except for the accuracy whose definition is independent of the classes. Features were extracted with FFT window of 512 samples, 50% overlap, and 2, 4 and 5 slices. For the mel-spectrum 32 and 64 filters were used and for the MFCC 31 and 36 coefficient over a 64 mel-filtered spectrogram.

features	# spectral bands	# slices	best score	
			pilot whales	orchive
cepstrum	> 20	3-5	53.6 ± 0.1	NA
mel-spectrum	> 20	3-7	65 ± 1	52 ± 10
MFCC	> 15	4-7	73 ± 1	59 ± 7

Table 4.4: Summary of results. Best classification scores and region of optimal parameters for the classification of pilot whale calls and killer whale calls from the orchive dataset.

number of temporal slices. MFCC and mel-spectral allowed better temporal resolutions than the cepstral features as indicated by the number of slices of highest scores —which for the first two were higher than for the later. This may be the reason why MFCC and mel-spectral outperformed cepstral features. The superiority of the mel-scale over the linear scale does not mean that whales perceive frequencies according to the mel-scale. However, for the classification task carried out, the mel-scale proved to be more effective than the linear scale.

Part II

Quantifying animal vocal sequences

Chapter 5

Background

Many animals combine vocal units (e.g. parrot notes, whale calls, dolphin whistles) into sequences that can carry information about their identity, behaviour and context. Vocal sequences feature two main characteristics: (1) timing and (2) combinatorial. Here we propose methods to quantify animal vocal sequences using a non-parametric statistical approach. These methods are used to investigate vocal sequences of pilot whales (Chapter 6) and parrots (Chapter 7).

5.1 Why quantifying animal vocal sequences?

Humpback whales are perhaps the most famous whales. It was after them that the popular term "whale song" was coined [65]. Their songs have literally brought this species to the stars, featuring in the interstellar album "Voyager Golden Record" [66] on board both Voyager spacecraft launched by NASA in 1977 (Fig. 5.1). Humpback whales, however, have not always been so dear to humans. Only thirty years before the golden record these animals were mere marine beasts that supplied humans with oil. So, what made these whales and their songs so popular? Humpback whale songs were first recorded by an antisubmarine warfare station during World War II. Under the suspicion of coming from a Soviet submarine [67], these sounds were classified as top secret and only identified as whale sounds a decade later [68]. Yet, this was not what led humpback whales to the stars. Many animals produce sounds after all. It was not until the 70s that the complex structure of their songs was recognised [69]. Humpback whale songs are made up of units that are combined and repeated in a hierarchical

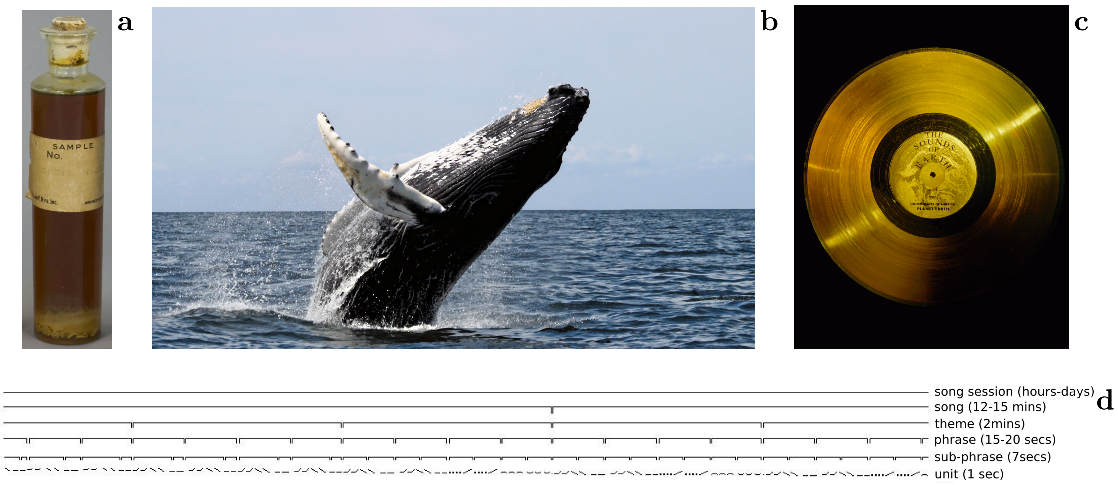


Figure 5.1: Humpbacks whales through history. **a**, bottle of whale oil. **b**, Humpback whale breaching. **c**, The Voyager Golden Record at both Voyager spacecraft. **d**, Diagram of the hierarchical structure of an idealised humpback whale song [64]. Song's base units combine into sub-phrases, that combine into phrases, that repeat for 2 to 4 minutes to form themes, that finally combine into songs. Diagram redrawn from [64], page 12. (All images were taken from Wikipedia, License CC-BY-SA-3.0).

manner [69]. Because of their beauty and complex structure, these vocal sequences were called songs [69].

Besides humpback whales, many other animals including birds, insects, frogs, primates, combine vocal elements into sequences (for review [70]). The reason why animals emit vocal sequences is often not clear [70]. Even though it is hard to decrypt the meaning of these sequences, it seems clear that these sequences carry information —much like other symbolic sequences occurring in biological and artificial contexts, like nucleotide sequences (DNA and RNA), amino acid sequences (proteins) and digital data (bit sequences). For animals, coding and decoding information out of vocal sequences could boost their fitness as a species. It has been hypothesised that species living in groups with complex social interactions, the complexity of their interactions is an important drive for the evolution of complex communication systems [71, 72]. This is known as the social complexity hypothesis and it has been tested in different species including rodents [73, 74], bats [75], non-human primates [76] and chickadees [72]. Investigating animal vocal sequences is important for understanding the forces driving their evolution and thereby the evolution of language [77].

Multiple studies have investigated vocal sequences on a diversity of taxa, yet there is very little agreement on how to approach this problem [70]. Some studies focus on the order of the vocal units as Markov chains [78, 79, 80, 81] or other models that account for sequential order [82, 83]; while other studies have recognised the importance of temporal dimension in the rhythm, calling rate and inter calling intervals [84, 85, 86].

Despite both dimensions —temporal and unit combinatorial— feature animal vocal sequences, very few studies investigate these two variables together. This chapter proposes a framework for quantifying animal vocal sequences taking these two dimensions into account. The framework consist of analysing recording annotations (described in Part I, see Fig. 2.1) using non-parametric statistical methods at two stages. In the first stage, the proposed methods are used to quantify temporal and combinatorial structures (Section 5.3) and in the second stage the proposed methods are used to compare the quantified structures (Section 5.4). Before presenting the mathematical tools we describe the approach to the problem (Section 5.2).

5.2 Temporal and combinatorial structures in vocal sequences

> *Man has an instinctive tendency to speak, as we see in the babble of our young children; whilst no child has an instinctive tendency to brew, bake, or write.* (Charles Darwin)

Vocal sequences, like speech and unlike written languages, are intrinsically temporal. Writing can be characterised in terms of words (semantic objects) and the arrangement of the words (syntax). Speech, on the other had, is a stream of utterances that beyond words and syntax, timing aspects —such as the speech's pace (also called speech tempo or speech rate) and the duration of words and pauses (silences)— play important communicative roles. For example, pause variance and syntax are strongly correlated [87]. The pauses at the end of sentences are longer than the pauses within a sentence [88, 89]. Speech rates can convey cues about the emotional state of the speaker. Slow rates are associated with low moods, while fast rates are associated with high levels of sympathetic arousal, during states of anger, fear, or excitement [90, 91, 92, 93].

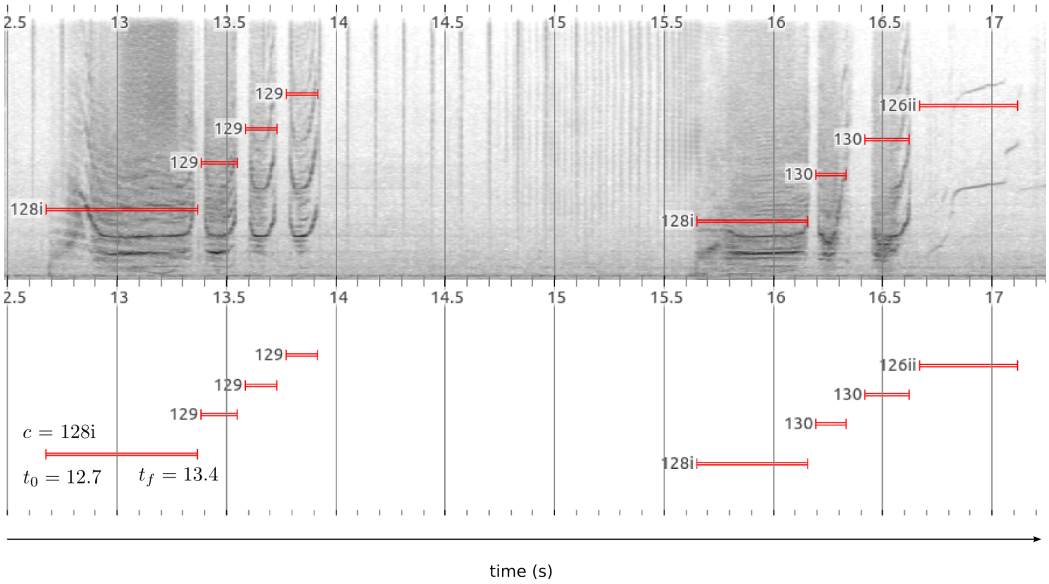


Figure 5.2: Annotations are segments of a time series. Sound is a time series that can be visualised as a spectrogram (upper panel). Annotations, represented with bars, indicate the temporal coordinates (t_0, t_f) of the calls, c; first bar shows the coordinates explicitly for call 128i. Lower panel shows only the annotations, which is the information used in Part II of this thesis. Image exported from sonic visualiser [99].

Calling rates of animals are neither stationary but have been observed to vary from context to context [94, 95, 96, 97, 98]. So, investigating animal vocalisations using speech-like approaches instead of pure syntax-like approaches might reveal structures important to the communication process. The proposed framework studies animal vocalisations as segments of a time series; by including the temporal coordinates of the vocal units, this is an approach closer to speech than that of most studies dealing only with the order of the units. In this part of the thesis we no longer deal with recordings themselves. Instead we will work with time series segments loaded from annotation files. These files are convenient for investigating animal vocal sequences because they summarise acoustic information in a light format and because of it being standard across audio processing software, errors can be easily checked at any point of the study. Below we define the variables of the problem and the kind of vocal structures we aim to quantify.

Segments of a time series

Recording annotations segment a time series with sounds of particular interest, in our case vocal units (Fig. 5.2). Each segment is characterised by a categorical variable that encodes the type of vocal unit c (e.g. a pilot whale call, Chapter 4) and their **temporal coordinates** (Fig. 5.2): stating t_0 and ending time t_f.

Vocal units

There are diverse criteria for defining vocal units [70], and the one chosen may depend on the taxa and the hypothesis being tested [70, 100]. Examples of commonly investigated units are songbird syllables, parrot notes, whale calls and dolphin whistles, among others. This examples are defined as vocal segments separated by silence gaps. However, units can also be defined in terms of sub-elements, or as a collection of elements, for instance humans' letters, syllables, words, sentences, etc. One should be careful not to take the human analogy too far, because while words have clear semantic meanings and are combined into sentences with more complex meanings, we do not know if animals vocal sequences share this characteristic[1]. Testing this hypothesis requires the assistance of ethological experiments, such as the study of an animal's response to playback sounds.

The next chapter uses the methods presented here to quantify structures in sequences of pilot whale calls. To save one word, throughout this thesis we use the term **call** to refer to the **vocal units**, unless indicted different (e.g. in Chapter 7 parrot vocal units are studied, which are called notes). However, one should bare in mind that the methods presented here are not exclusive to calls and without loss of generality they may be used to quantify the structures using other vocal units.

Quantified structures

The vocal structures quantified in this thesis can be separated into three types. The first type are **timing** structures, for patterns quantified using only the temporal coordinates of the annotations (Fig. 5.2). Timing patterns can be encountered in the distribution of call durations, the distribution of inter-call intervals (ICIs), and how the

[1] This property of language is called compositional syntax and up to last year was believed to be exclusive to human languages; demonstrated in [101] for bird calls.

calls chunk temporally. The second type are **combinatorial** structures, for patterns quantified using only the labels of the annotations, in our case the call types. Combinatorial patterns can be encountered in the ordering of calls in a sequence. This kind of structures are sometimes referred as **syntax** and are commonly investigated in bird vocalisations who tend to combine vocal units in highly structured sequences [102, 103]. The third type are **timing-combinatorial** structures, for patterns that combine the two kind of variables, temporal and call type coordinates. Investigating both variables together would allow us to explore correlations between the two (as have been observed in speech), i.e. in whether the distribution of call lengths depends on the call type, or whether certain call combinations have characteristic time intervals.

Having identified the variables of the problem: two continuous temporal coordinates and one categorical encoding the call type; and the structures we are aiming to quantify we are ready to go for the mathematical tools.

5.3 Quantifying vocal sequences

The framework here proposed aims to quantify timing and combinatorial structures embedded in animal vocal sequences. In this section we describe mathematical tools for such, starting with the timing patterns and the methods for quantifying them and continuing with the combinatorial patterns and the methods for quantifying them.

5.3.1 Timing patterns

We use the term timing patterns to refer to structures embedded in the temporal coordinates of the calls. Given a call c, the temporal coordinates are onset time $t_0(c)$ and the offset time $t_f(c)$ of the sound (Fig. 5.2). Out of these coordinates one can explore vocal aspects such as the **call and silence duration**, and the way calls **chunk temporally**.

Duration of a call c with temporal coordinates $t_0(c)$ and $t_f(c)$ is given by the difference $t_f(c) - t_0(c)$.

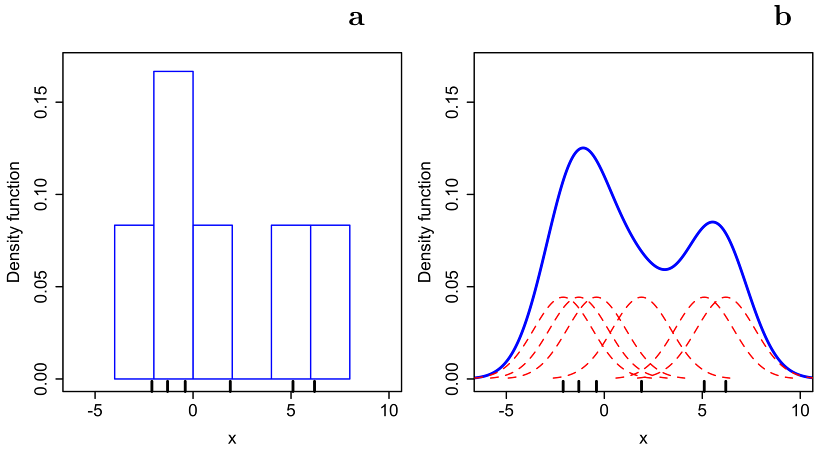


Figure 5.3: Comparison of a histogram and kernel density estimate (KDE). Estimate of the pdf of a random variable x with an histogram (**a**) and a KDE (**b**). Sticks at the bottom of both distributions indicate the data points. In the KDE, each data point contributes with a Gaussian kernel (dashed red line) to the whole distribution. (Image from Wikipedia, License CC BY-SA 3.0).

Inter-call interval (ICI) is the duration of the silence gap between two calls. Given the consecutive calls c_i and c_{i+1} indexed by i, i.e. $t_0(c_i) < t_0(c_{i+1})$, the ICI between the calls is

$$ICI_i = t_0(c_{i+1}) - t_f(c_i). \tag{5.1}$$

ICIs are sometimes defined with respect to the start of two consecutive calls. Here we opted to use the end and the start times to detach the length of the first call from ICI, which, as we will see in the next chapters it is often larger than the ICI itself. Another advantage of this definition is that it allows us to keep track of overlapping calls, for which ICIs are negative.

Call duration and ICIs were defined as differences of continuous variables, so both of them are continuous variables. One can get an overview of the values a continuous variable takes by looking at its distribution, or its normalised version: the probability density function (pdf). Below is presented an approach for estimating pdfs out of a sample.

Kernel density estimation (KDE)

Histograms are a common approach for estimating the probability density of a random variable from a sample (Fig. 5.3a). The problem with this approach is that histograms are highly dependent on the bin size and can be discontinuous even when the pdf is continuous. An alternative to histograms is to estimate the probability density by fitting a kernel density function (Fig. 5.3b). The method consist of parametrising a pdf as a sum of kernels (often times a Gaussian kernels is chosen) centred at each point of the sample. Given a sample with n points $X_1, X_2, \dots X_n$ the KDE at the point x is given by

$$\hat{\rho}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) \tag{5.2}$$

where $\frac{1}{nh}$ is a normalisation factor. For a Gaussian kernel, K takes the form

$$K(x; h) \propto \exp\Big(-\frac{x^2}{2h^2}\Big). \tag{5.3}$$

Chunk structure

Animal vocalisations are emitted with different rates that can be linked to behavioural or environmental factors [94, 95, 96, 97, 98]. Changes in the calling rate additionally structure the vocalisations into temporal chunks.

We can investigate the temporal chunks in terms of a thresholding silence gap τ, as follows. Given τ, a **call chunk** is a **sequence of calls**

$$c_1 c_2 \dots c_k, \tag{5.4}$$

indexed by i with inter-calling intervals ICI_i, such that $ICI_i < \tau \; \forall \, i \in 1, \dots, k-1$, i.e. the ICIs between the calls in a sequence are smaller than τ; and $t_0(c_1) < t_0(c_2) \dots < t_0(c_k)$, i.e. the order of the calls is determined by the initial time.

Notice that the temporal coordinates were only used to define a sequence, and this no longer carries temporal information. Also, in the case of overlapping calls, the one that starts earlier will appear before in the sequence.

From the definition of call chunks it follows that the number of calls in a sequence k, also called the sequence length, depends on the value of τ. The number of calls conforming a sequence is more restricted the smaller the value of τ; increasing τ relaxes the condition yielding longer sequences.

5.3.2 Combinatorial patterns

All methods described so far use only the temporal coordinates of the annotations, here we turn toward the categorical variable: the call type. In Section 5.3.1 we defined sequences of calls; a follow up question is whether the calls in a sequence occur in a particular order, e.g. whether the probability of having a particular call type depends on the former call. If we think of the call sequences as Markov chains we can compute the transition probabilities between consecutive calls. Determining ordering patterns in a sequence is a common query in the study of natural languages (e.g. the order of letter, syllables, words, etc.), where a pair of consecutive elements is called a **bigram**.

Bigrams probabilities

Given a sequence of calls $c_1 c_2 \ldots c_k$, the probability of having a call b in the n-th position, given the previous call was a is

$$p_{a,b} = P(c_n = b | c_{n-1} = a). \tag{5.5}$$

The transition probabilities $P(c_n|c_{n-1})$ can be estimated through a maximum likelihood [104], which normalises the conditional counts of consecutive calls in a sequence $N(c_{n-1}c_n)$ by the counts of the starting call $N(c_{n-1})$

$$P(c_n|c_{n-1}) = \frac{N(c_{n-1}c_n)}{N(c_{n-1})}. \tag{5.6}$$

5.4 Comparing vocal structures

Often, the motivation behind quantifying things is the possibility to compare them through the solid bedrock of numbers. By comparing the probabilities and pdfs assessed with the methods presented in the section above we can: insight on the significance of the probabilities, and assess dependences between the timing patterns and the call types. Below we present statistical methods for performing significance tests and comparing pdfs.

5.4.1 Statistical significance test

In statistics, significance tests are used for assessing the likeliness of an observation against a **null hypothesis** often symbolised as H_0. Null hypothesis are statements

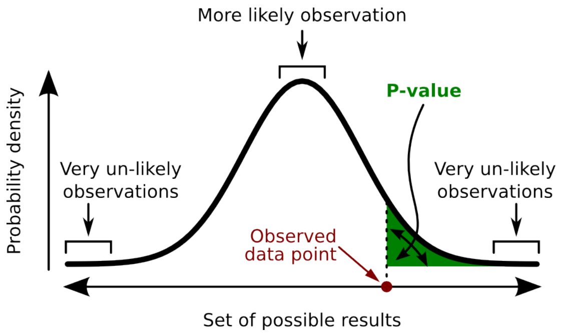


Figure 5.4: Illustration of a significance test. Null hypothesis is represented with the curve, the distribution of the test statistic. The p-value (shaded green area) is the fraction of times a value at least as extreme as the observed one was obtained assuming the null hypothesis is true. (Image modified from Wikipedia, License CC BY-SA 3.0).

one tries to evidence against, like the equality of means from two populations, or the relevance of the order of the items in a sequence. Besides the null hypothesis, important concepts in statistical hypothesis testing are: the test statistic, the p-value and the significance level α.

When carrying out a significance test, the null hypothesis is represented as the distribution of a random variable called the **test statistic** (e.g. the t-statistic, χ^2-statistic). Using confidence intervals one can measure the probability of obtaining a result at least as extreme as the one observed assuming the null hypothesis is true. This probability is called the **p-value** (Fig. 5.4). So, small p-values indicate a small chance that H_0 is true, in which case the null hypothesis is rejected. The p-value's threshold for rejecting or not the null hypothesis is the **significance level** α, which is often a number between 0.1 and 0.01 depending on the area of study.

Common test statistics such as the s-statistic, the t-statistic and the χ^2-statistic, rely on assumptions that are not always met by the data concerned, e.g. normality of the sample, homogeneity of variance, a minimum number of samples. An alternative for such cases is to generate the sampling distribution of the test statistic numerically using computer power. Synonyms for such tests are exact tests, permutations tests or randomisations test.

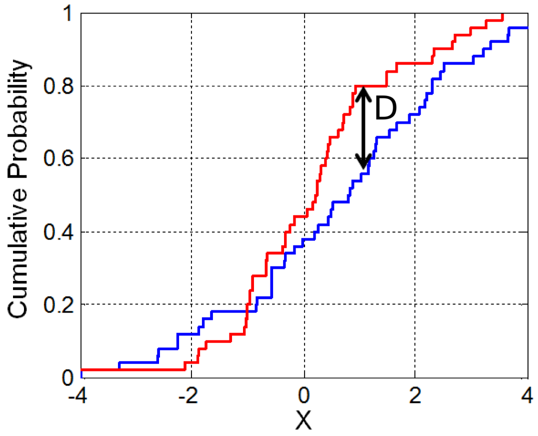


Figure 5.5: Illustration of a two sample Kolmogorov-Smirnov statistic. Red and blue curves are two empirical cumulative distributions and black arrow is the D-statistic. (Image: Bscan@wikipedia CC0).

Permutation test

In a permutation test the distribution of the null hypothesis for any test statistic is generated numerically by randomising the labels of the data. In such tests the p-value is computed as the fraction of times that randomising the labels yielded a value is at least as extreme as the observed one (Fig. 5.4). Permutation test produce good results (p-values) with 1000 randomisations [105].

Kolmogorov-Smirnov test (KS-test)

The Kolmogorov-Smirnov test or **KS-test** is a non-parametric statistical test for assessing the probability that two continuous variable samples were drawn from the same distribution. It is often used for testing the normality of a sample, but here we focus on the comparison of two generic samples.

The KS-test uses the D-statistic, which is the maximum distance between the cumulative distributions of the two samples being compared (Fig. 5.5). The p-values of the D-statistic indicate the probability of getting a D-statistic at least as large as the one observed.

5.4.2 Comparison of continuous variable distributions

Kullback-Leibler divergence

The KullbackLeibler divergence from Q to P, two discrete probability distributions, is given by

$$D_{\mathrm{KL}}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{5.7}$$

The KL-divergence is an information theoretical measure of the divergence of two probability distributions. It is non negative and equal to zero if and only if $P = Q$ so something like a distance —in fact it often called KL-distance— but not quite since is not symmetric and does not satisfies the triangle inequality.

Chapter 6

Vocal sequences of pilot whales

In collaboration with Heike Vester from *Oceansounds Norway*[1]

Long-finned pilot whales are highly social and vocal dolphins. They are often seen in large groups that can exceed hundreds of individuals [106]; they display altruistic behaviours like alloparenting (babysitting the calves from other group members) [107]; and they have been even seen socialising with other dolphin species [108]. Like humans (and most mammals), these animals are unable to survive on their own[2].

Little is known about long-finned pilot whales, catalogued as data deficient (a category for species whose conservation status cannot be properly assessed due to insufficient information) by the International Union for Conservation of Nature (IUCN) [111]. Studying these animals is challenging because they spend most of the time deep underwater where visibility is limited; and playback experiments —the best ally for studying animal behaviour— are not feasible. Fortunately pilot whales, as many other marine mammals, are highly vocal and their acoustic signals may reflect aspects about their social structures [60, 112] and contextual behaviour [113].

Pilot whales depend on acoustic communication and have complex vocal repertoires with clicks, buzzes, whistles and a range of calls mixing tonal and noisy sounds [61,

[1] www.ocean-sounds.org

[2] Do not worry, no brutal experiments were conducted to test this hypothesis, but pilot whales' social behaviour suggest this to be the case (more details in the next section). The strong social bonds and herding behaviour of pilot whales make these animals highly vulnerable to massive stranding [109, 110]. Single standers are rare and have only been observed in ill whales, whereas massive standings tend to be of mostly healthy individuals.

113, 114, 115]. Therefore studying their sounds is a good alternative for improving our understanding on this species. In the present chapter we focus on calls, deem relevant in keeping group cohesion and coordination [112, 116], both important social roles.

Pilot whales emit calls in rhythmic repetitions, while this aspect has been indicated in early bioacoustic studies [114], the focus has been mostly on the vocal repertoires [61, 113, 115, 117, 118], disregarding the temporal structure of the vocal sequences.

This study quantifies temporal and combinatorial call patterns in vocal sequences of Norwegian pilot whales deepening into their rhythmic structures. Patterns are quantified over annotations extracted from recordings with pilot whale calls using non parametric statistical methods (described in chapter 5). Before entering into the vocal structures some words are said about the whales and our dataset.

6.1 Long-finned pilot whales

The long-finned pilot whale (*Globicephala melas*) is a dolphin species encountered in the north Atlantic and southern hemisphere oceans. These animals feed mainly on squid [119], reason why they spend long time deep under the sea surface.

Pilot whales have a social structures that resembles those of Killer whales [120, 121, 122]. They form tight knit matrilines of *c.a.* 10 whales [123] and are often seen travelling with other matrilines forming groups with hundred individuals [123]. The social structure of pilot whales is more fluid than that of killer whales, similar to the fission-fusion societies encountered in many dolphins, but at the level of group instead of individuals [123]. It has been observed that the acoustic communication of whales reflects aspects of their social structures [60]. Thus, pilot whales acoustic signals may have: signature whistles [21, 124], often encountered in dolphins living in fission-fusion societies; and group specific calls similar to those of resident killer whales [125].

6.1.1 Dataset

Free living pilot whales were recorded in July, 2010 at the Vestfjord by the Lofoten archipelago in Norway (Fig. 6.1) with an estimate of 45 whales present during the encounter. Sounds were recorded with a hydrophone Reson TC4032 with a sampling frequency of 92kHz and downsampled to 46kHz prior to analysing.

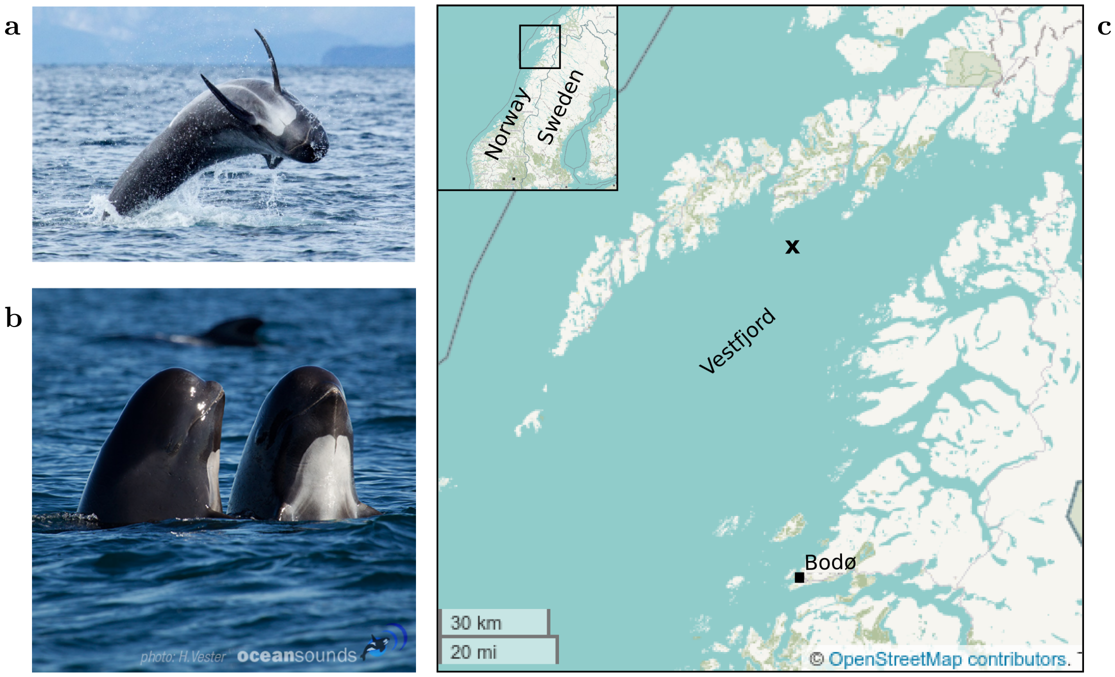


Figure 6.1: Pilot whales and location of the encounters. **a**, Pilot whale jumping (photo Heike Vester). **b**, Pilot whales raising their heads out of the water, this position is common during social contexts and is called spy hopping (photo Heike Vester). **c**, Map of the Vestfjord by the Lofoten archipelago in Norway where the whales were recorded indicated with an **x** (image edited from OpenStreetMap® CC BY-SA).

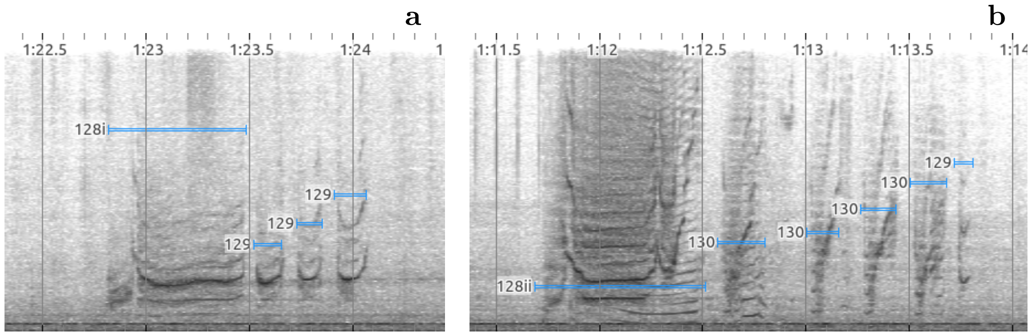


Figure 6.2: Examples of annotated recordings. Segment of a recording with a sequence starting with (**a**) call 128 subtype *i* and (**b**) call subtype *ii*. Subtype *ii* has an U-shape element at the end. Height of the labels was arranged for visualisation purposes and has no other meaning. Spectrogram exported from Sonic Visualiser [99].

The recordings were annotated with the call type (see chapter 4) and its starting time and duration (Fig. 6.2). Continuous sound segments were labeled as call types, with a 2-3 digit code, according to the catalogue in [61]. Pilot whales often use calls with similar spectrotemporal features, these were not placed into different call categories but were named as call subtypes and labeled with the same digit code and different number of *i*'s at the end (Fig. 6.2).

Annotations were generated automatically, with the algorithms described in chapters 3 and 4 and curated by a human observer. I revised the temporal labels of the calls and Heike Vester revised the call labels.

In this chapter we use two datasets: one with only the temporal coordinates of the calls and a fully annotated one with call types and their temporal coordinates. The first dataset has 497 calls from 4 recordings and we use it in section 6.2 to investigate the temporal patterns of the calls. The second dataset has 127 calls from a single recording and we use it in section 6.3 to investigate the combinatorial patterns of the calls in connection with their temporal structure.

6.2 Rhythm and temporal structure

Here we analyse the temporal structure of 497 calls regardless of the call type by looking at patterns within the call lengths, inter-call intervals (ICIs) and correlations between these two variables.

Call length

The distribution of call lengths ranged from in 0.1 s to 2.5 s in our sample with 3 characteristic call lengths roughly separated as follows (Fig. 6.3a): short calls, < 0.4 s, with 52% of the calls; medium length calls, between 0.4 s and 1 s with 42% of the calls; and long calls, > 1 s, with 1.5% of the calls.

Inter call intervals

Our sample had a wide distribution of ICIs; with a minimum at -0.5 s, corresponding to overlapping calls; and a maximum above 4 minutes. Despite the wide rage, more than 50% of the ICIs are smaller than 1 s. Thus, it makes sense to look at the distribution of the logarithm of the ICIs (log-ICIs) instead of the distribution of ICIs, in order to

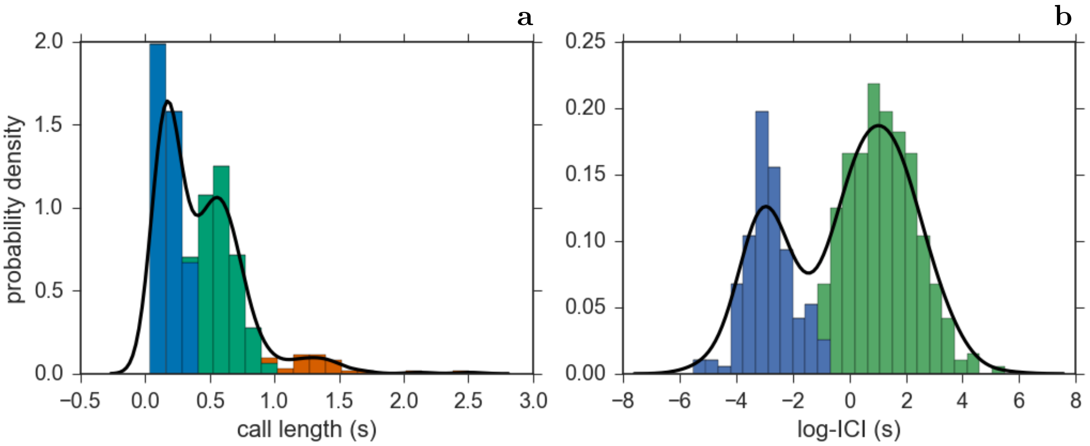


Figure 6.3: Distribution of call lengths and ICIs. **a**, Distribution of call lengths shows three dominant sizes indicated with different colours. **b**, Distribution of the logarithm of the ICIs with two characteristic lengths indicated with different colours. Probability density function fitted with a Gaussian kernel density estimate shown with solid line.

capture the short and long time scale structures together (Fig. 6.3b). The log-ICIs can be roughly split into two kinds: short < 0.4 s with 41% and long > 0.4 s with 59%.

Correlation between the call length and the length of the following silence

Call duration carries information about the length of subsequent silence. This can be illustrated by plotting the joint distribution between call length and following ICI (Fig 6.4b) and comparing it with the joint distribution of call length and following ICI assuming no dependence; destroying correlations by randomising the data (Fig 6.4c). Differences in the observed and the randomised distributions suggests that the call length and the following ICI are indeed correlated. Such differences may be quantified using the KL-divergence (see chapter 5), $D_{KL}(P||Q)$, with P being the observed joint distribution and Q the shuffled distribution (Fig. 6.4d). Because the $D_{KL}(P||Q)$ measures the divergence from Q to P (it can be thought as a distance), larger values indicate a stronger correlation between the call length and the following ICI. Comparing this distance with the distance between two uncorrelated variables (Fig 6.4c), we observe that the call length and the following ICI are more correlated than expected by chance. This correlation suggests that short calls are more often followed by short

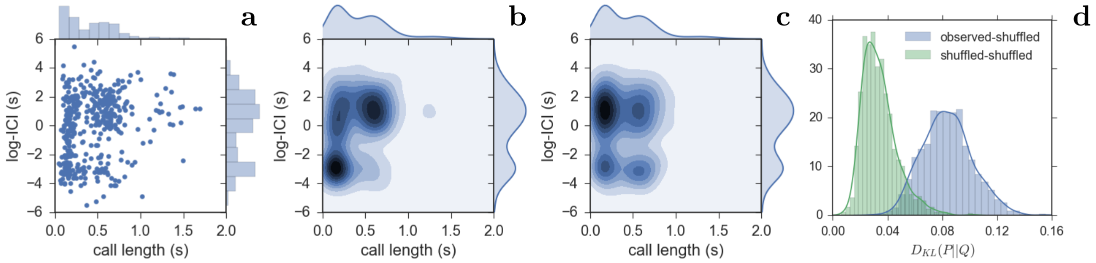


Figure 6.4: Correlation between call length and following ICI. **a**, Scatter plot of the call length and following log-ICI. **b**, Joint probability of the call lengths and the log-ICIs obtained with a two dimensional Gaussian kernel estimate of the points in panel **a**. **c**, Joint probability of the call lengths and log-ICIs assuming no correlation, with randomised data. Probability density function was estimated with a Gaussian kernel. **d**, KL-divergence of P, the joint probability between the call length and the ICI (panel **c**) and Q, the joint the probability with shuffled data (blue distribution). In green the distribution of the KL-divergence between two shuffled distributions. For both distributions data was randomised 1000 times.

ICIs and long calls are more often followed by long ICIs (comparison of panels (b) and (c) from Fig. 6.4).

Correlation between inter-call intervals

As the call length and the following ICI are correlated, the duration of nearby silences is also correlated. Analogous to our previous analysis, we plot the joint distribution of consecutive silences (Fig. 6.5b) and compare it with its randomised counterpart (Fig. 6.5b) and do the same for silences separated by two calls (Fig. 6.5c-d). The observed distributions look different from the randomised ones. These differences are quantified using the KL-divergence (see chapter 5), measuring the distance between the observed and randomised distributions for consecutive silences and for silences k calls away. We observe that the strength of the correlation peaks for consecutive ICIs, yet for silences up to 20 calls away the correlation is still stronger than expected by chance (Fig 6.5). This correlation suggests that close ICIs are clustering into time scales; short ICIs are more often followed by short ICIs and long ICIs are more often followed by long ICIs (Fig 6.5a-d).

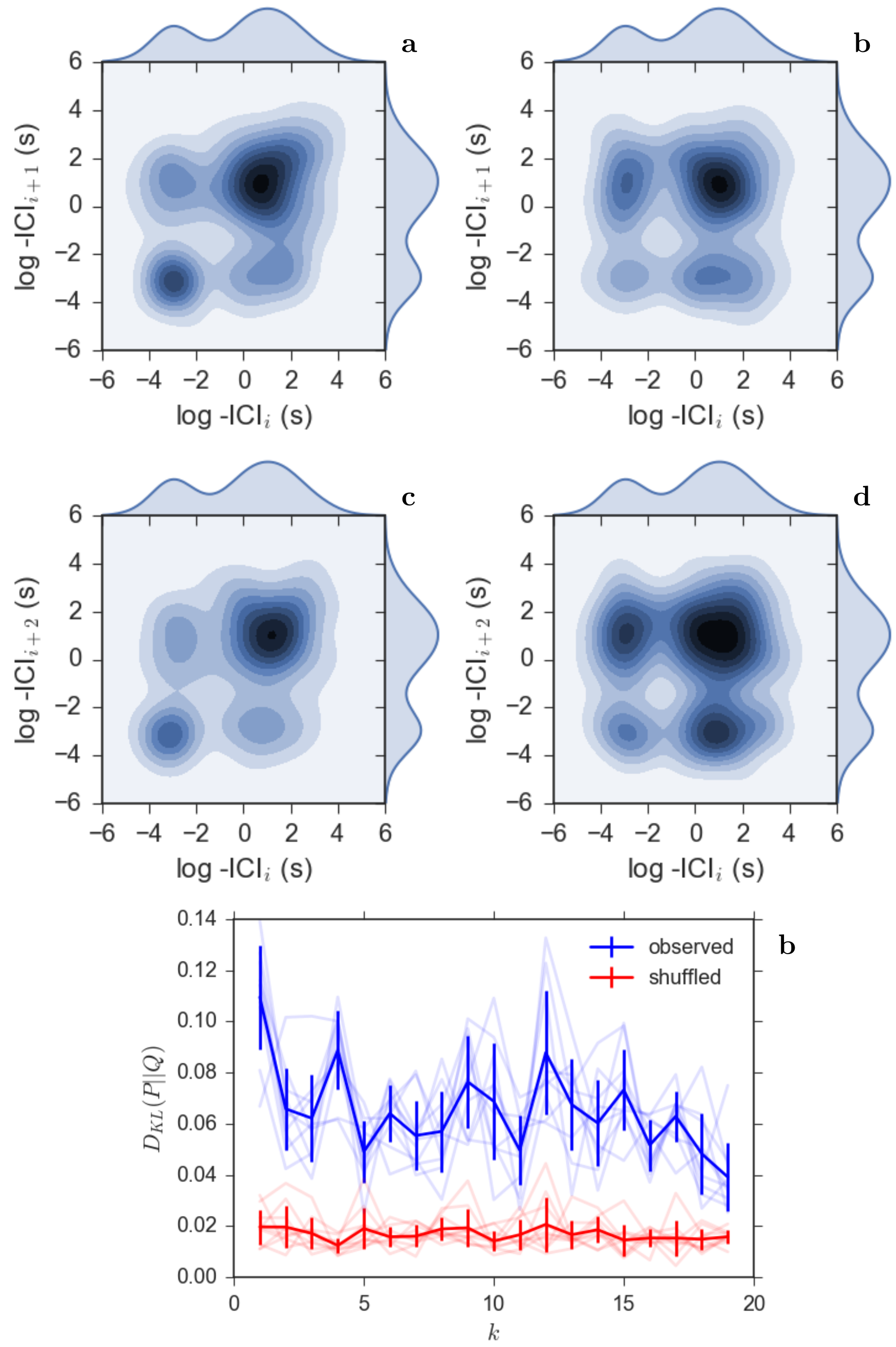


Figure 6.5: Correlation between close ICIs. Joint probability between (**a**) consecutive ICIs (**b**) consecutive ICIs assuming no correlation (with randomised data) (**c**) between ICIs 2 calls away and (**d**) between ICIs 2 calls away assuming no correlation (with randomised data). **d**, KL-divergence of the joint probability between silences k-calls away and the joint probability with randomised data (blue). In red the KL-divergence between two joint probabilities with randomised data. Data was shuffled 10 times for each k.

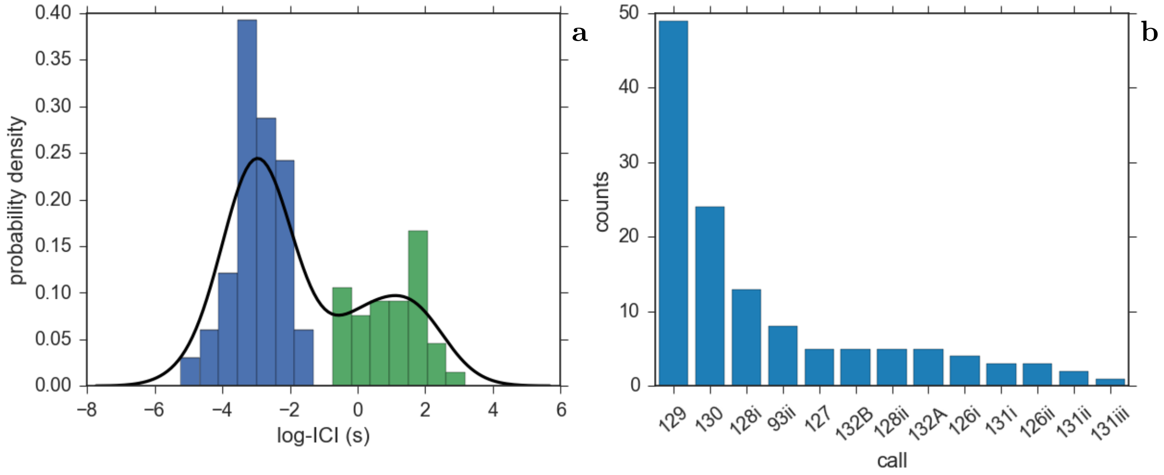


Figure 6.6: Dataset. **a**, Distribution of the logarithm of the ICIs. Dominant ICIs are indicated with different colours: blue for ICIs smaller than 0.4 s and green for ICIs larger than 0.4 s. **b**, Call composition in the dataset sorted by counts.

6.3 Temporal and combinatorial patterns

In section 6.2 we investigated vocal structures based exclusively on the temporal coordinates of calls. Here we take the call types into account to investigate combinatorial patterns between the calls in connection with the temporal variables. For this I use a sample of 127 calls, all from the same tape, with 13 call types (6.6b). Call counts for different call types were not the same; 4 most frequent calls follow a power law relation with their rank as can shown in Fig. 6.6b. Like in the sample from section 6.2, ICIs in this sample also show a bimodal distribution; where 68% of the ICIs are smaller than 0.4 s (Fig. 6.6).

Call types and their timing patterns

Different call types have different temporal distributions. Call types have characteristic lengths that can be *ad hoc* separated into: Brief calls shorter than 0.2 s; short calls between 0.2 s and 0.4 s; and long calls, longer than 0.4 s(Fig. 6.7a).

ICIs also depend on the call type; the distribution of silences prior and after the call differs for most calls (Fig. 6.7b). The asymmetries in the distributions of calls 128i, 128ii, 123A and 131ii (and possibly 131ii and 131iii but we have very few samples for these two calls) suggests that these are used as introductory calls; mostly preceded by

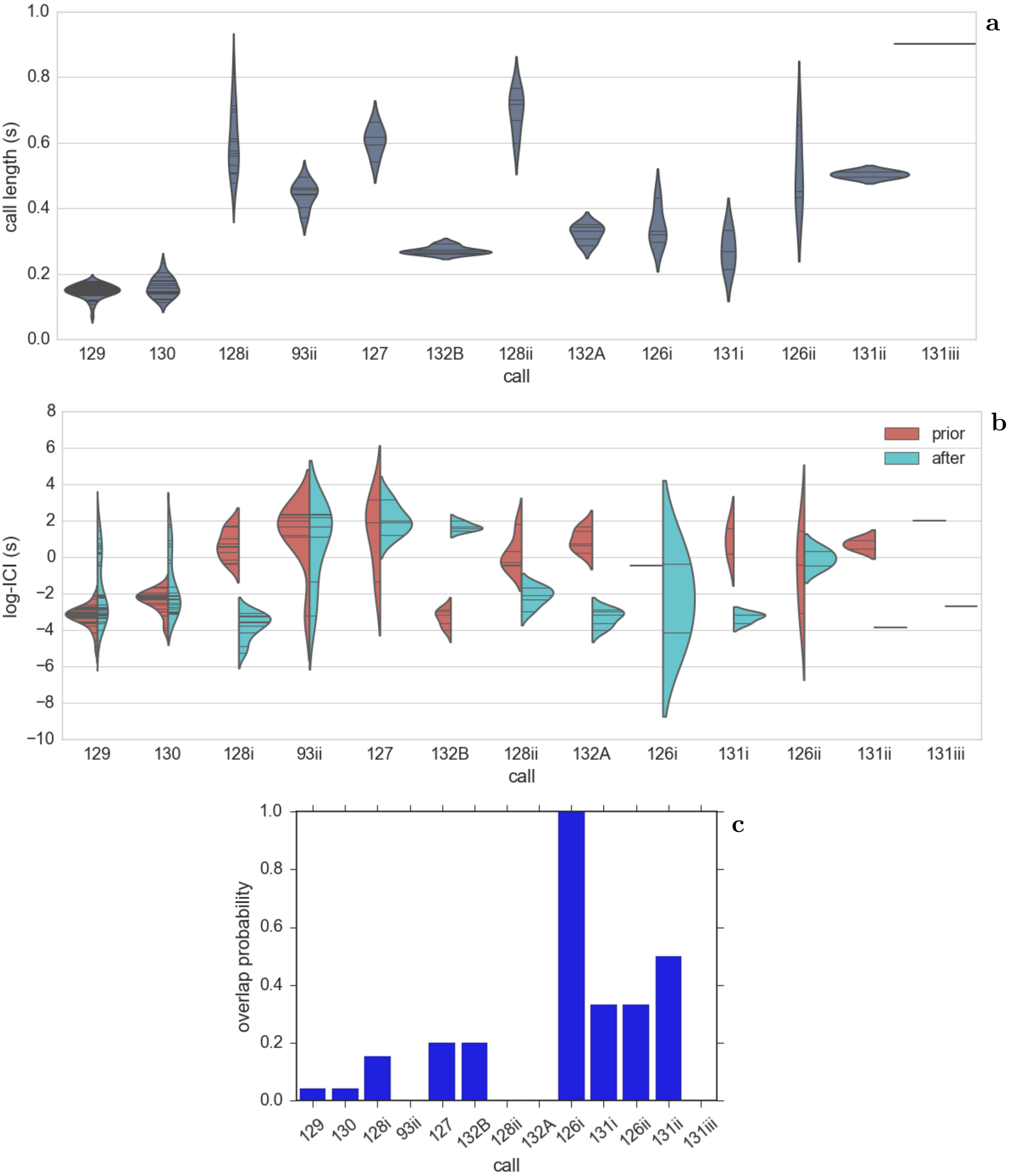


Figure 6.7: Temporal properties of the calls. **a**, Distribution of call lengths with a Gaussian kernel fitting. **b**, Distribution of the log-ICIs with a Gaussian kernel fitting. Sticks inside the distributions show observed values. **c**, Fraction of times a call was recorded in overlap with another call.

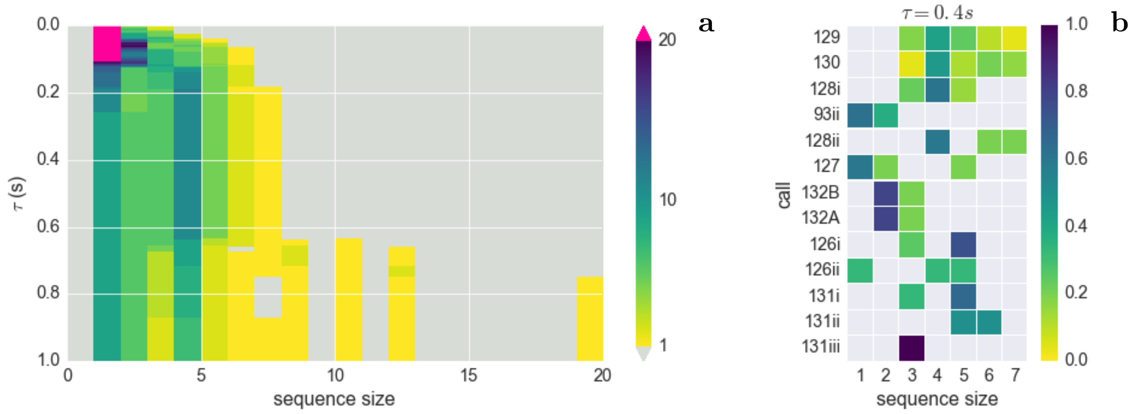


Figure 6.8: Chunk structure of the call. a, Call grouping as a function of τ in the interval 0 s to 2 s. **b**, Call composition (normalised for each call) for different sequence sizes defined with $\tau = 0.4$ s. Calls sorted according to their rank.

long silences and followed by short ICIs. Unlike, call 132B seems to be an ending call. Calls 129 and 130 have similar distributions; both calls are often preceded by short ICIs (of approximately 0.1 s) and succeeded by silences that range from small, of the order of cents of a second, to silences of the order of seconds.

The log-ICI does not include overlapping calls, so we looked at these calls separately, computing the probability of call to occur overlapped (Fig. 6.7c). Some calls like 126i, 131i, 126i and 131ii occur overlapped more often than others, say 93ii, 128ii and 132A, despite the latter occurring much more often in our sample.

6.3.1 Chunk structure

Here we investigate how the calls are chunked in time and whether different call types have different ”chunknesses” (the likeliness of a call to be grouped). Clearly the way calls temporally chunk depends on the largest interval between consecutive calls in a sequence, τ. (chapter 5). Figure 6.8a shows the distribution of sequence sizes as a function of τ, where we observe that for $\tau \in [0.2, 0.6]$ s, the distribution of chunk remains reasonably stable, in agreement with our observation from Fig. 6.6a. Within this region, calls are grouped in chunks of up to seven calls, with a higher number of chunks of 4 calls.

Different call types tend to occur in different group sizes (Fig. 6.8). Taking $\tau = 0.4$ s we observe that most calls are emitted within the vicinity of another call, and only calls

93ii, 127 are more likely to occur in isolation than grouped. Out of the grouped calls, types 132A and 132B often occur in sequences with 2 calls while 129, 130 occur more often in sequences with 3 to 5 calls.

6.3.2 Call combinations

Having noticed that calls are emitted in groups, we are ready to investigate whether calls follow an order within these groups or whether whales just dump them randomly.

Transition probabilities between consecutive calls in sequences with $\tau = 0.4$ s indicate that call types are preferentially combined (Fig. 6.9). By assessing the significance of these transition probabilities with a permutations test, we obtain that the observed probabilities cannot be explained by pure chance (Fig. 6.9b-c). Three kind of patterns stand out from Fig. 6.9: (1) bigrams occurring more often than expected by chance (128ii 129, 128ii 130, 132A 132B, 131i 129 and repetitions of calls 129 and 130) (2) calls likely to occur at the beginning of a sequence (128i, 128ii and 132) and (3) calls occurring in isolation (93ii and 127 potentially do, but we do not have enough samples to confirm the latter).

Some call types pattern similarly. For instance, calls 129 and 130 are most likely followed by repetitions or by each other, and despite being the two most frequent call types, they never occur at the beginning of a sequence. Calls 128i and its variation 128ii were never observed at the end of a sequence and occurred most times at the beginning.

Call repetitions occur more often than expected by chance (Fig. 6.10a) and the probability of having a repetition is affected by the call type and the ICI (Fig. 6.10)b.

6.3.3 Transitioning times of the bigrams

Along the chapter we have seen that call types and timing properties are strongly connected; through the distribution of call lengths and ICIs and the way calls get grouped. This suggests a call type explanation for the bimodal distribution of ICIs we observed in Fig. 6.6b. Looking at the distribution of ICIs for specific bigrams, we distinguish two groups (Fig. 6.11bc): one group with ICIs shorter than 0.3s; and a second group with ICIs typically of the order of seconds. This observation favours the hypothesis that different bigrams have different transitioning times, which is confirmed by carrying a KS-test (Fig. 6.11bc).

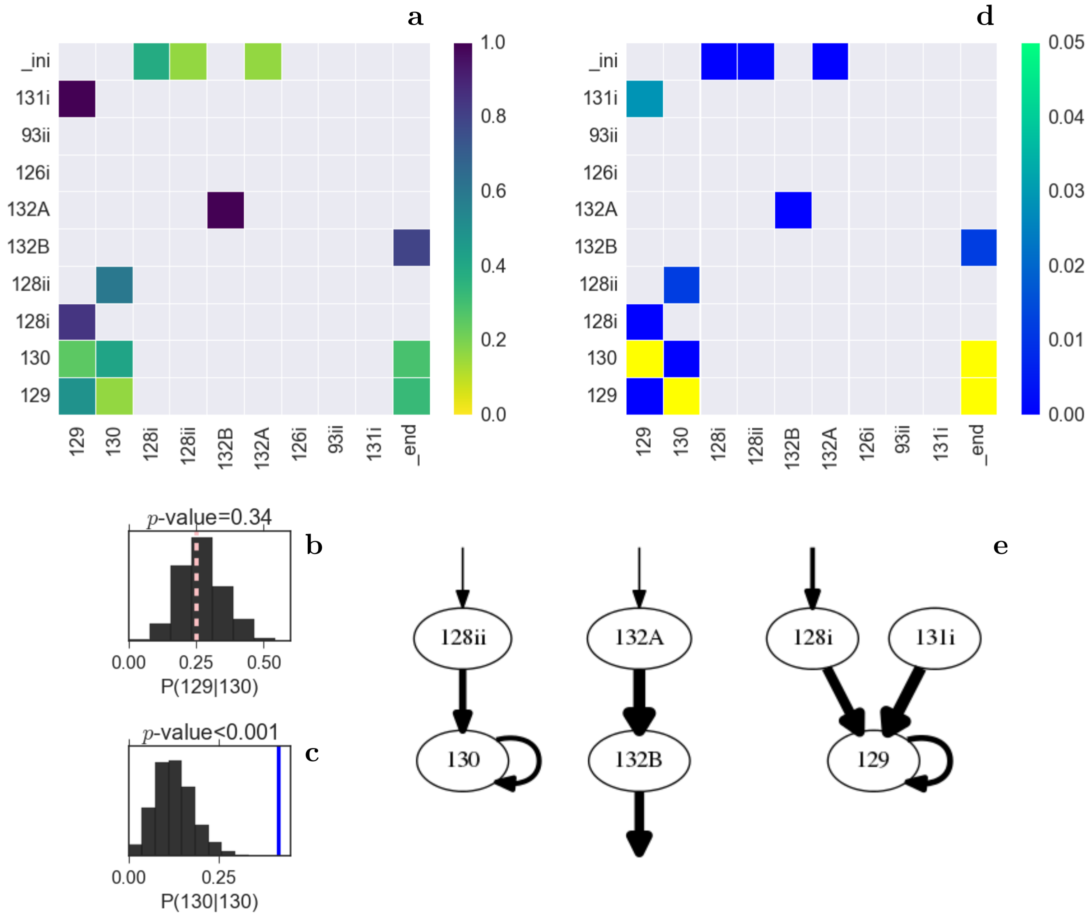


Figure 6.9: Transition probabilities between the calls in sequences with $\tau = 0.4$ **s. a**, First order transition probabilities between the calls in our sample. Bigrams observed less than three times are masked and shown on grey. Call labels are sorted according to their rank. Labels _ini and _end indicate the beginning and end of a sequence for a given τ. Significance for the bigram probabilities is assessed with a permutations test, where the observed transitions probabilities are compared with the distribution of the null hypothesis that calls order is irrelevant, obtained by randomising the calls 1000 times. Transition probabilities for bigrams 130 129 (b) and 130 130 (c); vertical line indicates observed values and in black the null hypothesis distribution. **d**, p-values for all bigrams observed at least three times. Bigrams with p-values larger than 0.05 are shown on yellow. **e**, Diagram of call (nodes) transitions represented as arrows with widths proportional to the transition probabilities. Only transitions with statistically significant probabilities are shown.

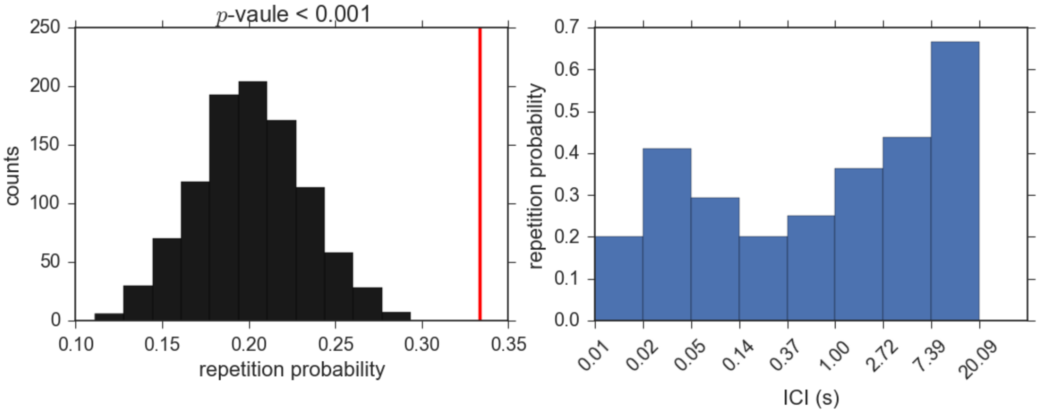


Figure 6.10: Call Repetitions. **a**, Probability of having two consecutive calls of the same type within 10 s: observed (vertical line) and null hypothesis that call order is irrelevant. Null hypothesis distribution obtained by permuting the calls within a tape 1000 times and then computing the proportion of repetitions for each realisation. **b**, Proportion of bigrams with the same calls for different time intervals, using logarithmically increasing bin widths.

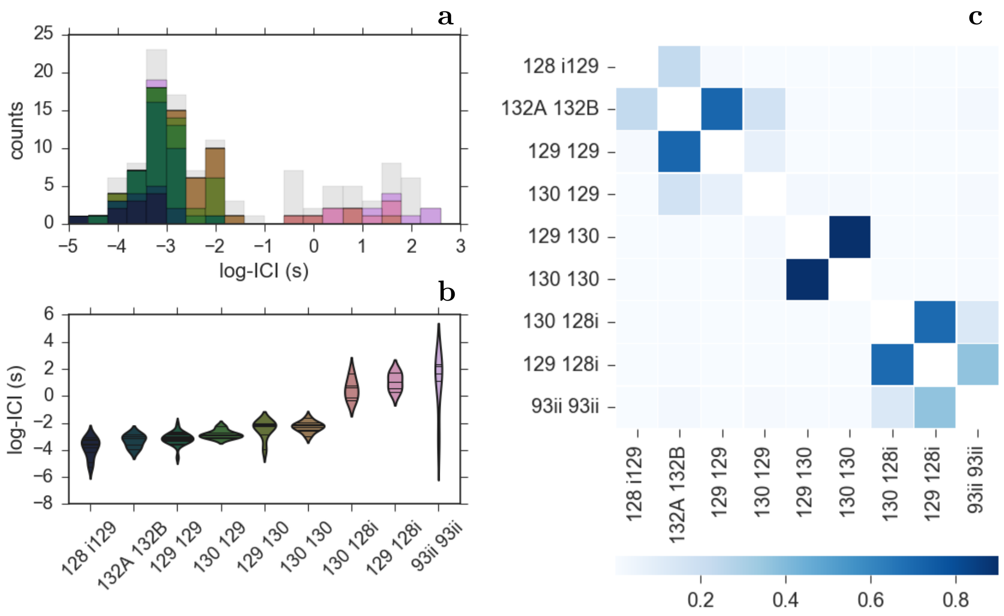


Figure 6.11: Inter-call intervals of the bigrams. **a**, In grey the histogram of the log-ICIs in the interval zero to 20 seconds. Overlaid the histogram of log-ICI for specific bigrams observed more than five times indicated with different colours. **b**, Distribution of the log-ICIs for the bigrams occurring more than five times with Gaussian kernel fitting. Lines inside the distribution indicate observed values. **c**, p-values of a KS-test comparing the distribution of ICIs for all bigrams against each other.

Bigrams with long ICIs have a larger variance than those with short ICIs. In other words, the ICIs of the bigrams with typically short ICIs are more precise than those with typically long ICIs.

6.4 Summary, discussion and outlook

Pilot whales produce calls in highly structured vocal sequences, so far, mostly described qualitatively. Here, I quantified temporal and call combinatorial patterns (summarised in table 6.1) in these animal's sequences. The significance of the encountered patterns was confirmed in spite of our small sample size by using non parametric statistical methods. These results are discussed below, first in terms of similar vocal patterns reported in pilot whales and other marine mammals; and later with respect to the context and possible functions these patterns may have for the animals.

6.4.1 Structured vocal sequences of marine mammals

Repeated call types are often encountered in delphinids. This kind of vocal sequences have been reported in: northern right whale dolphin *Lissodelphis borealis* [126], Guiana dolphin *Sotalia guianensis* [127], melon headed dolphins [128], short-finned pilot whales, *Globicephala macrorhynchus* [129], log-finned pilot whales *Globicephala melas* [130]. Our sample also had a high amount of call repetitions with a higher probability than expected by chance. Besides, it was found that the likeliness of a repetition depends on the call type and the ICI.

The analysis carried here disclosed a strong connection between the call types and the inter call intervals (ICIs). This connection has been recognised before yet very few studies have delve into this dimension. One of these studies investigated the transitioning times in whistles from bottlenose dolphins [131]; Janik *et al.* found that signature whistles [124] were emitted in bouts with characteristic ICIs between 1 s and 10 s whereas chirps (shorter whistles) [124] have shorter ICIs. Sayigh *et al.* also looked into the distribution of ICIs in vocal sequences from short-finned pilot whales and found that call types occurring often tend to have shorter ICIs ($<$ 1 min) than less frequently occurring call types [129]. Here, it was found that the distribution of silences, prior and after a call, is highly dependent on the call type and that bigrams —specific call combinations— have characteristic transitioning times. Additionally, bigrams with short ICIs have more precisely defined ICIs than bigrams with typically larger ICIs. Esch *et al.* observed a similar pattern in the ICIs of bottle-nose dolphin signature whistles [132]. Because of the broad nature of the silences between calls, examining them in terms of the logarithm of the ICIs instead of the ICIs is more convenient, since

Observed pattern	Figure
Temporal	
• The distribution of call lengths shows three characteristic sizes, *ad hoc*:	6.3a
– short calls < 0.4 s	
– medium calls between 0.4 s and 1 s	
– long calls, larger than 1 s	
• ICIs have a broad distribution with 2 characteristic time scales, *ad hoc*:	6.3b
– short ICIs < 0.4 s	
– long ICIs > 0.4 s	
• Call lengths are correlated with the ICIs	6.4
• Close ICIs are more correlated that expected by chance	6.5
Call combinatorial	
• Some call types are emitted more often that others	6.6a
• Call type 126i always occurred overlapped	6.7c
• Call types can be preferentially combined in sequences ($\tau = 0.4$ s)	
– bigrams 128i 129, 128ii 130, 131i 129 and repetitions of 130 and 130 occurred more often than expected by chance	6.9
• Repetitions of consecutive calls occur more often than expected by chance ($\tau = 10s$)	6.10a
• Likeliness of having a call repetition depends on the ICI	6.10b
Temporal and call combinatorial	
• ICIs depend on the transitioning calls	6.11
– a fraction of the bigrams have transitioning times shorter than 0.3 s	
– another fraction has transitioning of the order of a second	
• 128i, 128ii and 132A are introductory calls	6.7b, 6.9
• Call types (1) 129 and 130, and (2) 128 i and subtype ii share various similarities	
– in their distribution of call lengths	6.7a-b
– their distributions of ICIs, prior and after	6.7a-b
– 128i and 128ii are both introductory calls followed most often by calls 129 and 130	6.9
– 129 and 130 are often followed by repetitions or by each other	6.9

Table 6.1: Temporal and call combinatorial patterns. Summary of the temporal and combinatorial patterns encountered in our sample. Patterns are separated by category: temporal, call combinatorial and temporal and call combinatorial together. The sequence definition parameter τ thresholds the maximum ICI between consecutive calls in a sequence.

this transformation enables to appreciate the rich structure in the small time scales together with the few, yet not negligible, ICIs in the large time scales.

The distribution of call types in our sample showed different counts for different call types; with the four most frequent calls following a power law relation with their rank. Words in human languages have this trait known as Zipf's law [133], and underneath it a least effort lexical principle has been proposed [133, 134]. Our sample size is too small to claim that pilot whale calls follow Zipf's law, but this scaling was found in bottle-nose dolphin whistles [79]. The implications of Zipf's law are debatable [135, 136, 137, 138, 139], yet deepening into which features characterise life signals is an important matter with a deep potential in the search for extraterrestrial intelligence [37].

Pilot whale calls were found to occur in preferred ordered combinations, here called bigrams. These kind of call associations have been reported for a variety of marine mammals including: bottlenose dolphins (*Tursiops*) [79], killer whales (*Orcinus orca*) [140, 141], short-finned pilot whales (*Globicephala macrorhynchus*) [129], humpback whales (*Megaptera novaeangliae*) [142]. The significance of the call transition probabilities was assessed here in spite of the small sample by carrying an exact test.

It is possible that the probability of having a particular call type is not only dependent on the call immediately before, but on the previous n calls. With the analysis carried here we would not be able to detect these higher order associations. As in other studies [79, 129, 141], we were limited by the size of the sample. Ferrer I Camacho and McCowan turned around the sample size problem measuring the long range correlations in dolphins whistles [143]. Their analysis neither takes into account the exact arrangement of previous whistles but the authors found that dolphin whistles can be correlated up to the 4th whistle using a local randomisation test.

Sequences with a specific introductory sound have been widely reported for birds [144, 145] and monkeys [146] but not much for marine mammals. Here, three call types used at the beginning of sequences were identified.

Overlapping calls are common in recordings of social marine mammals, yet these are not investigated and often omitted from studies. Here we saw that some call types have a higher likeliness to occur overlapped. Call type 126i always happened in overlap with at least another call. Our sample size is too small to draw conclusions out of this observation, but it suggest that overlapping calls might have valuable information.

6.4.2 Relation to context and possible function of the patterns

Pilot whales produce dense sequences of repeated call types that resemble the way bootlenose dolphins produce signature whistles [147]. Given this similarity, it was hypothesised that pilot whale calls might function as bottlenose signature whistles [129]. However a recent study found no evidence in support of this hypothesis [130] so the function of this the repetitive signals is still an open question.

One of the challenges in studying pilot whale vocal repertoires is the diversity of similar calls, here referred as call subtypes. Call subtypes are essentially the same call type with a small modification such as the addition of an independent frequency component or a noise segment at the beginning or the end of the calls [148]. Zwamborn [148] inquires on the possible function of this call embellishments, whether they are emotional indicators or have the potential to convey location —this last hypothesis is based in one for killer whales, where the directional nature of upper frequency is believed to be used for coordination [149]. While the function of these call modifications is not clear, evidence in this chapter suggests that call subtypes might have similar ways of patterning. For instance, call 128i and 128ii both function as introductory calls followed by calls 129 or 130, these last two calls are also similar, being the two shortest calls in our sample (< 0.2 s). Quantifying the temporal patterns of the calls, as done in this study can lead to new ways of assessing similarities in the call usage, an interesting hypothesis to investigate in the future on a larger sample.

We detected several vocal patterns with potential prosodic content. On example is the wide distribution of inter-call intervals and its large span of correlations. Another example is the large variance in the length of calls of the same type. Particularly outstanding are call types 128i and 128ii whose length varied more than 50% in our sample. Regardless of these aspects being intentional or unintentional information about the signaller is cued acoustically [3]. The relevance of this information may be investigated further with larger datasets and conducting behavioural studies.

Valuable insights on the origin of language can be achieved through comparative studies in animal communication [3, 150]. Whales and humans are closely related, we both are mammals, share similar brain structures, are capable of vocal mimicry and vocal learning. Hence studying pilot whale communication can reveal phylogenetic traits that could have led to the acquisition of language [3, 150].

6.4.3 Outlook

In this chapter we describe diverse temporal and call combinatorial patterns in pilot whale call sequences. While these patterns might not appear exactly the same in further samples, the structural aspect will most certainly be present and its quantification gets us closer to assessing the function of these organized sequences. Calls of resident killer whales were initially thought of as simple curious sound that later turned up to be a fingerprint of these animals tight-knit social structures [140]. Bottlenose dolphins, who live in fission-fusion societies develop individual distinctive whistles and exchange them when meeting at the sea [151]. These two examples illustrate how intricately related sounds can be to the animals' social structure [112]. Vocal repertoires only capture one aspect of the vocalisations —the sound types— living temporal aspects aside. However, important biological cues can be encoded in the organisation of the vocal sequences. For instance, phylogeny on songbirds is correlated with syntactic patterns in the songs [152]. Therefore, quantifying vocal sequential patterns opens paths, beyond sound type, for advancing our understanding on the mysterious mammals living under the sea.

Chapter 7

Vocal sequences of parrots

In collaboration with Christian Montes-Medina and Katherine Renton.
Estación de Biología Chamela, Instituto de Biología, Universidad Nacional Autónoma de México.

Many birds combine vocal units into sequences following certain syntactic rules [145, 153, 154, 155, 156] (derived from the term linguistic syntax, set of rules that govern the structure of a sentence, see chapter 5). Quantifying structures within the animal signals can be an important step towards determining their function [70, 157].

Parrot vocal units —so called notes— can carry information like: identity, sex, and micro-geographic differences in their composition and syntax [95, 158]. Few studies have analysed the syntactic structure in parrot vocalisations and these are limited to mated pairs of yellow-naped amazons, *Amazona auropalliata* [158, 159].

In this chapter we aim to identify structures —both temporal and note combinatorial— in the vocal sequences of lilac-crowned amazons (*Amazona finschi*), a parrot species whose syntax has not been studied yet. More specifically, we want to know: whether the notes are emitted with any rhythmic pattern; whether all notes occur with the same frequency or if this depends on the note type; whether certain notes are more likely to be combined than others; among other structural characteristics of the vocal sequences. We use non-parametric statistical methods (described in Chapter 5) to assess syntactic rules in vocalisations from lilac-crowned amazons. Before we enter into the results section some words are said about the parrots and our dataset.

7.1 Lilac crowned amazon

Lilac crowned amazons are an endangered species [111] endemic to the Pacific slopes of Mexico (Fig. 7.1) [160]. These parrots are most vulnerable during early life stages. Eggs and chicks are mainly threatened by predators that include mammals like the coati (*Nasua narica*) and the virginia opossum (*Didelphis virginiana*); reptiles; scorpions; and illegal trade [161]. Adults are neither safe, haws lurk these birds, especially threatening young birds of less than three weeks after leaving the nest [161]. Parrots form flocks to sleep in places known as dormitories [161]. Seasonality in the rainy and dry season affects the birds migration patterns [161] and diet, which consist of seeds and fruits [162, 163, 164].

Pairs of parrots nest inside tree cavities —a limited resource in the forest— where they raise 1-3 chicks for 3 months [165]. Acoustic signals are especially important for cavity nesting birds since these are the only means females have for identifying the male when coming back to the nest after foraging.

7.1.1 Dataset

Vocalisations from 18 free-living lilac-crowned parrots were recorded in the Biosphere Reserve Chamela-Cuixmala (Fig. 7.1), on the coast of Jalisco, Mexico (research permits granted by the *Secretaria del Medio Ambiente y Recursos Naturales*). Parrots were recorded during opportunistic encounters along the nesting season, using Marantz PMD 660 or Marantz PMD 670 solid state digital recorders, and a directional ME66/k6 microphone (Sennheiser Electronic) on a shock-mount pistol-grip.

The recordings were manually annotated [1] using audacity [59], indicating the note type and the temporal coordinates: initial time and duration of the note. Notes were defined as continuous sounds delimited by silences [166] and classified visually according to qualitative spectro-temporal characteristics. In total, 2845 notes were identified and categorised into 17 types (labelled with one or two capital letters). Similar note types with low observation frequency were label as: NL for long notes; NP for shrieking notes; and BF for all other notes observed in 3 or less occasions. Figure 7.2 shows a spectrogram with an annotated tape section and table 7.1 summarises our dataset.

[1] Collection and annotation of the data was done by Montes-Medina.

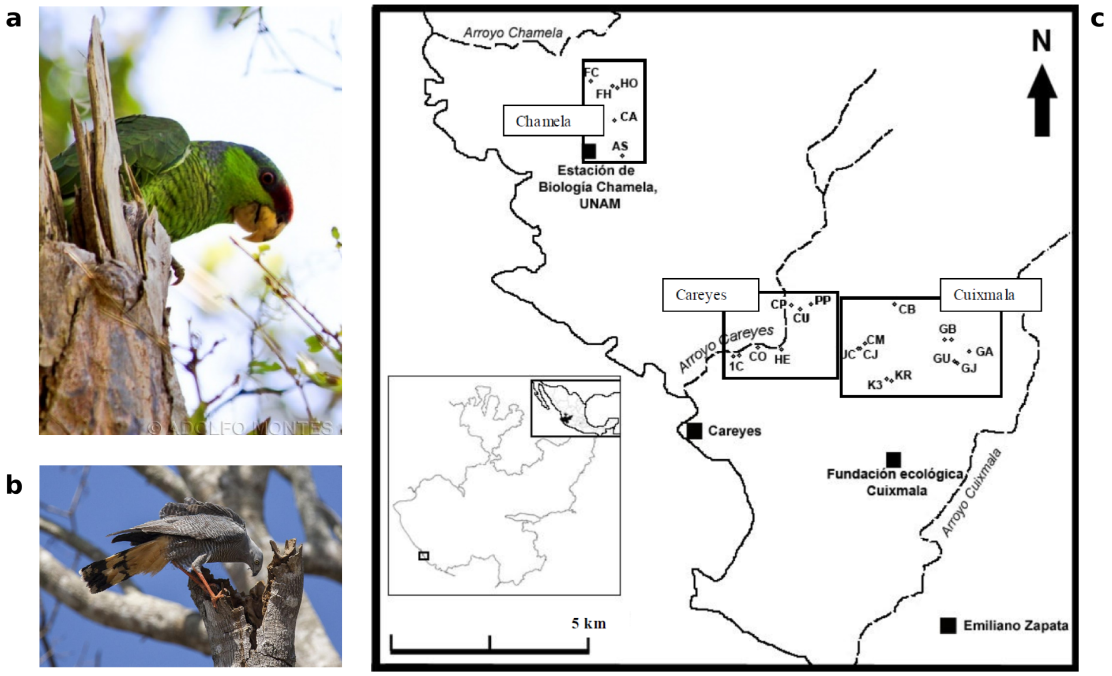


Figure 7.1: Lilac crowned amazon, predator and nest locations. **a**, Picture of a lilac-crowned amazon peering out of the nest (Photo: Montes-Medina) **b**, Picture of a Hawk, a predator bird (CC BY-SA 2.0). **c**, Biological reserve Chamela (19° 22'N 104° 56'W to 19° 35'N 105° 03'W) with the location of the nests in the three regions of Chamela, Careyes and Cuixmala.

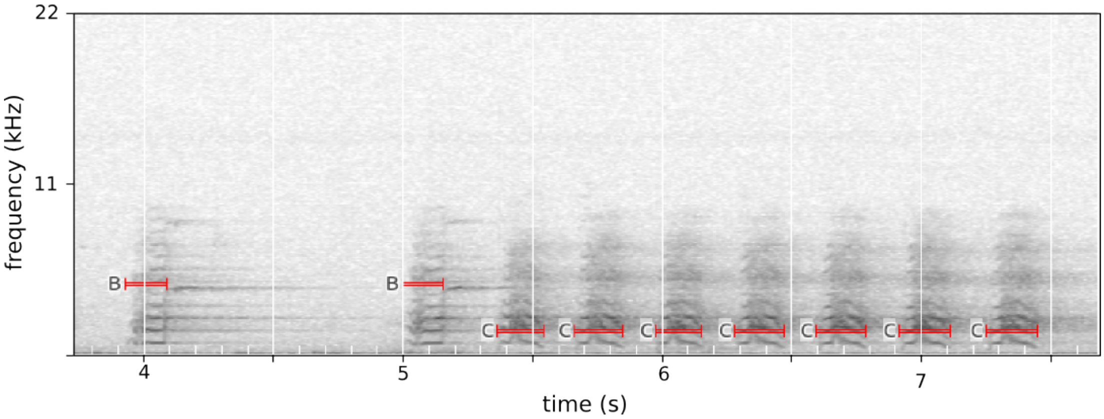


Figure 7.2: Annotated recording. Spectrogram of a recording section, ca. four seconds long. Spectrogram annotated with the parrot's notes. Annotations represented as horizontal bars, indicating the temporal coordinates of the note and the note type. Spectrogram exported from Sonic Visualiser [99].

7.2 Structure in the vocal sequences

Combinatorics is a way to achieve large numbers rapidly. Consider the 17 note types we have in our sample; there are 289 possible combinations of two notes sequences; 4913 combinations of three notes sequences; and the number grows rapidly with the sequence size. From the sample in Fig. 7.2 we know that parrots can at least produce sequences 7 notes long, yielding to 410338673 combinations. However, it is very unlikely that all these combinations occur as birds were to vocalise randomly —it is neither the case for human languages nor for most animal communication systems studied so far. But, which are those patterns? and why do they occur? Are the questions propelling this chapter (yet we only progress on the first one, as for the second question there are multiple theories, a popular one is the Zipf's least effort theory [134, 167]).

Using the statistical tools from chapter 5 we explore parrot vocalisations aiming to shrink the explosive number of combinations to a set of more comprehensible principles shaping their vocal sequences. We start with the descriptions of the temporal structure and note diversity to then move on to the structure within the vocal sequences.

7.2.1 Timing

In this section we consider only the temporal coordinates (ignoring the note type) of the vocalisations to focus on their temporal structure. We present patterns concerning:

parrot	area	# notes	# note types	# recordings
CJ	Cuixmala	181	9	4
GU	Cuixmala	253	9	4
CM	Cuixmala	117	14	4
CB	Cuixmala	107	8	4
CA	Cuixmala	160	9	3
K3	Cuixmala	561	13	6
KR	Cuixmala	184	12	4
GB	Cuixmala	227	9	3
GA	Cuixmala	265	12	4
GJ	Cuixmala	140	13	3
UC	Cuixmala	102	11	4
CO	Careyes	85	7	4
CP	Careyes	17	5	2
1C	Careyes	11	6	2
HE	Careyes	64	11	4
FH	Chamela	155	10	4
AS	Chamela	95	7	3
FC	Chamela	128	13	3

Table 7.1: Summary of the dataset. List of the 18 recorded parrots with the number of: notes, note types and recordings.

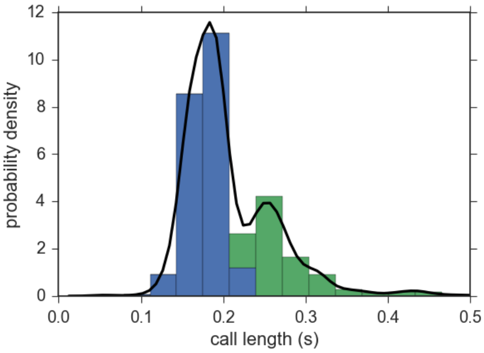


Figure 7.3: Note lengths. Distribution of note lengths in the range zero to 0.5 seconds. Distribution presents two dominant note lengths indicated with different colours, short notes in blue and long notes in green. Short and long scales were ad hoc split at 0.22 s.

note duration, inter-note intervals (here referred as ICI to keep consistency with the previous chapters) and the correlations between these two.

Note length

Our sample had notes with durations that ranged from 0.05 s to 1 s, with 96.7% of them between 0.12 s and 0.35 s (Fig. 7.3). Within this interval, the distribution of note lengths presents two modes: one for short notes around 0.18 s and another one for longer notes around 0.25 s (Fig. 7.3).

Inter-note intervals

Time intervals between consecutive notes ranged from 0.04 s to 129 s in our sample. To disclose the structure over this large range we look at the distribution of the logarithm of the ICIs (Fig. 7.4b). This trimodal distribution suggests that the parrots exploit the time resource emitting their vocalisations using three different time scales roughly separated as follows: one for short ICIs shorter than 0.4 s, one for medium ICIs between 0.4 s and 2 s, and one for long ICIs longer than 2 s. Despite the ICIs having a very wide distribution, 50% of the notes have ICIs shorter than 1.2 s (Fig. 7.4c). Zooming into this range reveals three peaks: two in the short time scales around 0.12 s and 0.2 s (Fig. 7.4b), and another in the medium size time scale around 0.8 s (Fig. 7.4c).

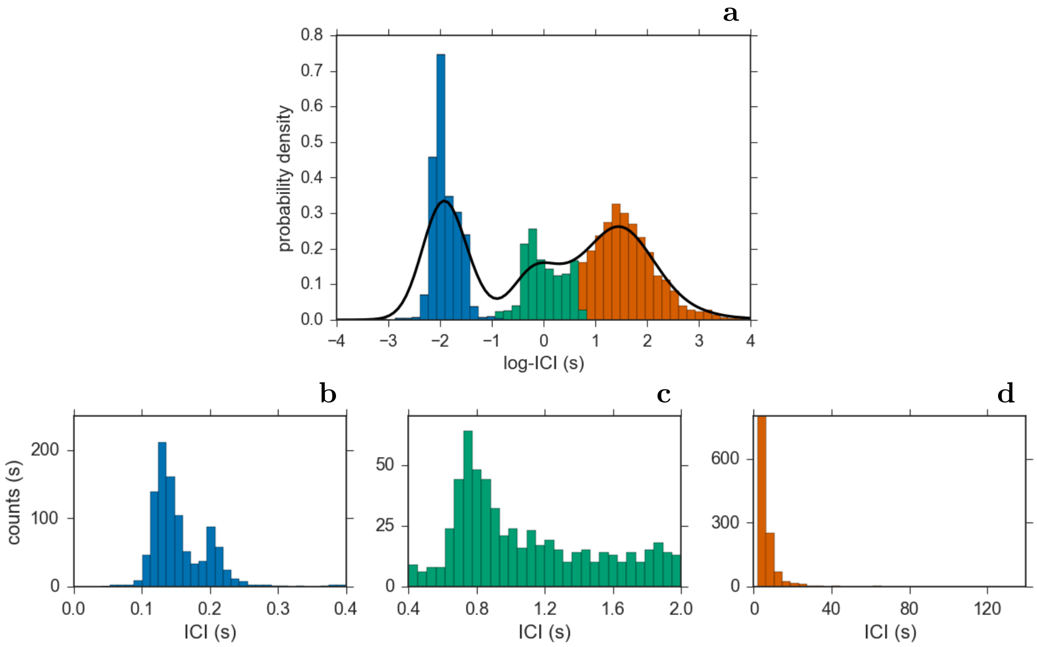


Figure 7.4: Distribution of inter-note intervals (ICI). **a**, Distribution of log-ICIs, highlighting three time scales with different colours. Distribution of ICIs for the time scales: (**b**) short, ICIs < 0.4 s; (**c**) medium, $0.4 <$ ICIs < 2 s and (**d**) long, $2 <$ ICIs.

Correlation between note and silence lengths

A scatter plot between the note length and the succeeding ICI suggests that the two variables are correlated (Fig. 7.5a). The high density of points obscures the patterns in the plot, which can be better appreciated in terms of a joint probability (Fig. 7.5b). This plot shows that: short notes (< 0.4 s) are often followed by very short silences (< 0.4 s), but may also be followed by longer ones; while long notes (> 0.2 s) are rarely followed by short silences and are most likely followed by very long silences (> 1.2 s). As a baseline for our observations in Fig. 7.5c we plotted the joint probability between the note lengths and ICIs assuming no relation between the two (shuffling the data). Differences between the observed distribution and the shuffled one stresses the strength of the correlation we observe between the note and the duration of the following silence (Fig. 7.5b).

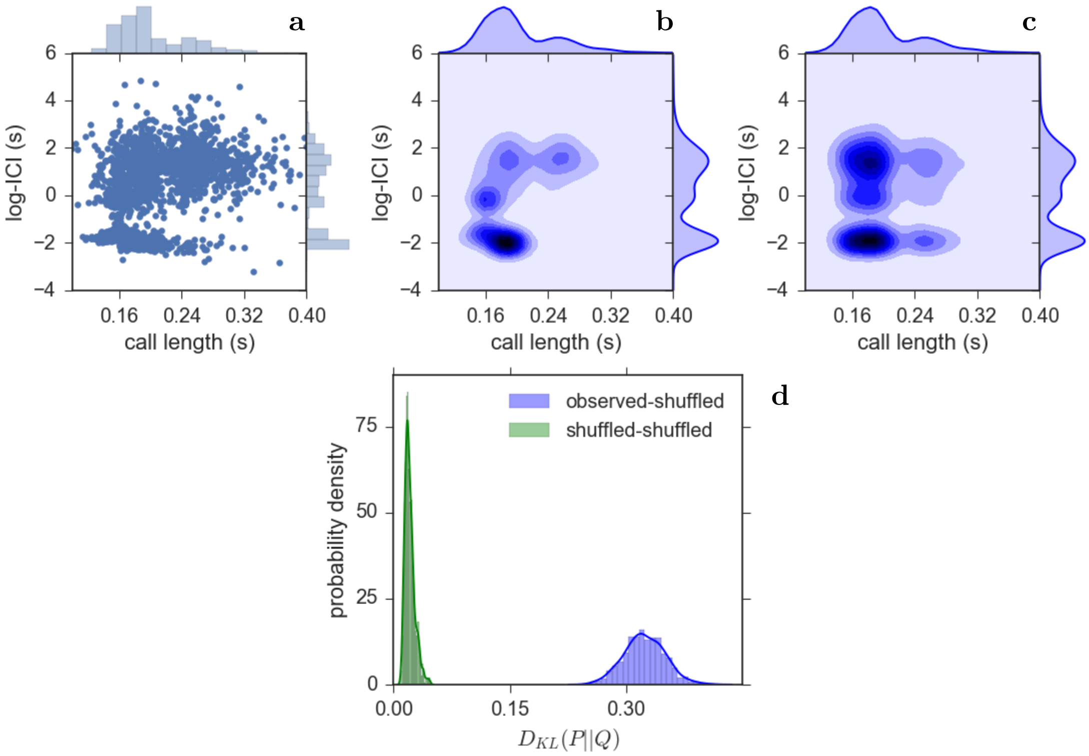


Figure 7.5: Correlation between the length of a note and the length of the succeeding silence as the log-ICIs. **a**, Scatter plot of the note lengths and the log-ICIs. **b**, Joint probability of the note lengths and the log-ICIs. **c**, Joint probability of the note lengths and log-ICIs assuming no correlation between this two, with randomised data. **d**, Distribution of the KL-divergence obtained by comparing the joint probability between the note lengths and the log-ICIs P (panel b) and the joint the probability between these two assuming no correlation, randomising the data (blue distribution). In green the distribution of the KL-divergence between two randomised distributions. For both distributions data was randomised 1000 times.

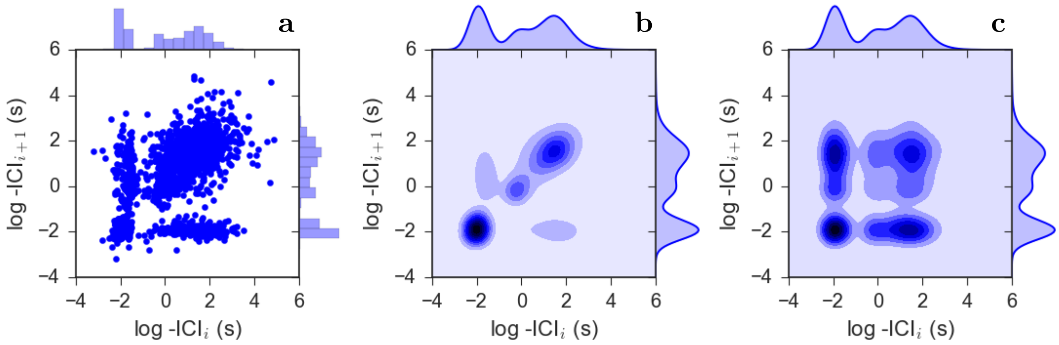


Figure 7.6: Correlation between consecutive inter-note intervals (ICIs). **a**, Scatter plot of consecutive ICIs. **b**, Joint probability of two consecutive ICIs in log scale. **c**, Joint probability of two consecutive ICIs assuming no relation, randomising the data.

Correlation between the ICIs

In Figure 7.6 we explore the correlation between the length of consecutive silences or ICIs. The joint probability in Fig. 7.6b shows that consecutive silences are clustered into time scales, i.e.: short ICIs are most likely followed by short ICIs, medium ICIs by medium ICIs, and long ICIs by long ICIs. Moreover, the joint probabilities are qualitatively different than those expected assuming no correlation (Fig. 7.6c) sustaining the importance of correlation between consecutive ICIs.

The correlation we observe is not limited to consecutive ICIs only but it extends to silences several notes away (Fig. 7.7c). The correlation between ICIs more than 10 notes away stagnates around 0.1 s, but is still higher than expected by chance.

7.2.2 Note composition

The 2852 notes were classified into 17 types. The note's frequencies are not homogeneous but depends on the note type (Fig. 7.8) with notes C and B the most frequent ones representing more than 50% of the notes in our sample.

We observe a large amount of note sharing between the birds, especially for notes A, B and C (Fig. 7.8b and c). Note H5 was only recorded from 4 birds yet these covered the three areas (Fig. 7.8b). All note types were recorded from at least two birds so no bird specific note was observed in our data (Fig. 7.8c).

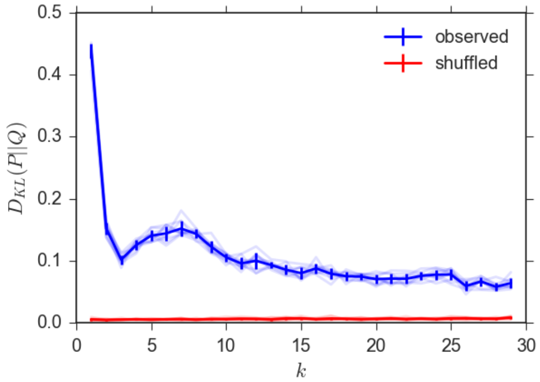


Figure 7.7: Span of correlations between inter-note intervals (ICIs). We use the KL-divergence to compare: the observed correlation between silences at a k-distance (via the joint probability of the two); and the joint probability assuming no correlation between the silences (shuffling the ICIs). Blue line shows the distance between the joint probability of the ICIs at a k-distance $P(k) = p(\tau_i, \tau_{i+k})$ and the joint probability of the shuffled data $Q(k) = p^*(\tau_i, \tau_{i+k})$. The red line acts as a baseline showing the distance between two uncorrelated joint probabilities of ICIs at a k-distance, both obtained by shuffling the ICIs. Data was randomised ten times for each k.

Timing properties —note length and ICI— depend on the note type. The distribution of notes' duration is pretty stable for each note type as it can be confirmed from the proximity of the quartile lines in Fig. 7.9a. Inter-note intervals also depend on the note type (Fig. 7.9b). Firstly, we see that different notes roughly occur at different time scales: notes C, E and C2 mostly occur within short ICIs, note B mostly occurs at short and medium ICIs and the rest of the notes occur mostly within long ICIs. Secondly, the distribution of ICIs prior and after a note is fairly symmetric in most cases, only notes A, B and C, differ from this pattern. Note C belongs to the short time scale regime, but the distribution of silences prior to the note is bimodal while the distribution after the note has only one mode. For note B, the preceding silences are generally longer than the succeeding ones, which also have a bimodal distribution with one mode in the short time scale and another one in the medium time scale. Similarly, the silences after note A are shorter than silences prior to the note. This asymmetry in the distribution of prior and after note silences might be related to the preceding or the succeeding note type respectively. Section 7.2.5 investigates further this hypothesis. Finally, comparing panels **a** and **b** from figure Fig. 7.9 we see that long notes, longer than 0.22 s, rarely occur within short (< 0.4 s) ICIs, in agreement with our observation

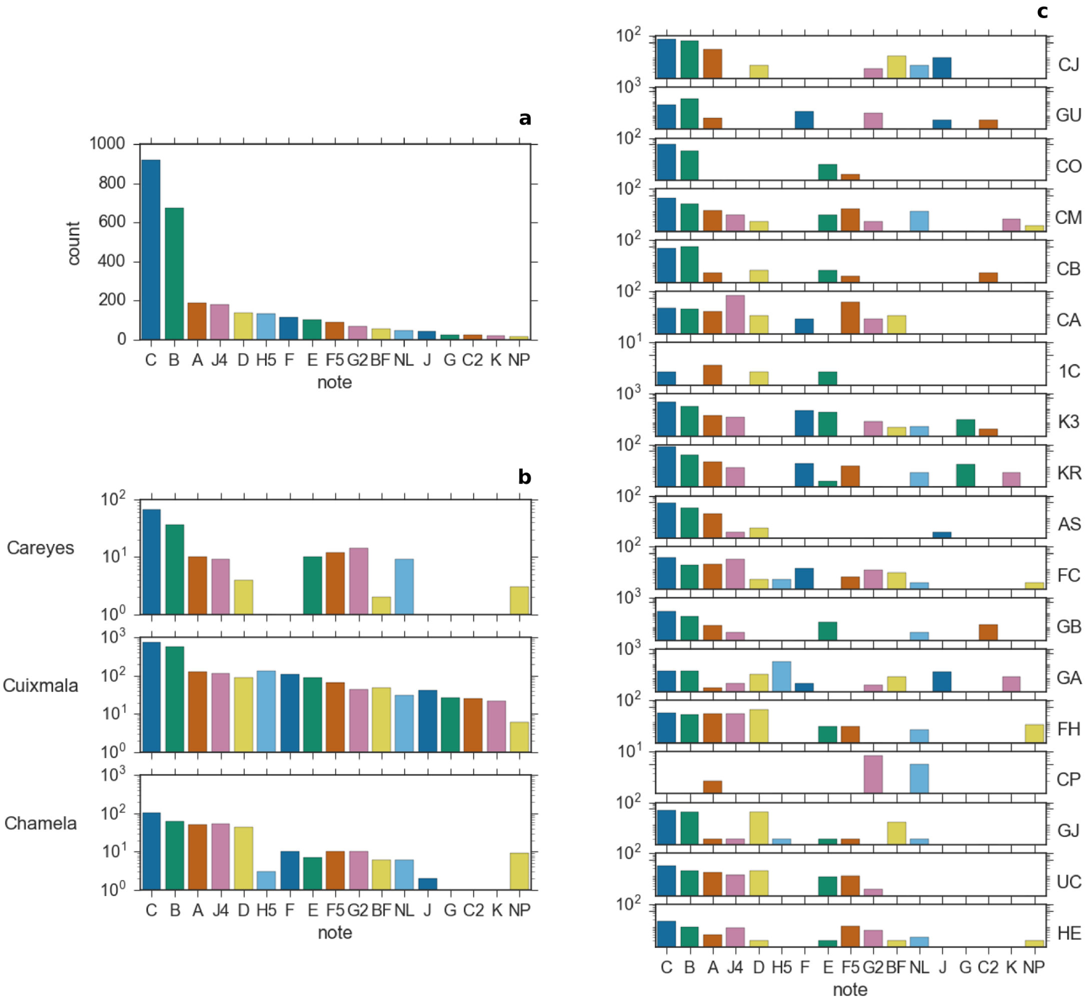


Figure 7.8: Note composition. Note counts for: (**a**) all data, (**b**) by area and (**c**) by bird. Counts in plots b and c are shown in log scale.

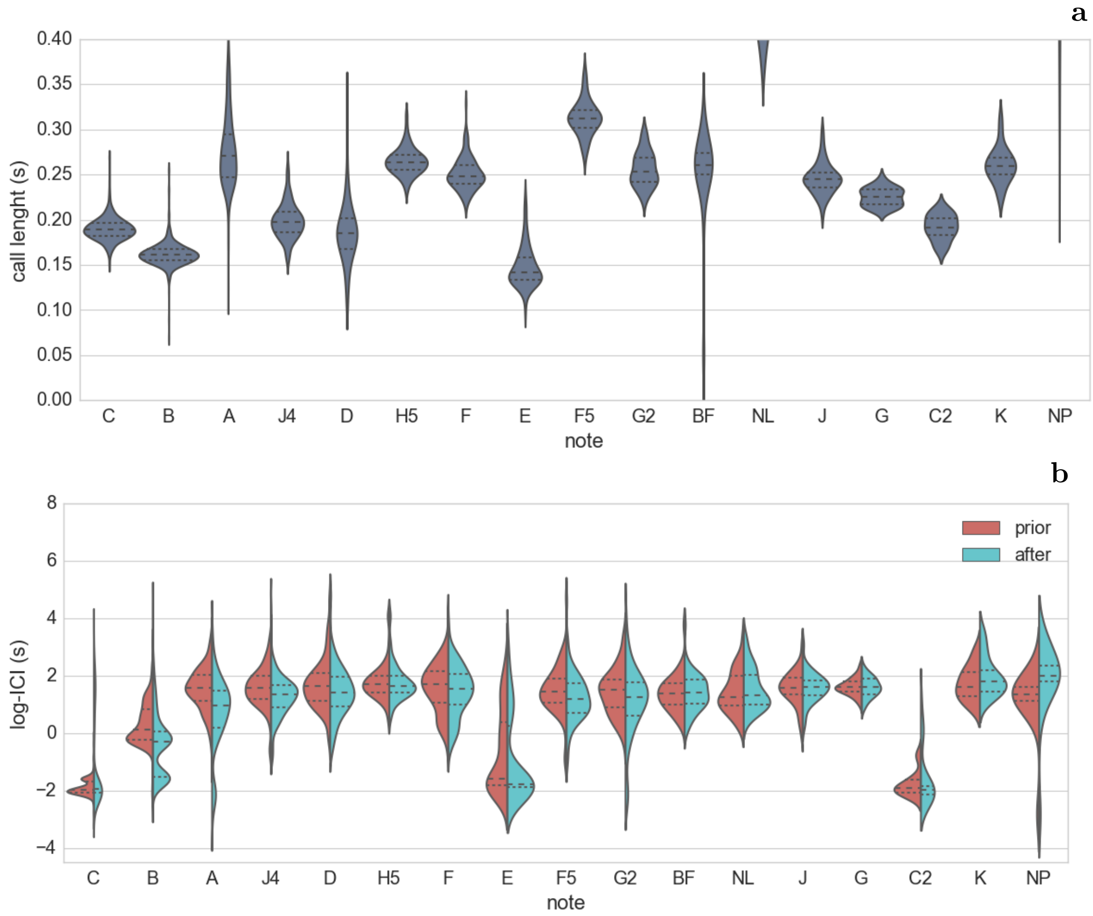


Figure 7.9: Timing properties of the notes. **a**, Note lengths in the range 0 s to 0.4 s with a Gaussian kernel fitting. **b**, Inter-note intervals (ICIs) prior (red) and after (blue) with a Gaussian kernel. Dashed lines inside the distributions indicate the median and doted lines the quartiles. Notes are sorted according to frequency.

from the last section (Fig. 7.5).

7.2.3 Chunk structure

In section 7.2.1 we found that notes are produced either in isolation or within a small vicinity (typically shorter than 0.4 s) of another note. Here we investigate groups of shortly spaced notes in terms of their size and note composition.

Sequences are chains of consecutive notes whose size, or number of chained notes, depends on how sequences are defined. Let τ be the threshold for the longest ICI between consecutive notes in a sequence. A small τ, being very restrictive, would

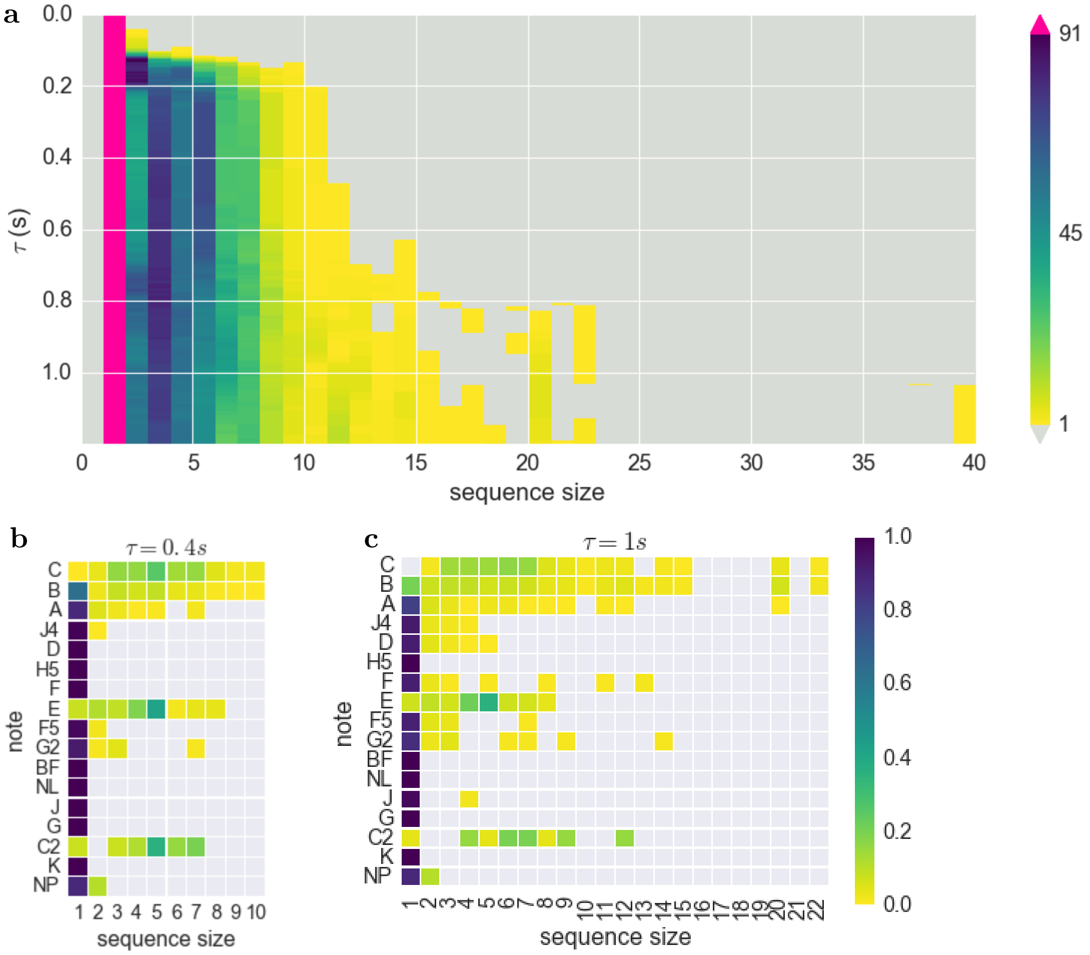


Figure 7.10: Chunk structure of the notes. a, Note grouping as a function of τ in the interval 0 s to 1.2 s. Note composition (normalised for each note) for different sequence sizes at (**c**) $\tau = 0.4$ s and (**c**) $\tau = 1$ s.

hardly yield sequences larger than one, whereas a large τ would group many notes together missing intrinsic structure. From section 7.2.1 we know that more than half of the notes in our sample are separated by ICIs shorter than 1.2 s. Inspecting the chunk structure of the notes in the interval 0 to 1.2 s we observe preference for sequences with 7 or less notes (Fig. 7.10a). For τ between 0.2 s and 0.4 s we notice that the distribution of sequence sizes remains steady (consistently with our findings from section 7.2.1).

For $\tau = 0.4$ s, grouped notes represent 45% of the total notes, with 87% of them in sequences of sizes 3 to 7. Notes occurring in groups are: C2, E and C, and sometimes A and B (a result in agreement with our findings in section 7.2.2). Notes C2, E and C are rarely emitted isolatedly or in sequences of size two and most frequently in sequences of 5 notes (Fig. 7.10b). Most other notes occur only in isolation with the exception of notes J4, F5 and G2 which were recorded in sequences at least once (Fig. 7.10b). At $\tau = 1$ s more than half of the notes in our sample appear grouped. As for note composition with $\tau = 1$ s, the most drastic change happens for note B which occurs mostly in isolation at $\tau = 0.4$ s, while at $\tau = 1$ s occurs mostly grouped (Fig. 7.10c).

Summarising, some notes tend to be produced in sequences of size 3 to 7, whereas other notes are more likely to occur in isolation. The remaining part of the chapter is dedicated to the question of how notes are combined in sequences.

7.2.4 Note combinations and ordering

For $\tau = 1.2$ s most notes appear isolated (Fig. 7.11**a**), and only bigrams C C, B C, C2 C, B B, A B, D B, F F, E E, C2 J, C2 C2 and C2 E (bigram NP NP was not included in the list because NP is a class with different note types) occur more often than expected by chance. Out of these 11 bigrams, almost half of them are note repetitions.

Note C is the most frequent in our sample, but it occurs in very precise ways: either followed by B or in repetitions, and (essentially) never at the beginning of a sequence. Notice also that while B C occurs more often than expected by chance, C B does not (Fig. 7.11b).

For $\tau = 20$ s we observe that notes are often produced in repetitions as indicated by the prominent diagonal in Fig. 7.12 (a and b). With the exception of notes A and J (and possibly note K, but more data would be needed to confirm the pattern), the transition probabilities for the repetition of all other notes are significantly higher (with a significance level $\alpha = 0.1$) than expected by chance.

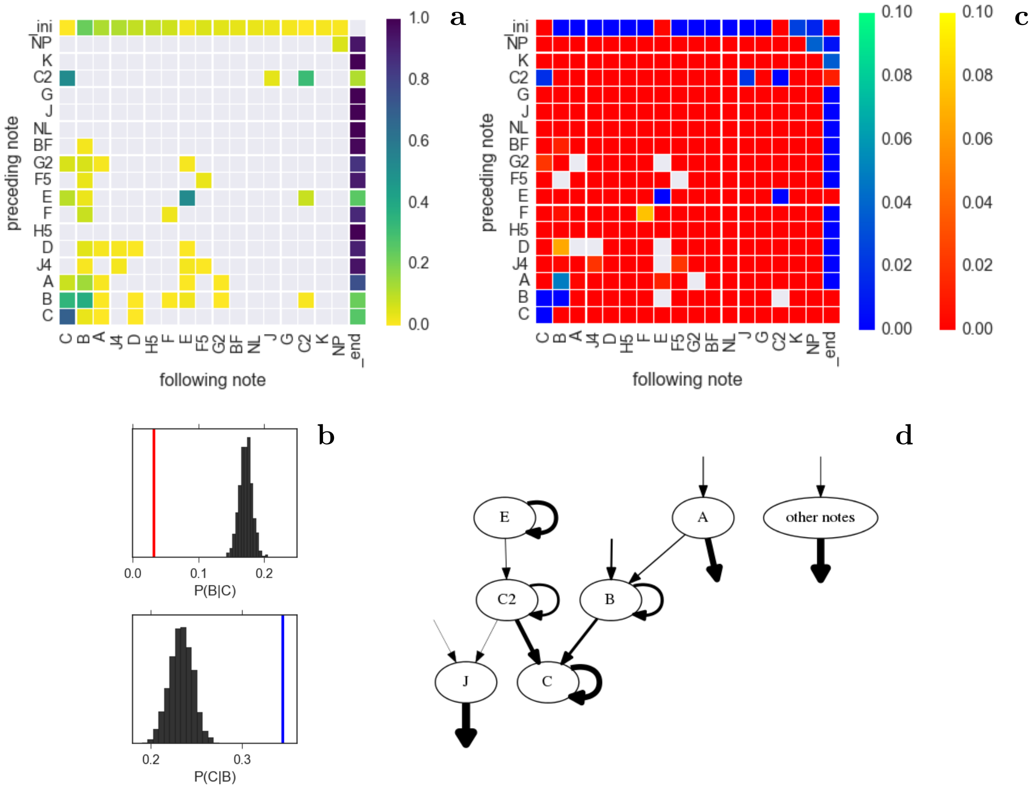


Figure 7.11: Transition probabilities between notes in sequences with $\tau = 1.2$ **s.** **a**, First order transition probabilities between the notes in our sample. Significance for the bigram probabilities is assessed with a permutations test, where the observed transition probabilities are compared with the distribution of the null hypothesis that note order is irrelevant, obtained randomising the notes 1000 times. **b**, Transition probabilities for bigrams C B (top) and B C (bottom); vertical lines indicate observed value, and in black the null hypothesis distribution. **c**, Significance of the observed probabilities; red-yellow colourmap shows the p-values for the bigrams less likely than expected by chance, and blue-green colourmap shows the p-values for the bigrams more likely than expected by chance. p-values larger than 0.1 are shown in grey. Note labels are sorted according to their rank. Labels _ini and _end indicate the beginning and end of a sequence. **d**, Diagram of note (nodes) transitions represented as arrows with widths proportional to the transition probabilities. The node “other notes” represents the rest of the notes not appearing in the diagram which are more likely to occur in isolation. Only transitions with statistically significant arrows are shown.

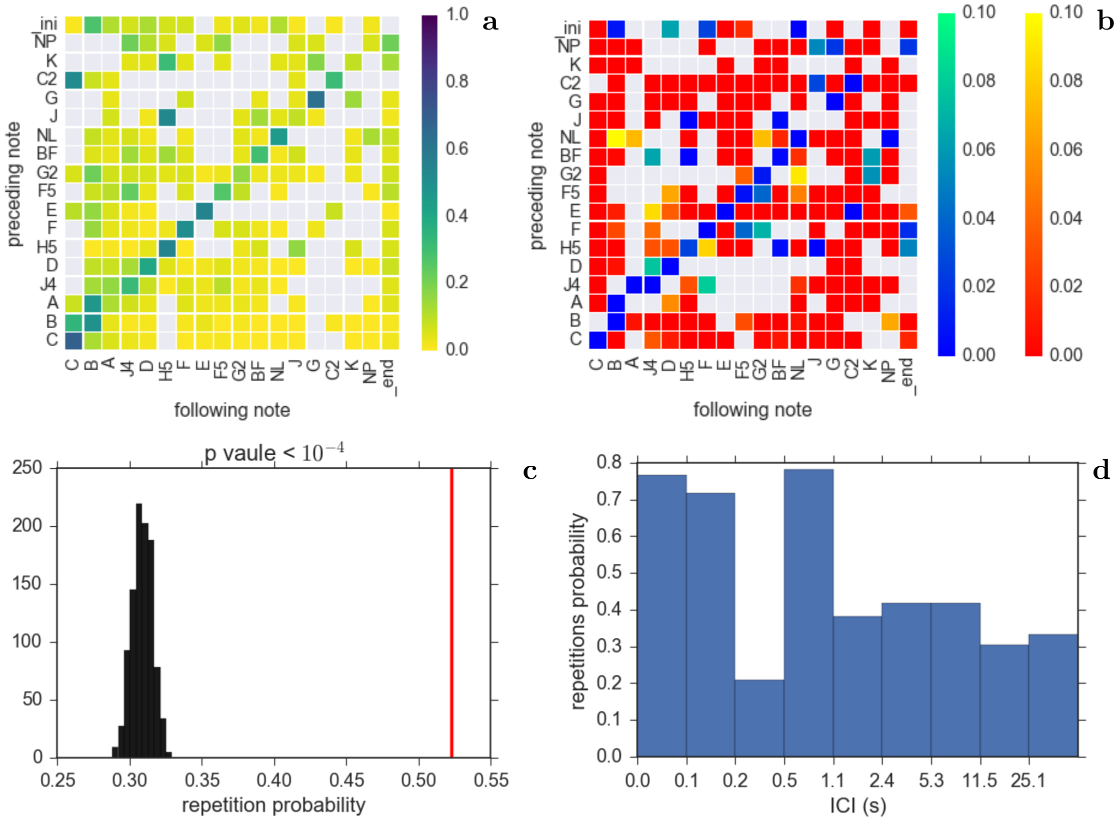


Figure 7.12: Note repetitions. (**a**) Transition probabilities with $\tau = 20$ s and (**b**) p-values of a permutations test carried out to assess the significance of the bigrams. **c**, Probability of having two consecutive notes of the same type within 10 s: observed (vertical line) and distribution of the null hypothesis that notes order is irrelevant. Distribution of the null hypothesis was obtained permuting the notes within a tape 1000 times and then computing the proportion of repetitions for each realisation. **d**, Proportion of bigrams with the same notes for different time intervals, using logarithmically increasing bin widths.

Note repetitions are a prominent pattern. A permutations test over all bigrams in sequences with $\tau = 20$ s confirms that the probability of having a note repetition is significantly higher than expected by chance (Fig. 7.12c). Examining this probability as a function of the ICI (Fig. 7.12d) we obtain that for short ICIs the probability of having repetitions fluctuates largely and for long ICIs the probability of having repetitions stabilises around 0.4 s for ICIs between 1 and 11 s, and around 0.3 s for longer ICIs.

7.2.5 Inter-note intervals of the bigrams

Throughout this chapter we encountered multiple connections between the note types and the duration of the silences between the notes. Here we examine the distribution of ICIs in connection with the transitioning notes.

In section 7.2.1 we found that the distribution of ICIs has two prominent peaks in the short time scale and one more in the medium time scale (Fig. 7.4). Looking at the ICIs by bigram reveals that these peaks correspond to particular note transitions: 0.12 s for C C, 0.2 s for B C and 0.9 for B B (Fig. 7.13a and b). The fact that these peaks are so prominent may be due to the amount of B and C notes occurring in our dataset, yet this suggests that note transitions might be characteristic ICIs. To test the significance of this hypothesis we carry out a KS-test (chapter 5) to compare the distribution of ICIs for each bigram with the distribution of all bigrams and obtain p-values way below 10^{-4}, hence we can safely say that the ICIs depend on the bigram. To get a sense on the similarity of the bigrams in terms of their transitioning times we plot the p-values of the KS-test obtained from comparing the distribution of ICIs for all bigrams (Fig. 7.13c).

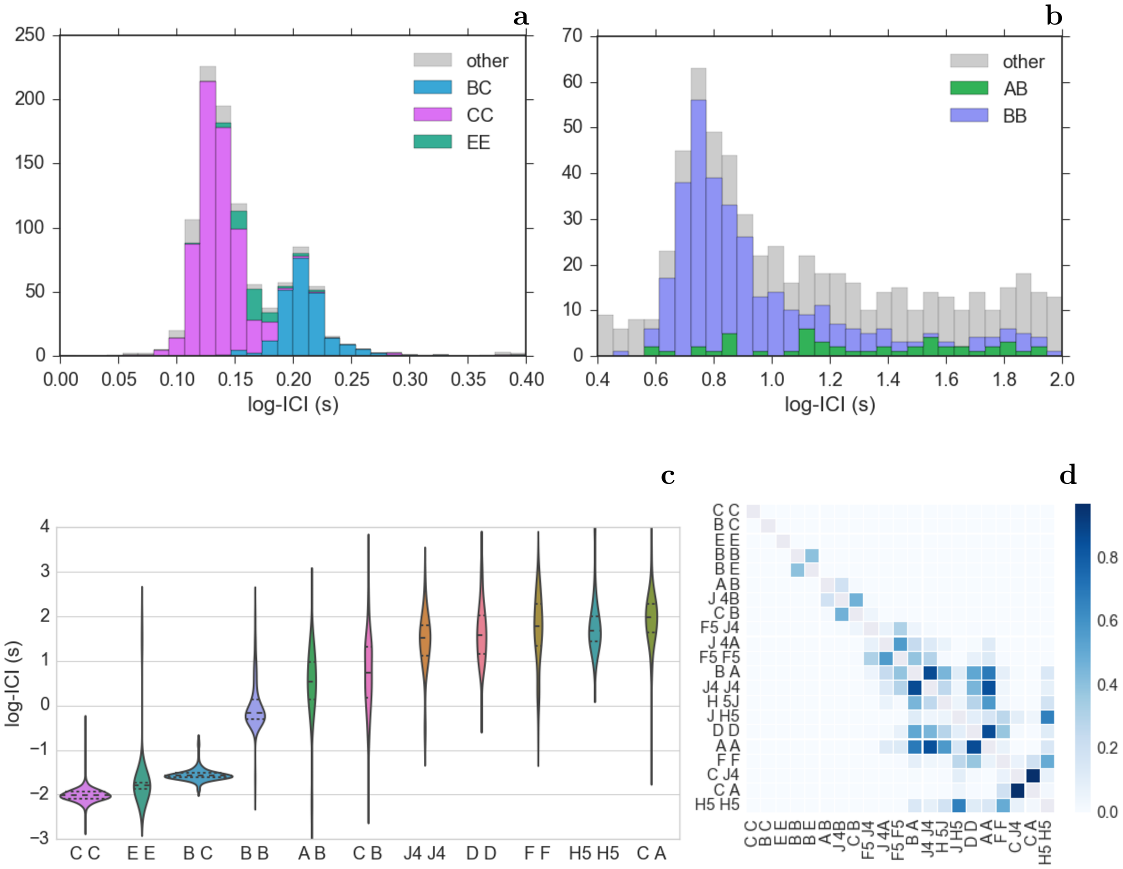


Figure 7.13: Inter note intervals (ICIs) by bigram. Distribution of ICIs with bigrams observed 20 or more times indicated with different colours in ranges (**a**) 0 to 0.4 s and (**b**) 0.4 s to 2 s. **c**, Distribution of the ICIs with a Gaussian kernel fitting for the bigrams with 30 or more observations. **d**, Comparisons of the distribution of ICI by bigram, using the *p*-values of the KS-test between the distribution of ICI of the bigrams against each other. Only bigrams with more than 20 observations.

7.3 Discussion

We found that parrot vocal sequences embed a rich temporal and note combinatorial structure. The encountered patterns are summarised in table 7.2. This list boils down the explosive number of possible note combinations we started with, to a more manageable set of syntactic principles shaping the vocalisations.

We found that inter calling intervals are strongly correlated with the transitioning notes. Similar rhythmic patterns have been observed in songs from sedge warblers [96] and humpback whales [97, 98]. We also found that close inter note intervals are correlated. The strength of the correlation peaks for consecutive ICIs, but remains higher than expected by chance for ICIs as far as 30 notes away. This aspect might be linked to the general mood of the parrot; whether it was agitated, or in a more relaxed state [95].

Vocal signals with one or more repetitions of an element are common across animals [128, 129, 130, 168, 169] and parrots are not an exception. Our sample also contained many note repetitions with a rate much higher than expected by chance. However, addressing note repetitions as a general statement misses important aspects of the anatomy of this common form of animal signalling. In addition to the high amount of repetitions in our sample we encountered that its likeliness depends on the note types, e.g. note A is more likely to be followed by note B than by the same note. Besides the note type, the ICI also affects the likeliness of having repeating notes. Repetitions of note B occur with a high frequency in our dataset and more often than expected by chance, however, for short ICIs around 0.2 s, note B is almost certainly not followed by another note B, but by note C.

One syntactic pattern we encountered often is note B followed by repetitions of note C. This kind of syntactic rule of an introductory note followed by note repetitions is common among birds [144, 170]. For chickadees (*Poecile atricapillus*) it is believed that notes types encode different qualitative information whereas note repetitions encode intensity [144, 170]. An interesting question to investigate in the future would be whether similar syntactic patterns have similar functions across different species.

Animal vocalisations are a form of response to contextual stimulus, finding the link between these two is central to behavioural biology. The function of the syntactic

observed pattern	Figure
Temporal	
• ICI have broad distribution with 3 characteristic time scales:	7.4
– short ICIs < 0.4 s	
– medium ICIs between 0.4 s and 2 s	
– long ICI > 2 s	
• Long note (> 0.2 s) are rarely followed by short ICIs	7.5
• Lengths of consecutive ICI are correlated	7.6
• The ICIs within a tape are more correlated that expected by chance	7.7
Note composition and diversity	
• Some notes are emitted more often that others	7.8
– most frequently occurring notes were C and B	
• No bird specific notes were observed	7.8
– notes A and B were observed for all birds and note C in all but one	
Temporal and combinatorial	
• Notes can be combined into sequences ($\tau = 0.4$ s)	7.10
– notes C, C2, E occur most often in sequences of 3 to 7 notes and rarely in isolation	
– notes A and B occur both in sequences and in isolation	
– all other notes occur in isolation (some outliers for notes NP, G2, F5 and J4)	
• Notes are preferentially combined	7.11
• Bigrams occurring more often that expected by chance ($\tau = 1.2$ s) are:	7.11
– B C, C2 C, E C2, C2 J	
– repetitions of notes: C, B, E and C2	
• Note repetitions occur more often than expected by chance ($\tau = 10$ s)	7.12
• Likeliness of having a note repetition depends on the ICI	7.12
• ICIs depend on the transitioning notes	7.13
– C C 0.12 s	
– B C 0.2 s	
– B B between 0.6 s and 1 s	

Table 7.2: Patterns identified in the parrots vocalisations. The patterns are separated by category: timing, note composition and diversity, and note sequences. The sequence definition parameter τ thresholds the maximum ICI between consecutive calls in a sequence.

patterns here described goes beyond the scope of this study, but is certainly an important question for its potential to recalibrate our place in the animal kingdom. In a recent study, Suzuki et al. [101] found that great tits are capable of compositional syntax; where element combinations are linked to the creation of more complex meanings. Campbell monkeys can alter the meaning of two alarm notes by adding a suffix [26]. The syntactic mechanisms described in the previous two examples are analogous to the way human languages combine words to create sentences [150] —also called lexical syntax— a property that was thought to be exclusive to humans [101, 150, 171].

Note types and inter calling intervals are deeply connected in the vocal sequences from lilac-crowned amazons. These structures open up a series of questions such as whether the strong dependence between the ICIs and the transitioning notes is important for the parrots, or whether parrots mind about the correlations we found between close ICIs. These questions regarding the parrots view point on the encountered patterns certainly deserve further investigation with the aid of playback experiments. While the function of the patterns presented in this study is not known, distinguishing them equips us with framework for asking further questions about the implications of these patterns.

Chapter 8

Discussion and outlook

This thesis proposes methods for studying bioacoustic signals, from the segmentation of animal sounds (part I) —by annotating recordings with vocal units— to the analysis of vocal sequences —by quantifying temporal and call combinatorial structures (part II). For part I, we used supervised machine learning methods to annotate recordings with whale calls and, for part II, we used non-parametric statistical methods to quantify vocal sequences of parrots and pilot whales. Below I summarise and discuss the main outcomes of this thesis, starting from the methods and results of part I, then continuing with the methods and results of part II, and ending with a discussion of how the developed methods and the study of parrots and pilot whales may contribute to the ongoing search for the origin of language.

8.1 Automatic annotation of bioacoustic recordings

In part I of this thesis, we propose a framework for automatically annotating sound files (Chapter 2) and we use it for the two bioacoustic tasks of: (i) detecting tonal sounds of pilot whales (Chapter 3), and (ii) classifying pilot whale call types (Chapter 4). Both tasks were supervised, trying different spectral features (Appendix B) and scanning over a large range of parameters that regulate the spectral and temporal resolution.

Optimal features depended on the task. For the detection task, Mel-spectral features outperformed MFCC by 4% to 10% (depending on the dataset) whereas for the call classification task MFCC outperformed Mel-spectral features by 8%. This asymmetry on the performance of the features can be understood in terms of the differences between the tasks. Whale calls are tonal sounds, so detecting them is mostly a matter of comparing the distribution of the power spectral densities over time. Call phonology is not relevant for the detection of whale calls, thus the MFCC's effectiveness in compressing spectral information does little for this task. On the other hand, the classification of whale calls is all about comparing the evolution of phonological properties. This is confirmed by the way MFCC features seized relevant information for this task —summarising effectively spectral information while allowing a better temporal resolution through the number of temporal slices. The way these two feature representations assist the learning tasks carried out here stresses the importance of carefully choosing features that suit a particular task.

The modular architecture of the code developed here enables to easily combine different feature extraction mechanisms and machine learning estimators for machine learning tasks with audio data. This framework is not limited to the analysis of bioacoustic signals but can be used for other machine learning tasks such as speech processing and music classification. Having a flexible framework is desirable because on the one hand there are multiple paths for processing audio signals, and on the other the outcome of a machine learning task depends on a variety of factors —the extracted features, the machine learning algorithm, the dataset, and the aim of the task— it is not easy then to know in advance which combination of parameters will yield best results.

Propelled by the big data revolution of the last years machine learning has gained great popularity. Similarly, audio signal processing is a popular science given the multiple applications in the processing of music, speech and other sounds. Standard algorithms for both tasks —machine learning and audio signal processing— are available on multiple programming languages. One of the contributions of this thesis was to develop code in the Python programming language to combine the sklearn Python module for machine learning [57] and the librosa Python module for audio signal processing [58]. The main features of the developed code are (i) the ease to process raw audio data together with annotation files through the stacking of preprocessing and feature extracting steps and (ii) the wrapping of these steps into machine learning experiments. This code can be cloned from the github repository *pylotwhale*.

8.2 Quantifying animal vocal sequences

The second part of this thesis proposes using non-parametric statistical methods to quantify temporal and combinatorial patterns in animal vocal sequences (Chapter 5). We applied these methods to investigate vocal sequences of pilot whales (Chapter 6) and parrots (Chapter 7). Below I discuss the main outcomes regarding part II of this thesis: first in terms of the proposed methods and then by comparing the structures found for parrots and whales.

8.2.1 Framework for quantifying vocal sequences

Investigating animal vocal sequences as a segmented time series —sound being the time series and the call types the segments— allowed us to delves into the temporal and combinatorial structures of the sequences in great detail. Each segment is characterised by a triad with one categorical variable for the call type and two numerical variables with the beginning and ending times of the call. The segments were handed as annotation files, a format that advantages other ways of investigating vocal sequences as, for instance, with the usage of cut audio files and excel tables. Annotation files summarise acoustic information in a format standard across audio processing software. Being simple text files information can be easily retrieved, facilitating error checking at any point of study —something that with cut audio files would otherwise be challenging. Here, we annotated the recordings with the call type, however they may be annotated with other kind of information such as the caller, as was done in [172] where the authors investigated communication networks in birds.

The distribution of inter calling intervals (ICIs) for both species studied here showed a wide distribution across multiple time scales that ranged from hundredths of second to minutes. This is an expected feature of the ICIs since animals adjust their calling rates, being more vocal in some contexts than in others [94, 95, 96, 97, 98]. Analysing the inter calling intervals in terms of the log-ICI revealed rich short time scale structures together with long time scale structures of the few —yet not negligible— large ICIs. Logarithmic transformations have also proved to be useful in neuroscience when analysing inter-spiking intervals [173].

Structures in animal vocal sequences have been reported for a diversity of taxa (for reviews, see [70, 100]). Identifying these structures can be an important step

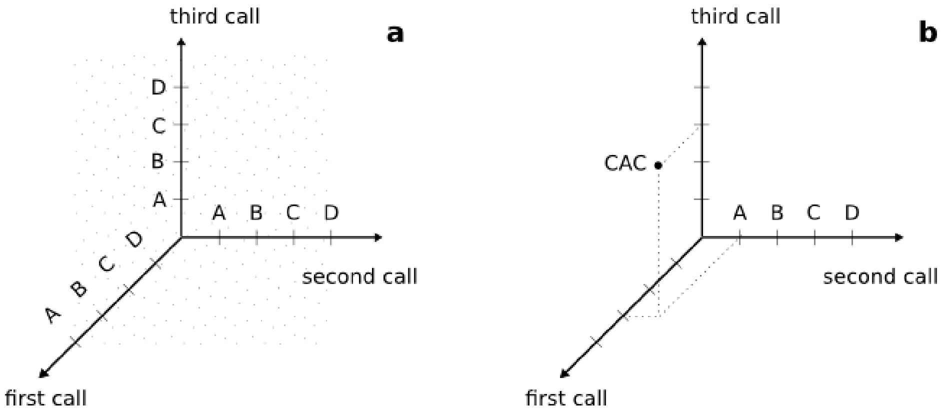


Figure 8.1: Phase space of vocal sequences. Illustration of a phase space for sequences with a maximum of three calls. **a**, calls produced randomly. **b**, sequence CAC.

in determining their function [70]. If we think of the vocal sequences as points in a discrete phase space of dimension equal to the number of elements in a sequence (Fig. 8.1), randomly combined units would populate the phase space uniformly. No animal is expected to vocalise randomly, so identifying sequential patterns shrinks the explosive number of points in the vocal phase space to a set of more comprehensible syntactic principles that might be related to the context and to the behaviour of the animals.

8.2.2 Comparison between the vocal sequences of whales and parrots

Both species studied in this thesis, long-finned pilot whales (Chapter 6) and lilac crowned amazons (Chapter 7) produce richly structured vocal sequences. In the samples of both species the temporal structure is strongly linked to the call/note types —two aspects rarely studied together. The distribution of ICIs, the likeliness of the calls to occur grouped, and the distribution of ICIs for specific bigrams were highly dependent on the call types.

The vocal patterns from pilot whales appear to be less rigid than those identified for the parrots. However we should not go too far with this comparison, first because sample sizes are different and second because the identified patterns have different interpretations. The parrots' dataset only contains notes emitted by males during the nesting phase with only one bird singing in each recording, so we know who the signaller is. In turn, the pilot whales' dataset comes from an encounter with an estimated 45 animals, so we do not know whether the vocal sequences come from one animals or several. Not knowing the caller is not an obstacle for quantifying such vocal structures —if a specific pattern occurs significantly often its prevalence is pointing at an intrinsic structure, regardless of whether it comes from one or more signallers. An example of structured sequences orchestrated by two individuals are duets reported for diverse taxa including whales [174, 175, 176] and birds [152, 159, 177]. Further, outliers of prevalent vocal patterns draw attention to extreme episodes that might be reflecting contextual outliers or behavioural outliers.

8.3 Outlook on the quantification of animal vocal sequences

The timing patterns quantified here —ICIs and call duration— are highly intertwined with the call types, and their combinatorial structures (syntax-like structures). Human speech also exhibits timing patterns as pausing, word lengthening, and rhythm. In the context of linguistics these aspects fall under the scope of prosody.

Prosody —sometimes referred as the musicality of speech— studies how acoustic patterns, such as rhythm, stress and intonation, contribute to meaning. Humans use prosody to communicate a variety of features of the speaker that range from involuntary emotional states to cues that aid comprehension through irony, sarcasm, or the form of a statement e.g. question or command. Additionally, prosodic modulations help coordinate communication by signalling turn taking between interlocutors [178, 179, 180] and aid syntax acquisition in infants [181, 182, 183].

Structures with potential prosodic content were quantified for both whales and parrots species studied here. Specifically, we observed that pauses (ICIs) are highly dependent on the transitioning calls. The correlations between the call lengths and pauses, and the long range correlations between the pauses, contribute to quantify rhythm. The length of some calls varied widely while other call types had more precise lengths. For instance, the length of call 128i from the pilot whales and note A from the parrots varied more than 50%. Whether these prosodic variations convey information to the receiver is a question that may be addressed conducting behavioural experiments.

In addition to the temporal prosodic dimension quantified here, sound's amplitude and spectrum are as well prosodic dimensions important in animal interactions [3]. So, further studies may find additional structures within these dimensions.

Animals can interact emotionally through prosody. High levels of arousal in primates [184, 185, 186] pandas [187] and elephants [188, 189] can be expressed through prosodic features such as an increase in call duration and elevation of pitch. Another form of prosodic interaction are the duets reported for a variety of birds [190] and the antiphonal vocalisations reported for some species of primates [191, 192, 193] and elephants[188].

Prosodic modulations in the voice have been suggested as an evolutionary precursor for language [1, 194, 195, 196, 197, 198, 199]. Insight into this hypothesis can be achieved through comparative studies across animals [3]. In which, similarities in closely related species, like primates and other mammals, might be derived from a common feature [3, 150] whereas similarities in distant species, like birds, may point at environmental forces favouring the convergent evolution of a trait [3, 150]. Quantifying vocal sequences of animals by including prosodic content may contribute to understanding how language originated. Quoting Darwin

> The suspicion does not appear improbable that the progenitors of man, either the males or females, or both sexes, before they had acquired the power of expressing their mutual love in articulate language, endeavoured to charm each other with musical notes and rhythm.

Appendix A

Physics of sound

Sound is an important communication channel for humans and most animal species. Consider for example a quiet garden with birds and insects singing in the background in comparison to a bar filled with bouncy, chatty echoes on a music background. According to the sounds around us we adjust our speech's volume, tone and pace. Sound also predisposes our brain. For instance, a newborns' perception of speech sounds is influenced by prenatal maternal speech [200]; sound can alter our visual motion perception [201] and even induce visual illusions [202]. Sound perception (hearing) is an involved neurophysiological process specific to each animal. However sound is a physical phenomenon, and understanding its physics can be insightful in trying to adopt an animal's perspective.

Sound is a vibration that propagates as a longitudinal wave when the particles of a medium oscillate, deforming it into regions of high and low pressure (Fig. A.1). This deformation travels from its source at a speed c transporting acoustic energy.

At the point of reception —e.g. ear or microphone— sound is determined by its pressure levels as a function of time, $p(t)$. Bringing this function to the frequency space helps us understand sound better. Take for instance the tone A of the third octave whose frequency is 220Hz. This sound is described by the equation

$$p(t) = a\cos(2\pi f t). \tag{A.1}$$

where a is the amplitude, associated to the loudness, and f is the frequency of the wave, associated to the pitch. While loudness and pitch are two subjective perceptual characteristics, the amplitude and the frequency of a wave are two objective physical

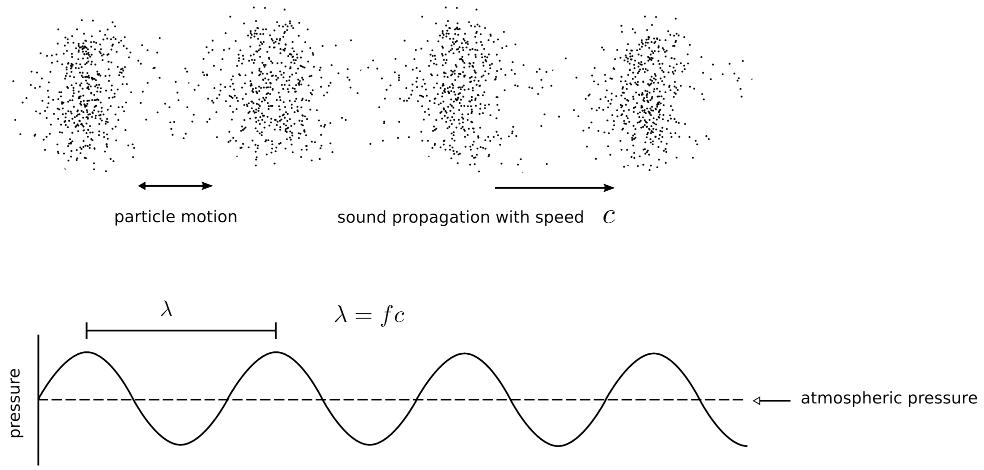


Figure A.1: Sound propagation. Illustration of a longitudinal pressure wave with wavelength λ, frequency f and speed c.

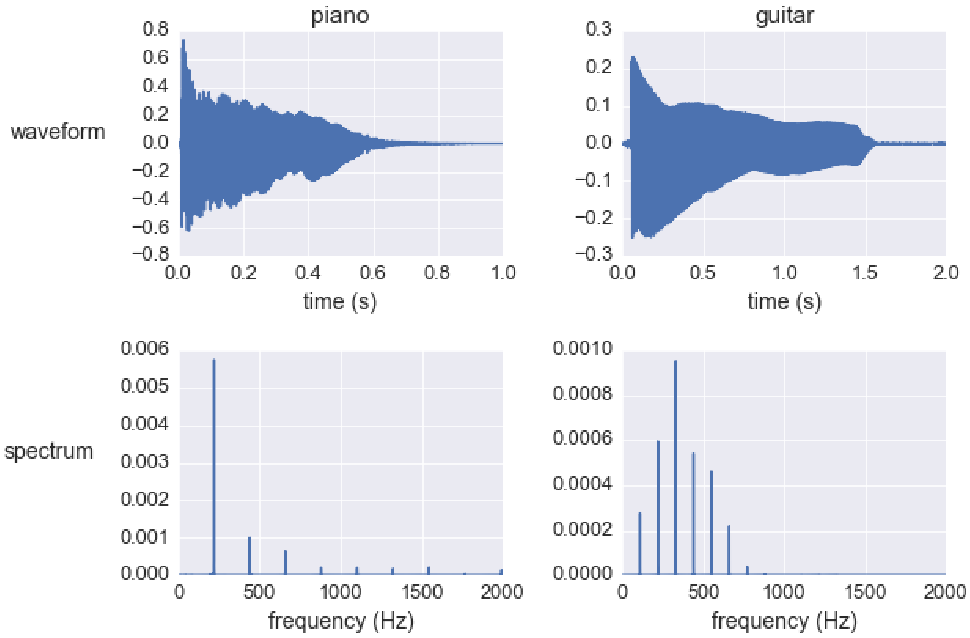


Figure A.2: Note A of the third octave (220Hz) of a synthetic piano and a guitar. The timber is the characteristic that allows us to differentiate the sounds from two instruments even when playing the same note. Physically this is related to the temporal (upper row) and the spectral (lower row) envelopes of the sound. The spectral envelope is characterised by the relative intensities of the harmonics.

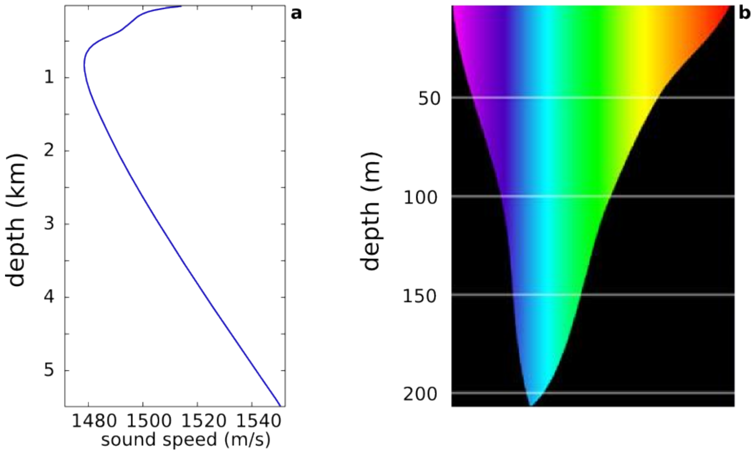


Figure A.3: Underwater sound speed and light absorption as a function of the depth. **a**, Speed of sound underwater as a function of the depth. Physicochemical conditions of sea water (e.g. pressure, temperature and salinity) change with the depth which affects the speed of sound. (Image by Nicoguaro CC BY-SA 4.0, via Wikimedia Commons). **b**, Light gets attenuated underwater due to light absorption which differs for the wavelength. (Image by National Oceanic and Atmospheric Administration, U.S. (Public domain).

characteristics. The tones produced by instruments are not composed of a single frequency, as the one of Eq. A.1, but of a collection of them. These frequencies are usually related harmonically with specific intensities for each instrument, which is why we can differentiate between an A when played on a piano or on a guitar (Fig. A.2).

Our ears respond to sound with frequencies between 20 Hz and 20kHz [203]. Sound is not limited to the human spectral range; many insects, baths and toothed whales can hear ultrasonic sounds (above 20kHz), and, on the other side of the spectrum, animals such as elephants and some baleen whales can communicate with infrasonic sounds (below 20Hz).

The human ear can resolve sounds with pressures between 20 μPa and 200 Pa. Because of this wide range —along 8 orders of magnitude— and our almost logarithmic perception of sound intensity, it is customary to measure sound pressure levels (SPL) in decibels (dB$_r$), a logarithmic scale relative to a reference pressure p_r. Thus a pressure p in dB$_r$ is given by

$$SPL = 20 \log_{10}\left(\frac{p}{p_r}\right). \tag{A.2}$$

The logarithms has no units so the decibels are not units in the typical sense but indicate a relative sound level; thus the importance of mentioning the reference level when reporting measurements in decibels. The decibels in air are different from the decibels in water, for the former $p_r = 20\mu$Pa and for the latter $p_r = 1\mu$Pa.

Another difference between sound in air and in water is its propagation speed. The speed of sound is determined by the physical properties of the material and is given by the equation

$$c = \sqrt{\frac{K}{\rho}}, \tag{A.3}$$

where K is the elasticity bulk modulus of the medium and ρ the density. Water is almost 1000 times denser than air but its elasticity bulk modulus is more than 15000 time greater than that of air, yielding underwater sound speeds more than 4 times faster than those of air.

Electromagnetic waves get quickly absorbed underwater due to the high density of the medium (Fig. A.3). Sound, instead, waves can travel long distances [38, 39]. Humans use sound for underwater communication and navigation. Underwater acoustics is used in oceanographic and marine animals research, industrial applications such as fishing and seismic oil exploration and military defence with sonars. Marine animals also use sound for communication and navigation. Marine mammals have evolved specialised ears for listening underwater; some toothed whales can echolocalise being able to reconstruct 3D spaces out of sound. The excellent conditions for sound propagation underwater also make this medium highly vulnerable to acoustic pollution.

Appendix B

Spectral representations of sound

B.1 Spectral features

Having the right features is for a machine learning task like having the right glasses for a shortsighted like me. Certainly the right features depend on the classification task, and when working with sound, spectral features capture well the characteristics perceived by humans, such as the pitch. So analysing sound in the frequency domain often does a good job in machine learning tasks involving sound data.

One can decompose sound into its spectral components (Fig. B.1a-b) using Fourier analysis. For digital signals this is done with the discrete Fourier transform (DFT). Given a digital signal one dimensional $x(n)$ with N samples $n = 1, \dots, N$, its discrete Fourier transform is

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi kn/N}, \tag{B.1}$$

where $e^{-i2\pi kn}$ is the k-th element of the Fourier basis with $k \in \bar{N}$. Each basis element describes a complex sinusoid with a particular frequency k. $X(k)$ is a complex function that weights k-th spectral component of $x(n)$. The power spectral density $P(k)$ measures the energy content of the signal in the frequency domain and is given by

$$P(k) = \frac{1}{N}|X(k)|^2 \tag{B.2}$$

The fast Fourier transform is an algorithm for numerically computing the discrete Fourier transform effectively [204]. The algorithms requires the number of samples N to be powers of 2.

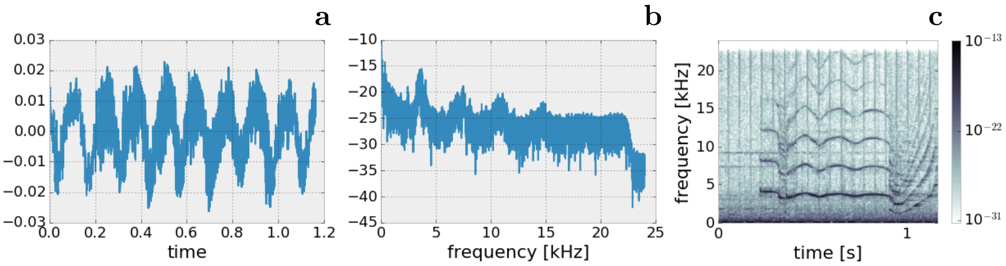


Figure B.1: Sound of a whale in the (**a**) time domain (waveform), (**b**) in the frequency domain (power spectral density) and (**c**) in a mix of the temporal and frequency domain, as a spectrogram. Power spectral density in log scale, computed with $N = 256$ and 50% overlap.

Below we describe 4 different feature representations based on the spectral decomposition of sound.

B.1.1 Spectrogram

The ear processes sound spectral information dynamically. So, neither representing sound in the time domain (Fig. B.1a) nor in the frequency domain (Fig. B.1b) is adequate. While the first one lacks spectral information the later lacks temporal information. An in between point is the short time power spectral density —or spectrogram.

The spectrogram is computed by framing a signal into short time windows and computing the power spectral density for each frame. Given $x_i(n)$, the i-th frame of a signal $x(n)$, its power spectral density $P_i(k)$ is given by

$$P_i(k) = \frac{1}{N}|X_i(k)|^2, \tag{B.3}$$

where $X_i(k)$ is the DFT of $x_i(n)$. The number of samples per frame N determines the spectro-temporal resolution. Larger windows increase the spectral resolution and decrease the temporal resolution whereas smaller windows increase the temporal resolution and decrease the spectral resolution. The datasets we work in this thesis have sampling rates around 40 kHz, for which window sizes between 256 samples and 1024 samples are a good trade off between frequency and temporal resolution. Framing the signal with a rectangular window causes power leakage, which can be reduced using windows like the Hanning or Hamming. Additionally, window effects in the borders can be counteracted analysing overlapping frames.

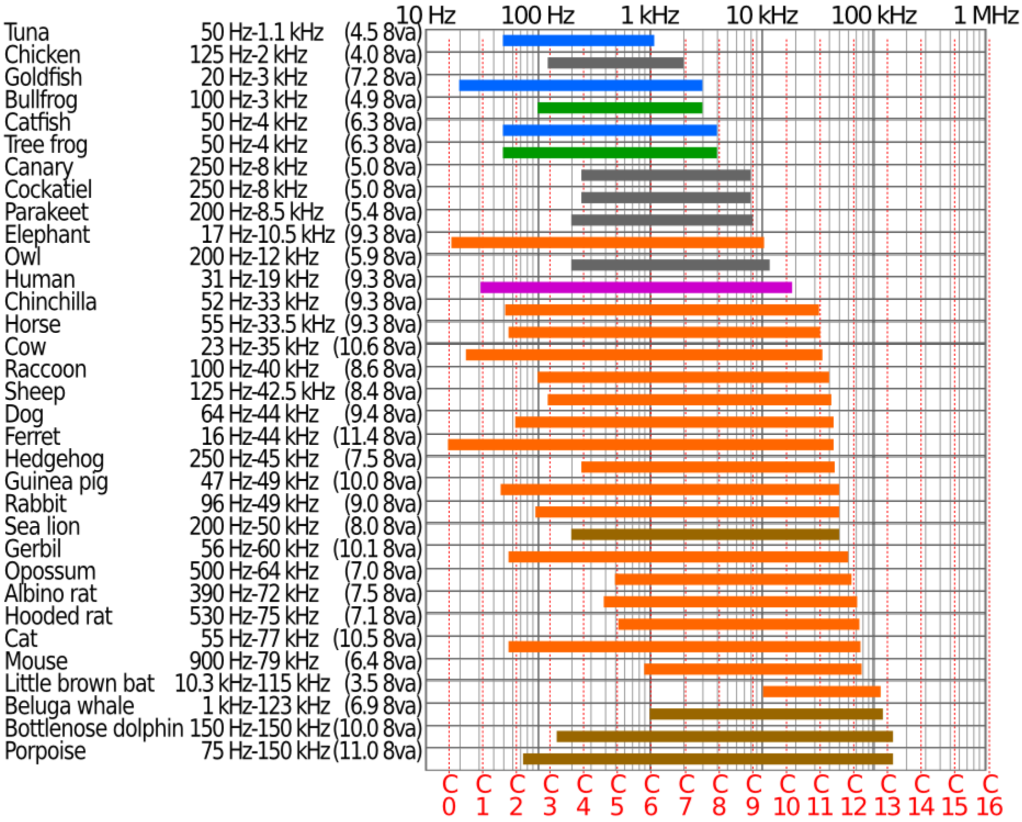


Figure B.2: Hearing ranges of animals. (Image By Cmglee - Own work, CC BY-SA 3.0, via Wikipedia)

One disadvantage of the spectrogram is its high dimensionality. Take for instance a one second signal with a sample rate of 44kHz and $N = 512$, without overlap. Encoding this signal with a spectrogram requires a 21760-dimensional vector (85 time bins times 256 frequency bands). The temporal axis can be summarised by averaging the frequencies over the time frame (more a about it will me mentioned in section B.2) but this still leaves 256 frequency components. The features discussed below comprise different ways of summarising the spectral information.

B.1.2 Mel-spectrogram

The range of frequencies humans and many mammals can resolve extends over several orders of magnitude (Fig. B.2). The range of frequencies differs for different species but the wideness of the frequency range is present across species. Humans can hear sounds with frequencies between *ca.* 20 Hz and 20 kHz; below our hearing range sound is referred as infrasound and above our hearing range sound is referred as ultrasound.

The sensitivity is not the same for all the frequencies along this wide range. We resolve low frequencies better than we do for high frequencies. The mel-scale is a logarithmic frequency scale that tries to resemble human perception [205], mapping frequencies to mel-scale frequencies (Fig. B.3a)

$$M(f) = 1125 \log_{10}(1 + f/700), \tag{B.4}$$

with f the frequency in Hz. f is related to k through the sampling frequency f_s via

$$f(k) = \frac{kf_s}{N}. \tag{B.5}$$

The mel-scaling can be implemented with a filterbank. A filterbank splits a signal into different regions of the spectrum (Fig. B.3b). For digital signals, the mel-scale filterbank can be implemented as a collection of triangular filters centred around f_c [206]

$$H_m(k) = \begin{cases} 0 & \text{for } f(k) < f_c(m-1) \\ \frac{f(k)-f_c(m-1)}{f_c(m)-f_c(m-1)} & \text{for } f_c(m-1) \leq f(k) < f_c(m) \\ \frac{f(k)-f_c(m+1)}{f_c(m)-f_c(m+1)} & \text{for } f_c(m) \leq f(k) < f_c(m+1) \\ 0 & \text{for } f(k) \geq f_c(m+1) \end{cases} \tag{B.6}$$

where $m = 1, \ldots M$, with M the number of filters. The mel-spectrum of the i-frame is given by

$$\tilde{X}_i(m) = \log_{10}\left(\sum_{k=0}^{N-1} P_i(k) H_m(k)\right). \tag{B.7}$$

Notice that the high frequencies are not dismissed but encoded with less resolution than lower frequencies (Fig. B.3c).

B.1.3 Mel-frequency cepstrogram

The mel-frequency cepstrogram (MFC), commonly referred as the mel-frequency cepstral coefficients (MFCC) are widely used in speech recognition systems for they can efficiently encode sound features. Their efficiency lays in the compression of redundancies in the harmonics by taking the discrete cosine transform (DCT) of the mel-scale spectrogram (Fig. B.1a). The l-th MFCC for the i-th frame is given by

$$c_i(l) = \mathrm{DCT}\{\tilde{X}_i(m)\}_l \tag{B.8}$$

For speech recognition tasks typically only the first 20 cepstral coefficients are used.

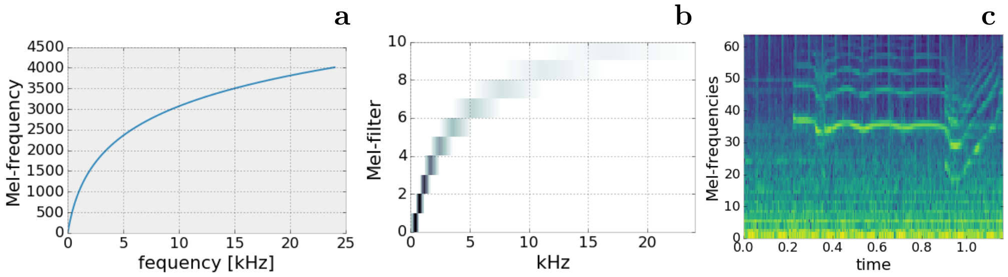


Figure B.3: **a** Graph of the mel scale mapping. **b** Mel filterbank with 10 triangular basis elements. **c** Mel-spectrogram in log scale from the soud in Fig B.1 with 68 mel-scale frequency bands.

B.1.4 Cepstrogram

The cepstrogram is similar to the MFCC but without using the mel-scale (Fig. B.1b). So, it is given by the cosine transform of the logarithm of the power spectral density

$$g_i(q) = \mathrm{DCT}\{\ln(P_i(k))\}. \tag{B.9}$$

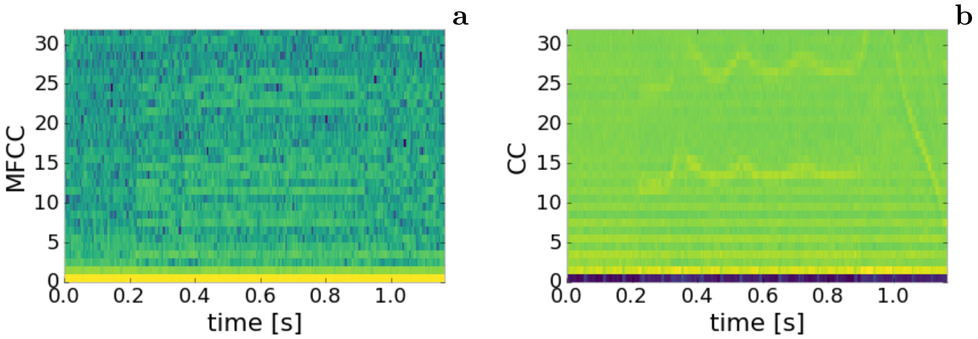


Figure B.4: (**a**) Mel-frequency cepstrogram (**b**) Cepstrogram of the whale sound in (Fig. B.1)

B.2 Temporal summarisation

The mel-spectrum, the MFCC and the cepstrogram are ways of compressing spectral features. Now we talk about how to compress features in the temporal dimension.

The aim is to reduce the number of temporal features. Let L be the size of the summarisation window. We can summarise the temporal information computing the mean and the standard deviation over L frames along the frequency bands of some spectral features. Where the frequency bands can be given as frequencies, mel-scale frequencies, mel-scaled cepstral coefficients or cepstral coefficients.

Therefore, the summarised features are given by

$$\mu_l(k) = \frac{1}{L} \sum_{i=l}^{l+L-1} s_i(k) \tag{B.10}$$

$$\sigma_l(k) = \sqrt{\frac{1}{L} \sum_{i=l}^{l+L-1} (s_i(k) - \mu_l(k))^2} \tag{B.11}$$

where $s_i(k)$ is the k-th frequency band of the i-frame and l is the index of the new time-summarised features.

Appendix C

Support vector machines

Support vector machines (SVM) are models that can be used to classify datasets separating them with hyperplanes called **decision boundaries**. Hyperplanes are such that maximise the **margin** between sets with different labels (Fig. C.2). Because SVM classifiers seek to separate datasets based on predefined labels, they belong to the kind of supervised machine learning models.

Given an unknown classification instance $\vec{u}$, the predicted label $y_{\vec{u}}$ will depend on which side of the hyperplane $\vec{u}$ is. For a dataset with binary labels $\{-1, +1\}$ this can be expressed mathematically as

$$y_{\vec{u}} = \begin{cases} +1 & \text{if} \quad \vec{w} \cdot \vec{u} + b > 0 \\ -1 & \text{if} \quad \vec{w} \cdot \vec{u} + b \leq 0 \end{cases} \tag{C.1}$$

where the decision boundary is defined points x in satisfying the plane equation $\vec{w} \cdot \vec{x} - b = 0$, with $\vec{w}$ a normal vector and b a scalar.

Training a SVM classifier consists of finding the hyperplane, $\vec{w} \cdot \vec{x} + b = 0$, that maximises the margin between two sets, subject to the correct classification (Fig. C.1b). For training samples $\vec{x}^{(i)}$ (indexed by i) with binary labels $y^{(i)} \in \{-1, +1\}$, the constraint of correctly classifying samples can be expressed mathematically as $y^{(i)} \, (\vec{w} \cdot \vec{x}^{(i)} + b) \geq 1 \; \forall \; i$; which states that samples with different labels should lay in different sides of the hyperplane. Therefore, the problem to solve is

$$\text{maximise margin subject to } y^{(i)} \, (\vec{w} \cdot \vec{x}^{(i)} + b) \geq 1. \tag{C.2}$$

Below we develop on the definition of the margin and the classification constraint.

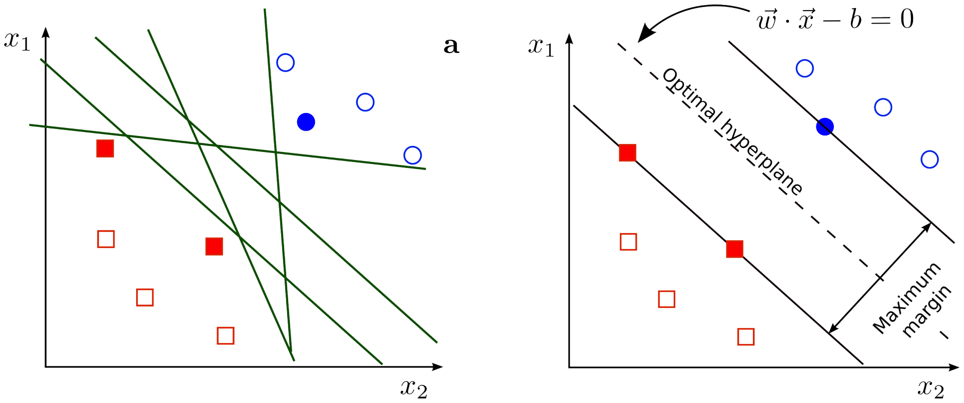


Figure C.1: Optimal hyperplane. Illustration of a binary dataset in a two dimensional feature space. The SVM classifier searches the hyperplane with the maximum margin subject to the least misclassification error. **a**, Various lines separating the two sets. **b**, Hyperplane with maximum margin, given by the distance between the planes with the support vectors (filled symbols) (Image modified from opencv.org CC BY-SA 3.0).

Defining the margin of SVM classifier involves two important concepts: the **support vectors** and the decision boundary (already mentioned). The support vectors are the points of each class closest to the decision boundary. The support vectors of each class define parallel planes that sandwich the decision boundary equidistantly (Fig. C.1b). The distance between the support vectors and the decision boundary is the margin whose width is $\frac{1}{||w||^2}$[1].

Separating datasets with hyperplanes may not always be possible, especially for real datasets. In such cases is important to relax the constraint to ensure the convergence of the optimisation problem under the adequate penalisation for misclassification. This is done introducing a "slack" variable ξ, that is zero for correctly classified samples.

Gathering the concepts above, the optimisation problem for training a SVM consists of finding the support vectors that

$$\text{minimise } \frac{1}{2}||\vec{w}||^2 + C\sum_i \xi^{(i)} \quad \text{subject to } y^{(i)}(\vec{w} \cdot \vec{x}^{(i)} + b) \geq 1 - \xi^{(i)}. \tag{C.3}$$

The first term corresponds to the margin maximisation, expressed as the minimisation of the inverse of the margin. The second term is a penalisation for misclassification tuned by the parameter C; with larger values of C penalising more than smaller values.

[1] this expression is derived around the web and in multiple machine learning books such as [207] using some algebra and geometry. Here we focus on the overall characteristics of the SVM classifiers.

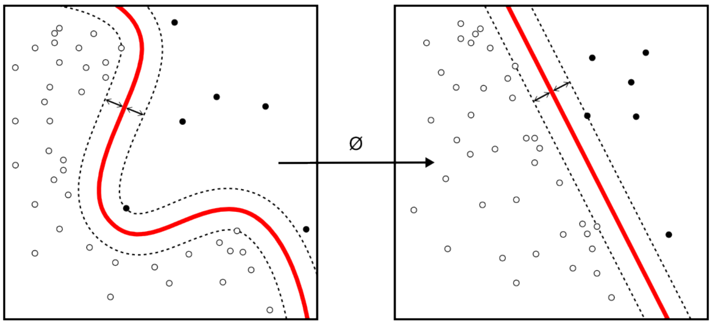


Figure C.2: Kernel trick. Comparison of a linear and a Gaussian kernel (Image by Alisneaky, svg version by Zirguezi Own work CC BY-SA 4.0, via Wikimedia Commons)

The last term is the constraint for correctly classifying labels as in expression C.2) but considering ξ.

SVMs as described above can only separate samples linearly by drawing hyperplanes. A nice aspect of SVMs is that they can easily be extended to non linear classification through the kernel trick [208]. The trick consists of mapping the samples to a space where they can be separated with hyperplanes. Because the optimisation problem C.3 only depends on inner products between the feature vectors $\vec{x^{(i)}} \cdot \vec{x^{(j)}}$, the linear functions can be replaced with non linear kernel functions $\phi(\vec{x^{(i)}}, \vec{x^{(j)}})$. A common choice the Gaussian radial basis function given by

$$\phi(\vec{x^{(i)}}, \vec{x^{(j)}}) = e^{-\gamma ||\vec{x}_i - \vec{x}_j||^2}. \tag{C.4}$$

The parameter γ sets the influence of the instances. Larger values of γ give more influence to the samples and smaller give less influence to the samples thus being more prone to overfitting.

Training a SVM classifier requires tuning the parameter C that penalises misclassification and γ when using radial basis function (Fig. C.3).

SVM classifiers were proposed in the 90's [208, 209] and quickly gained great popularity. Advantages of SVMs over other machine learning classifiers include that they do not get stuck in local minima given their convex cost function and that they are easy to kernelise allowing to implement non linear classification.

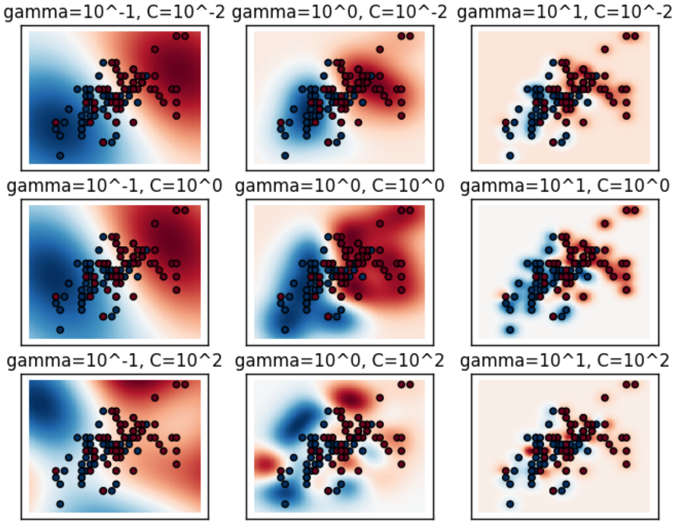


Figure C.3: Parameters of the SVM with radial basis function. Influence of parameters C and γ for a binary classification problem (Image from the Sklearn online documentation, BSD License).

Acknowledgements

Schrödinger's cat was the only animal that had showed in my career up to the moment I started this PhD. Having had the opportunity to study animal vocalisations was an extraordinarily beautiful research experience. This would not have been possible without the interdisciplinary team that supported me in this process. I owe this opportunity to Heike Vester, Christian Montes-Medina, Sarah Hallerberg, Kurt Hammerschmidt and Marc Timme. Thank you for letting me learn from you.

I thank my supervisor Marc Timme for his support throughout this PhD. He listened attentively and responded resourcefully to the many challenges that arose in these four years. I am also grateful for his insights on how to prepare talks and write abstracts and applications and for supporting me to participate in conferences; a very important part of my development these years.

I would like to thank Florentin Wörgötter for keeping the focus and his critical comments during the committee meetings. I am grateful for financially supporting me during the last months of my PhD. I would like to acknowledge as well the members of his group for the machine learning input during the computational neuroscience seminars.

I am very grateful with Kurt Hammerschmidt for his sincere support every time I needed it; during the committee meetings, commenting about sections of this thesis and abstracts. Your point of view was very valuable for the development of this project.

I thank my collaborators Heike Vester and Christian Montes Medina for sharing with me your field work days thereby cheering up my Bunsenstrasse work days.

During my first two years of PhD I had the fortune to share the office with Harold Gutch; whose computer powers with root access, the quietest keyboard typing and his witty sense of humour overload the concept an awesome officemate. I am grateful for the conversations about machine learning. I also thank Sarah Hallerberg and David Hofmann for their eagerness to discuss articles on the subject in the various machine learning seminars and other encounters. I thank David for carefully reading and commenting part I of this thesis.

I thank Bernhard Altaner for his critical comments about Chapter 5 which helped me improve its presentation. I thank Chiara Entradi and Jonathan Dawson for kindly listening the story of this thesis and for your questions which helped me to develop these ideas better. I thank Aura Montemayor, Dimitra Maoutsa, José Casadiego, Nahal Sharafi, Nora Molkenthin, Tania Garfias, Tomos Parry, Gunter Weber who helped me proof-reading this thesis.

This doctorate would not have been possible without the financial support of CONACYT, GGNB, and the Physics Institute of the University of Göttingen.

I am very grateful with my graduate school GGNB for supporting me during my PhD studies through courses, travel grants and for the opportunity of organising a conference.

I am enormously grateful with the staff that keeps the institutes running. I thank Barbara Kutz, Ayse Bolik, Viktoryia Novak, Frauke Bergmann, Ursula Wörgötter, Antje Erdmann for the help with the administrative work. I thank Yorck-Fabian Beensen, Denny Fliegner, Harold Gutch, Sven Jahnke, Debsankha Manik for maintaining the IT infrastructure. The personnel from the Otto-Hahn library that was incredibly helpful and accessible finding literature for this thesis.

I thank Andreas Sorge for having kindle the dialogue about reproducible research in the group. From those conversations I learned a great deal of good coding practices that boosted my coding effectiveness allowing me to focus on the research questions. This thesis is fully reproducible thanks to Andreas. I am also grateful with the people who make the Advanced Python Programming Summer School, G-node possible. There I learned many good coding practices as well as from the python course people: Norma, Sven, Fernando, Markus, Frank, Tūreiti, Josephine, Deb, Alexander, Sebastian, Tariq. Teaming up for the course was always fun.

This thesis benefited enormously from open source software. I would like to acknowledge the community who does this wonderful job for free. Particularly I would like to thank the contributors of the python projects on which this thesis built: pandas, sklearn, seaborn and librosa.

I would like to acknowledge the people of network dynamics group. These last four years are marked with beautiful memories of travels, world dinner flavours, sunny breaks at the terrace and numerous contributions to J. of App. Bakery. It was wonderful to work in a group with such a fresh and stimulating environment.

I thank my Leonard-Nelson family, my friends and the *colectivo de mexicanos en Gttingen* for all the fun and support.

Agradezco a mi familia a mis padres, a mis hermanos y a mis bellos sobrinos Gael y Pia; su amor es la raíz de toda inspiración en mi.

References

[1] Charles Darwin. *The Descent of Man and Seletion in Relation to Sex.* John Murray, 1871. 1, 102

[2] Jane Goodall. **Tool-using and aimed throwing in a community of free-living chimpanzees**. *Nature*, **201**(4926):1264, 1964. 1

[3] Piera Filippi. **Emotional and Interactional Prosody across Animal Communication Systems: A Comparative Approach to the Emergence of Language**. *Frontiers in Psychology*, **7**, 2016. 1, 72, 101, 102

[4] Niels J Dingemanse, Anahita JN Kazem, Denis Réale, and Jonathan Wright. **Behavioural reaction norms: animal personality meets individual plasticity**. *Trends in Ecology & Evolution*, **25**(2):81–89, 2010. 1

[5] Whitney E Greene, Kelly Melillo-Sweeting, and Kathleen M Dudzinski. **Comparing object play in captive and wild dolphins**. *International Journal of Comparative Psychology*, **24**(3), 2011. 1

[6] Luke Rendell and Hal Whitehead. **Culture in whales and dolphins**. *Behavioral and Brain Sciences*, **24**(2):309–324, 2001. 1

[7] Michael Krützen, Janet Mann, Michael R Heithaus, Richard C Connor, Lars Bejder, and William B Sherwin. **Cultural transmission of tool use in bottlenose dolphins**. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(25):8939–8943, 2005. 1

[8] Andrew Whiten, Jane Goodall, William C McGrew, Toshisada Nishida, Vernon Reynolds, Yukimaru Sugiyama, Caroline EG Tutin,

Richard W Wrangham, and Christophe Boesch. **Cultures in chimpanzees**. *Nature*, **399**(6737):682–685, 1999. 1

[9] Carel P Van Schaik, Marc Ancrenaz, Gwendolyn Borgen, Birute Galdikas, Cheryl D Knott, Ian Singleton, Akira Suzuki, Sri Suci Utami, and Michelle Merrill. **Orangutan cultures and the evolution of material culture**. *Science*, **299**(5603):102–105, 2003. 1

[10] Allison B Kaufman and James C Kaufman. *Animal creativity and innovation*. Academic Press, 2015. 1

[11] Stan A Kuczaj, Radhika Makecha, Marie Trone, Robin D Paulos, and Joana AA Ramos. **Role of peers in cultural innovation and cultural transmission: Evidence from the play of dolphin calves**. *International Journal of Comparative Psychology*, **19**(2), 2006. 1

[12] Simon M Reader and Kevin N Laland. *Animal innovation*, **10**. Oxford University Press Oxford, 2003. 1

[13] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. **The faculty of language: What is it, who has it, and how did it evolve?** *science*, **298**(5598):1569–1579, 2002. 1

[14] Noam Chomsky. *Language and problems of knowledge: The Managua lectures*, **16**. MIT press, 1988. 1

[15] Steven Pinker and Paul Bloom. **Natural language and natural selection**. *Behavioral and brain sciences*, **13**(04):707–727, 1990. 1

[16] Clive K Catchpole. **Sexual Selection and the Evolution of Complex Songs Among European Warblers of the Genus Acr Ocephal Us**. *Behaviour*, **74**(1):149–165, 1980. 1

[17] Kirsten M Bohn, Barbara Schmidt-French, Sean T Ma, and George D Pollak. **Syllable acoustics, temporal patterns, and call composition vary with behavioral context in Mexican free-tailed bats**. *The Journal of the Acoustical Society of America*, **124**(3):1838–1848, 2008. 1

[18] Anne Marijke Schel, Sandra Tranquilli, and Klaus Zuberbühler. **The alarm call system of two species of black-and-white colobus monkeys (Colobus polykomos and Colobus guereza).** *Journal of Comparative Psychology*, **123**(2):136, 2009. 1

[19] Jeremy GM Robertson. **Male territoriality, fighting and assessment of fighting ability in the Australian frog Uperoleia rugosa**. *Animal Behaviour*, **34**(3):763–772, 1986. 1

[20] Gloriana Chaverri, Erin H Gillam, and Thomas H Kunz. **A call-and-response system facilitates group cohesion among disc-winged bats**. *Behavioral Ecology*, **24**(2):481–487, 2012. 1

[21] Vincent M Janik and Peter JB Slater. **Context-specific use suggests that bottlenose dolphin signature whistles are cohesion calls**. *Animal behaviour*, **56**(4):829–838, 1998. 1, 56

[22] Daniel T Blumstein. **The evolution, function, and meaning of marmot alarm communication**. *Advances in the Study of Behavior*, **37**:371–401, 2007. 1

[23] Kate Arnold and Klaus Zuberbühler. **The alarm-calling system of adult male putty-nosed monkeys, Cercopithecus nictitans martini**. *Animal behaviour*, **72**(3):643–653, 2006. 1

[24] Charles D Hockett et al. *The origin of speech*. Freeman, 1960. 1

[25] Klaus Zuberbühler. **Predator-specific alarm calls in Campbell's monkeys, Cercopithecus campbelli**. *Behavioral Ecology and Sociobiology*, **50**(5):414–422, 2001. 1

[26] Karim Ouattara, Alban Lemasson, and Klaus Zuberbühler. **Campbell's monkeys use affixation to alter call meaning**. *PloS one*, **4**(11):e7808, 2009. 2, 95

[27] VincentM Janik and PeterJ B Slater. **Vocal learning in mammals**. *Advances in the Study of Behavior*, **26**:59–99, 1997. 2

[28] Monica Gagliano, Stefano Mancuso, and Daniel Robert. **Towards understanding plant bioacoustics**. *Trends in plant science*, **17**(6):323–325, 2012. 3

[29] Rolf Bardeli, D Wolff, Frank Kurth, M Koch, K-H Tauchert, and K-H Frommolt. **Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring**. *Pattern Recognition Letters*, **31**(12):1524–1534, 2010. 3

[30] Miguel A Acevedo and Luis J Villanueva-Rivera. **Using automated digital recording systems as effective tools for the monitoring of birds and amphibians**. *Wildlife Society Bulletin*, **34**(1):211–214, 2006. 3

[31] Andrew Digby, Michael Towsey, Ben D Bell, and Paul D Teal. **A practical comparison of manual and autonomous methods for acoustic monitoring**. *Methods in Ecology and Evolution*, **4**(7):675–683, 2013. 3

[32] Jérôme Sueur and Almo Farina. **Ecoacoustics: the ecological investigation and interpretation of environmental sound**. *Biosemiotics*, **8**(3):493–502, 2015. 3

[33] Paola Laiolo. **The emerging significance of bioacoustics in animal species conservation**. *Biological Conservation*, **143**(7):1635–1645, 2010. 3

[34] Klaus Riede. **Acoustic monitoring of Orthoptera and its potential for conservation**. *Journal of Insect Conservation*, **2**(3):217–223, 1998. 3

[35] G Pavan, C Fossati, M Manghi, and M Priano. **Passive acoustics tools for the implementation of Acoustic Risk Mitigation Policies**. *European Cetacean Society Newsletter No*, pages 52–58, 2004. 3

[36] Max Deffenbaugh. **Mitigating seismic impact on marine life: Current practice and future technology**. *Bioacoustics*, **12**(2-3):316–318, 2002. 3

[37] Laurance R Doyle, Brenda McCowan, Simon Johnston, and Sean F Hanser. **Information theory, animal communication, and the search for extraterrestrial intelligence**. *Acta Astronautica*, **68**(3):406–417, 2011. 3, 71

[38] Maurice Ewing and J Lamar Worzel. **Long-range sound transmission**. *Geological Society of America Memoirs*, **27**:1–32, 1948. 3, 106

[39] Roger Payne and Douglas Webb. **Orientation by means of long range acoustic signaling in baleen whales**. *Annals of the New York Academy of Sciences*, **188**(1):110–141, 1971. 3, 106

[40] William A Yost and Richard R Fay. *Auditory perception of sound sources*, **29**. Springer Science & Business Media, 2007. 3

[41] David K Mellinger and Christopher W Clark. **MobySound: A reference archive for studying automatic recognition of marine mammal sounds**. *Applied Acoustics*, **67**(11):1226–1242, 2006. 9

[42] Stan Boutin, Diane L Haughland, Jim Schieck, Jim Herbers, and Erin Bayne. **A new approach to forest biodiversity monitoring in Canada**. *Forest Ecology and Management*, **258**:S168–S175, 2009. 9

[43] **Macaulay Library at the Cornell Lab of Ornithology**. 9

[44] Mark P Johnson and Peter L Tyack. **A digital acoustic recording tag for measuring the response of wild marine mammals to sound**. *IEEE journal of oceanic engineering*, **28**(1):3–12, 2003. 9

[45] Dan Stowell and Mark D. Plumbley. **Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning**. *PeerJ*, **2**:e488, 7 2014. 10

[46] Judith C. Brown, Paris Smaragdis, and Anna-Nousek McGregor. **Automatic identification of individual killer whales**. pages EL 93 – EL 98, 2010. 10

[47] Patrick J. Clemins, Michael T. Johnson, Kirsten M. Leong, and Anne Savage. **Automatic classification and speaker identification of African elephant (Loxodonta africana) vocalizations**. pages 1–8, 2004. 10

[48] Judith C. Brown and Patrick J. O. Miller. **Automatic classification of killer whale vocalizations using dynamic time warping**. pages 1201–1207, 2007. 10

[49] Federica Pace, Frederic Benard, Herve Glotin, Olivier Adam, and Paul White. **Subunit definition and analysis for humpback whale call classification**. *Elservier*, **71**:1107–1112, 2010. 10

[50] Lee Ngee Tan, Kantapon Kaewtip, Martin L Cody, Charles E Taylor, and Abeer Alwan. **Evaluation of a Sparse Representation-Based Classifier For Bird Phrase Classification Under Limited Data Conditions.** In *Interspeech*, pages 2522–2525, 2012. 10

[51] Volker B. Deecke and Vincent M. Janik. **Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls**. pages 645–653, 2005. 10

[52] Judith C. Brown and Paris Smaragdis. **Hidden Markov and Gaussian mixture models for automatic call classification**. *Brown and Smaragdis: JASA Express Letters*, pages EL221–224, 2009. 10

[53] Peter Rickwood and Taylor. Andrew. **Methods for automatically analyzing humpback song units**. pages 1763–1772, 2007. 10

[54] G. Picot, O. Adam, M. Bergounioux, H. Glotin, and Mayer F-X. **Automatic prosodic clustering of humpback whales song**. *University Orlans, France*, 2008. 10

[55] Sieve Analytics. **Arbimonautomated remote biodiversity monitoring network**, 2015. 10

[56] Steven Ness. *The Orchive: A system for semi-automatic annotation and analysis of a large collection of bioacoustic recordings*. PhD thesis, 2013. 10, 37

[57] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. **Scikit-learn: Machine learning in Python**. *The Journal of Machine Learning Research*, **12**:2825–2830, 2011. 13, 16, 29, 98

[58] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. **librosa: Audio and music signal analysis in python**. In *Proceedings of the 14th Python in Science Conference*, 2015. 16, 98

[59] Audacity Team. **Audacity(R): Free Audio Editor and Recorder**, 1999-2014. 21, 76

[60] Peter Tyack. **Population biology, social behavior and communication in whales and dolphins**. *Trends in ecology & evolution*, **1**(6):144–150, 1986. 27, 55, 56

[61] Heike Vester, Sarah Hallerberg, Marc Timme, and Kurt Hammerschmidt. **Vocal repertoire of long-finned pilot whales (Globicephala melas) in northern Norway**. *The Journal of the Acoustical Society of America*, 2017. accepted. 31, 55, 56, 58

[62] Steven Ness, Helena Symonds, Paul Spong, and George Tzanetakis. **The Orchive: Data mining a massive bioacoustic archive**. *arXiv preprint arXiv:1307.0589*, 2013. 37

[63] John KB Ford et al. *A catalogue of underwater calls produced by killer whales (Orcinus orca) in British Columbia.* Department of Fisheries and Oceans, Fisheries Research Branch, Pacific Biological Station, 1987. 37

[64] Roger Payne and Roger Payne. *Communication and behavior of whales.* Westview Press Boulder, 1983. 44

[65] Roger Payne. **Songs of the Humpback Whale**. CRM Records, August 1970. 43

[66] NASA. **The Voyager Golden Record**. Great Images in NASA Description. 43

[67] William N Tavolga. **Listening Backward: Early Days of Marine Bioacoustics**. In *The Effects of Noise on Aquatic Life*, pages 11–14. Springer, 2012. 43

[68] OW Schreiber. **Some sounds from marine life in the Hawaiian area**. *The Journal of the Acoustical Society of America*, **24**(1):116–116, 1952. 43

[69] Roger S Payne and Scott McVay. **Songs of humpback whales**. *Science*, **173**(3997):585–597, 1971. 43, 44

[70] Arik Kershenbaum, Daniel T Blumstein, Marie A Roch, Çağlar Akçay, Gregory Backus, Mark A Bee, Kirsten Bohn, Yan Cao, Gerald Carter, Cristiane Cäsar, et al. **Acoustic sequences in non-human animals: a tutorial review and prospectus**. *Biological Reviews*, **91**(1):13–52, 2014. 44, 45, 47, 75, 99, 100

[71] Todd M Freeberg, Robin IM Dunbar, and Terry J Ord. **Social complexity as a proximate and ultimate factor in communicative complexity**, 2012. 44

[72] Todd M Freeberg. **Social complexity can drive vocal complexity group size influences vocal information in Carolina chickadees**. *Psychological Science*, **17**(7):557–561, 2006. 44

[73] Daniel T Blumstein and Kenneth B Armitage. **Does sociality drive the evolution of communicative complexity? A comparative test with ground-dwelling sciurid alarm calls**. *The American Naturalist*, **150**(2):179–200, 1997. 44

[74] Kimberly A Pollard and Daniel T Blumstein. **Evolving communicative complexity: insights from rodents and beyond**. *Phil. Trans. R. Soc. B*, **367**(1597):1869–1878, 2012. 44

[75] Gerald S Wilkinson. **Social and vocal complexity in bats.** 2003. 44

[76] Karen McComb and Stuart Semple. **Coevolution of vocal communication and sociality in primates**. *Biology Letters*, **1**(4):381–385, 2005. 44

[77] RIM Dunbar. **Group size, vocal grooming and the origins of language**. *Psychonomic Bulletin & Review*, pages 1–4, 2016. 44

[78] John G Robinson. **An analysis of the organization of vocal communication in the titi monkey Callicebus moloch**. *Ethology*, **49**(4):381–405, 1979. 45

[79] BRENDA McCOWAN, Sean F Hanser, and Laurance R Doyle. **Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires**. *Animal behaviour*, **57**(2):409–419, 1999. 45, 71

[80] Laurance R Doyle, Brenda McCowan, Sean F Hanser, Christopher Chyba, Taylor Bucci, and J Ellen Blue. **Applicability of information theory to the quantification of responses to anthropogenic noise by southeast Alaskan humpback whales**. *Entropy*, **10**(2):33–46, 2008. 45

[81] Todd M Freeberg and Jeffrey R Lucas. **Information theoretical approaches to chick-a-dee calls of Carolina chickadees (Poecile carolinensis).** *Journal of Comparative Psychology*, **126**(1):68, 2012. 45

[82] Ryuji Suzuki, John R Buck, and Peter L Tyack. **Information entropy of humpback whale songs a**. *The Journal of the Acoustical Society of America*, **119**(3):1849–1866, 2006. 45

[83] Arik Kershenbaum, Ann E Bowles, Todd M Freeberg, Dezhe Z Jin, Adriano R Lameira, and Kirsten Bohn. **Animal vocal sequences: not the Markov chains we thought they were**. *Proceedings of the Royal Society of London B: Biological Sciences*, **281**(1792):20141370, 2014. 45

[84] Michel Andre and Cees Kamminga. **Rhythmic dimension in the echolocation click trains of sperm whales: A possible function of identification and communication**. *Journal of the Marine Biological Association of the UK*, **80**(01):163–169, 2000. 45

[85] H Carl Gerhardt and Franz Huber. *Acoustic communication in insects and anurans: common problems and diverse solutions.* University of Chicago Press, 2002. 45

[86] Marc David Hauser, Bryan Agnetta, and Christine Perez. **Orienting asymmetries in rhesus monkeys: the effect of time-domain changes on acoustic perception**. *Animal Behaviour*, **56**(1):41–47, 1998. 45

[87] Eric Brown and Murray S Miron. **Lexical and syntactic predictors of the distribution of pause time in reading**. *Journal of Verbal Learning and Verbal Behavior*, **10**(6):658–667, 1971. 45

[88] François Grosjean and Alain Deschamps. **Analyse contrastive des variables temporelles de langlais et du français: vitesse de parole et variables composantes, phénomènes dhésitation**. *Phonetica*, **31**(3-4):144–184, 1975. 45

[89] Francois Grosjean, Lysiane Grosjean, and Harlan Lane. **The patterns of silence: Performance structures in sentence production**. *Cognitive psychology*, **11**(1):58–81, 1979. 45

[90] Klaus R Scherer and James S Oshinsky. **Cue utilization in emotion attribution from auditory stimuli**. *Motivation and emotion*, **1**(4):331–346, 1977. 45

[91] William F Johnson, Robert N Emde, Klaus R Scherer, and Mary D Klinnert. **Recognition of emotion from vocal cues**. *Archives of General Psychiatry*, **43**(3):280–283, 1986. 45

[92] Heiner Ellgring and Klaus R Scherer. **Vocal indicators of mood change in depression**. *Journal of Nonverbal Behavior*, **20**(2):83–110, 1996. 45

[93] Rainer Banse and Klaus R Scherer. **Acoustic profiles in vocal emotion expression.** *Journal of personality and social psychology*, **70**(3):614, 1996. 45

[94] Christopher S Evans and Peter Marler. **Food calling and audience effects in male chickens, Gallus gallus: their relationships to food availability, courtship and social facilitation**. *Animal Behaviour*, **47**(5):1159–1170, 1994. 46, 50, 99

[95] Adolfo Christian Montes-Medina, Alejandro Salinas-Melgoza, and Katherine Renton. **Contextual flexibility in the vocal repertoire of an Amazon parrot**. *Frontiers in zoology*, **13**(1):40, 2016. 46, 50, 75, 93, 99

[96] Clive K Catchpole. **Temporal and sequential organisation of song in the sedge warbler (Acrocephalus schoenobaenus)**. *Behaviour*, **59**(3):226–245, 1976. 46, 50, 93, 99

[97] III Mercado, Louis M Herman, and Adam A Pack. **Stereotypical sound patterns in humpback whale songs: usage and function**. *Aquatic Mammals*, **29**(1):37–52, 2003. 46, 50, 93, 99

[98] Stephen Handel, Sean K Todd, and Ann M Zoidis. **Rhythmic structure in humpback whale (Megaptera novaeangliae) songs: Preliminary implications for song production and perception**. *The Journal of the Acoustical Society of America*, **125**(6):EL225–EL230, 2009. 46, 50, 93, 99

[99] C. Cannam, C. Landone, and M. Sandler. **Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files**. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy, October 2010. 46, 57, 78

[100] Carel ten Cate and Kazuo Okanoya. **Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning**. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **367**(1598):1984–1994, 2012. 47, 99

[101] Toshitaka N Suzuki, David Wheatcroft, and Michael Griesser. **Experimental evidence for compositional syntax in bird calls**. *Nature communications*, **7**, 2016. 47, 95

[102] Jack W Bradbury and Sandra L Vehrencamp. **Principles of animal communication**. 1998. 48

[103] Robert C Berwick, Kazuo Okanoya, Gabriel JL Beckers, and Johan J Bolhuis. **Songs to syntax: the linguistics of birdsong**. *Trends in cognitive sciences*, **15**(3):113–121, 2011. 48

[104] Daniel Jurafsky and H James. **Speech and language processing an introduction to natural language processing, computational linguistics, and speech**. 2000. 51

[105] David S Moore and George P McCabe. *Introduction to the Practice of Statistics.* WH Freeman/Times Books/Henry Holt & Co, 1989. 53

[106] D Bloch, G Desportes, R Mouritsen, S Skaaning, and E Stefansson. **An introduction to studies of the ecology and status of the long-finned pilot whale (Globicephala melas) off the Faroe Islands, 1986-1988**. *Report of the International Whaling Commission (Special Issue 14)*, pages 1–32, 1993. 55

[107] Joana F Augusto, Timothy R Frasier, and Hal Whitehead. **Characterizing alloparental care in the pilot whale (Globicephala melas) population that summers off Cape Breton, Nova Scotia, Canada**. *Marine Mammal Science*, 2016. 55

[108] Lisa S Baraff and Regina A Asmutis-Silvia. **LONG-TERM ASSOCIATION OF AN INDIVIDUAL LONG-FINNED PILOT WHALE AND ATLANTIC WHITE-SIDED DOLPHINS**. *Marine Mammal Science*, **14**(1):155–161, 1998. 55

[109] DO Cordes. **The causes of whale strandings**. *New Zealand Veterinary Journal*, **30**(3):21–24, 1982. 55

[110] Robert Nawojchik, David J St Aubin, and Anne Johnson. **Movements and dive behavior of two stranded, rehabilitated long-finned pilot whales (Globicephala melas) in the northwest Atlantic**. *Marine Mammal Science*, **19**(1):232–239, 2003. 55

[111] **The IUCN Red List of Threatened Species**. Version 2016-3. 55, 76

[112] Peter L Tyack. *Functional aspects of cetacean communication.* Cetacean societies: field studies of dolphins and whales. The University of Chicago Press Chicago, IL:, 2000. 55, 56, 73

[113] Linda S Weilgart and Hal Whitehead. **Vocalizations of the North Atlantic pilot whale (Globicephala melas) as related to behavioral contexts**. *Behavioral Ecology and Sociobiology*, **26**(6):399–402, 1990. 55, 56

[114] RenÉ-Guy Busnel and Albin Dziedzic. **Acoustic signals of the pilot whale Globicephala melaena and of the porpoises Delphinus delphis and Phocoena phocoena**. *Whales, dolphins and porpoises*, pages 607–646, 1966. 56

[115] Algis G Taruski. **The whistle repertoire of the North Atlantic pilot whale (Globicephala melaena) and its relationship to behavior and environment**. In *Behavior of marine animals*, pages 345–368. Springer, 1979. 56

[116] Frants H Jensen, Jacobo Marrero Perez, Mark Johnson, Natacha Aguilar Soto, and Peter T Madsen. **Calling under pressure: short-finned pilot whales make social calls during deep foraging dives**. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20102604, 2011. 56

[117] Leah Nemiroff and Hal Whitehead. **Structural characteristics of pulsed calls of long-finned pilot whales Globicephala melas**. *Bioacoustics*, **19**(1-2):67–92, 2009. 56

[118] Heike Vester, Kurt Hammerschmidt, Marc Timme, and Sarah Hallerberg. **Quantifying group specificity of animal vocalizations without specific sender information**. *Physical Review E*, **93**(2):022138, 2016. 56

[119] G Desportes and R Mouritsen. **Preliminary results on the diet of long-finned pilot whales off the Faroe Islands**. *Rep Int Whal Commn. Special Issue*, (14):305–324, 1993. 56

[120] Diethard Tautz Bill Amos, Christian Schlöktterer. **Social Structure of Pilot Whales Revealed by Analytical DNA Profiling**. *Science*, **260**:670, 1993. 56

[121] Hal Whitehead, Felicia Vachon, and Timothy R Frasier. **Cultural Hitchhiking in the Matrilineal Whales**. *Behavior Genetics*, pages 1–11, 2017. 56

[122] Bill Amos, Dorete Bloch, Genevieve Desportes, Tamsin MO Majerus, David R Bancroft, John A Barrett, and Gabriel A Dover. **A review of the molecular evidence relating to social organisation and breeding system in the long-finned pilot whale**. **special issue 14**:209–217, 1993. 56

[123] C Andrea Ottensmeyer and Hal Whitehead. **Behavioural evidence for social units in long-finned pilot whales**. *Canadian Journal of Zoology*, **81**(8):1327–1338, 2003. 56

[124] Melba C Caldwell. **Individualized whistle contours in bottle-nosed dolphins (Tursiops truncatus)**. *Nature*, **207**:434–435, 1965. 56, 69

[125] John KB Ford and H Dean Fisher. **Group-specific dialects of killer whales (Orcinus orca) in British Columbia**. In *Communication and behavior of whales*, **76**, pages 129–161. Westview Press Boulder, CO, 1983. 56

[126] Shannon Rankin, Julie Oswald, Jay Barlow, and Marc Lammers. **Patterned burst-pulse vocalizations of the northern right whale dolphin, Lissodelphis borealis**. *The Journal of the Acoustical Society of America*, **121**(2):1213–1218, 2007. 69

[127] Luciana Duarte de Figueiredo and Sheila Marino Simão. **Possible occurrence of signature whistles in a population of Sotalia guianensis (Cetacea, Delphinidae) living in Sepetiba Bay, Brazil**. *The Journal of the Acoustical Society of America*, **126**(3):1563–1569, 2009. 69

[128] Maxwell B Kaplan, T Aran Mooney, Laela S Sayigh, and Robin W Baird. **Repeated call types in Hawaiian melon-headed whales (Peponocephala electra)**. *The Journal of the Acoustical Society of America*, **136**(3):1394–1401, 2014. 69, 93

[129] Laela Sayigh, Nicola Quick, Gordon Hastie, and Peter Tyack. **Repeated call types in short-finned pilot whales, Globicephala**

macrorhynchus. *Marine Mammal Science*, **29**(2):312–324, 2013. 69, 71, 72, 93

[130] Elizabeth MJ Zwamborn and Hal Whitehead. **Repeated call sequences and behavioural context in long-finned pilot whales off Cape Breton, Nova Scotia, Canada**. *Bioacoustics*, pages 1–15, 2016. 69, 72, 93

[131] Vincent M Janik, Stephanie L King, Laela S Sayigh, and Randall S Wells. **Identifying signature whistles from recordings of groups of unrestrained bottlenose dolphins (Tursiops truncatus)**. *Marine Mammal Science*, **29**(1):109–122, 2013. 69

[132] H. Carter Esch, Laela S. Sayigh, and Randall S. Wells. **Quantifying parameters of bottlenose dolphin signature whistles**. *Marine Mammal Science*, **25**(4):976–986, 2009. 69

[133] George Kingsley Zipf. **The Psycho-Biology of Language. An Introduction to Dynamic Philology. 1935**, 1935. 71

[134] Ramon Ferrer i Cancho and Ricard V Solé. **Least effort and the origins of scaling in human language**. *Proceedings of the National Academy of Sciences*, **100**(3):788–791, 2003. 71, 78

[135] Anatol Rapoport. **Zipfs law re-visited**. *Studies on Zipfs Law*, pages 1–28, 1982. 71

[136] Ryuji Suzuki, John R Buck, and Peter L Tyack. **The use of Zipf's law in animal communication analysis**. *Animal Behaviour*, **69**(1):F9–F17, 2005. 71

[137] B. McCowan, L. Doyle, M. Jenkins, and S. F. Hanser. **The appropriate use of Zipfs law in animal communication studies**. *Elservier*, pages F1–F7, 2003. 71

[138] Gejza Wimmer and Gabriel Altmann. **Towards a unified derivation of some linguistic laws**. In *Contributions to the Science of Text and Language*, pages 329–337. Springer, 2007. 71

[139] Ramon Ferrer-i Cancho and Brenda McCowan. **A law of word meaning in dolphin whistle types**. *Entropy*, **11**(4):688–701, 2009. 71

[140] John K. B. Ford. **Acoustic behaviour of resident killer whales (Orcinus orca) off Vancouver Island, British Columbia.** *Canadian Journal of Zoology*, **67**:727–745, 1989. 71, 73

[141] Ruediger Riesch, John K. Ford, and Thomsen Frank. **Whistle sequences in wild killer whales (Orcinus orca)**. pages 1822–1829, 2008. 71

[142] Jennifer L. Miksis-Olds, John R. Buck, Michael J. Noad, Douglas H. Cato, and M. Dale Stokes. **Information theory analysis of Australian humpback whale song**. pages 2385–2393, 2008. 71

[143] Ramon Ferrer-i Cancho and Brenda McCowan. **The span of correlations in dolphin whistle sequences**. *Journal of Statistical Mechanics: Theory and Experiment*, **2012**(06):P06002, 2012. 71

[144] Jack P Hailman, Millicent S Ficken, and Robert W Ficken. **The chick-a-deecalls of Parus atricapillus: a recombinant system of animal communication compared with written English**. *Semiotica*, **56**(3-4):191–224, 1985. 71, 93

[145] Clive K Catchpole and Peter JB Slater. *Bird song: biological themes and variations*. Cambridge university press, 2003. 71, 75

[146] Agnes Candiotti, Klaus Zuberbühler, and Alban Lemasson. **Context-related call combinations in female Diana monkeys**. *Animal cognition*, **15**(3):327–339, 2012. 71

[147] Laela S Sayigh, H Carter Esch, Randall S Wells, and Vincent M Janik. **Facts about signature whistles of bottlenose dolphins, Tursiops truncatus**. *Animal Behaviour*, **74**(6):1631–1642, 2007. 72

[148] Elizabeth Zwamborn. *Repeated Call Sequences in Long-finned Pilot Whales: Social Setting, Modification, and Behavioural Context*. PhD thesis, 2016. 72

[149] Patrick J. O. Miller. **Mixed-directionality of killer whale stereotyped calls: a direction of movement cue?** *Behav Ecol Sociobiol*, **52**:262–270, 2002. 72

[150] Katie Collier, Balthasar Bickel, Carel P van Schaik, Marta B Manser, and Simon W Townsend. **Language evolution: syntax before phonology?** In *Proc. R. Soc. B*, **281**, page 20140263. The Royal Society, 2014. 72, 95, 102

[151] Nicola J Quick and Vincent M Janik. **Bottlenose dolphins exchange signature whistles when meeting at sea**. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20112537, 2012. 73

[152] Nigel I Mann, Kimberly A Dingess, Keith F Barker, Jeff A Graves, and Peter JB Slater. **A comparative study of song form and duetting in neotropical Thryothorus wrens**. *Behaviour*, **146**(1):1–43, 2009. 73, 101

[153] Jack P Hailman and Millicent S Ficken. **Combinatorial animal communication with computable syntax: chick-a-dee calling qualifies as languageby structural linguistics**. *Animal Behaviour*, **34**(6):1899–1901, 1986. 75

[154] Kathryn M Rusch, Carolyn L Pytte, and Millicent S Ficken. **Organization of agonistic vocalizations in Black-chinned Hummingbirds**. *Condor*, pages 557–566, 1996. 75

[155] Daniel W Leger. **First documentation of combinatorial song syntax in a suboscine passerine species**. *The Condor*, **107**(4):765–774, 2005. 75

[156] D Rendall and CD Kaluthota. **Song organization and variability in northern house wrens (Troglodytes aedon parkmanii) in western Canada**. *The Auk*, **130**(4):617–628, 2013. 75

[157] Christine R Dahlin and Timothy F Wright. **Does syntax contribute to the function of duets in a parrot, Amazona auropalliata?** *Animal cognition*, **15**(4):647–656, 2012. 75

[158] Christine R Dahlin and Timothy F Wright. **Duets in Yellow-Naped Amazons: Variation in Syntax, Note Composition and Phonology at Different Levels of Social Organization**. *Ethology*, **115**(9):857–871, 2009. 75

[159] Timothy F Wright and Christine R Dahlin. **Pair duets in the yellow-naped amazon (Amazona auropalliata): phonology and syntax**. *Behaviour*, **144**(2):207–228, 2007. 75, 101

[160] Joseph Michael Forshaw and William T Cooper. *Parrots of the world.* Blandford London, 1989. 76

[161] Katherine Renton and A Salinas-Melgoza. **Amazona finschi (Sclater 1864) Loro corona lila**. *Noguera, FA, JH Vega-Rivera & AN García-Aldrete, M. Quesada (Eds.). Historia Natural de Chamela. Instituto de Biología, Universidad Nacional Autónoma de México. Distrito Federal, México*, pages 341–342, 2002. 76

[162] Katherine Renton. *Reproductive ecology and conservation of the Lilac-crowned Parrot (Amazona finschi) in Jalisco, Mexico.* PhD thesis, University of Kent at Canterbury, 1998. 76

[163] Katherine Renton. **Lilac-crowned Parrot diet and food resource availability: resource tracking by a parrot seed predator**. *The Condor*, **103**(1):62–69, 2001. 76

[164] Katherine Renton. **Influence of environmental variability on the growth of Lilac-crowned Parrot nestlings**. *Ibis*, **144**(2):331–339, 2002. 76

[165] Katherine Renton and Alejandro Salinas-Melgoza. **Nesting behavior of the Lilac-crowned Parrot**. *The Wilson Bulletin*, pages 488–493, 1999. 76

[166] Ofer Tchernichovski, Fernando Nottebohm, Ching Elizabeth Ho, Bijan Pesaran, and Partha Pratim Mitra. **A procedure for an automated measurement of song similarity**. *Animal behaviour*, **59**(6):1167–1176, 2000. 76

[167] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books, 2016. 78

[168] Indrikis Krams, Tatjana Krama, Todd M Freeberg, Cecilia Kullberg, and Jeffrey R Lucas. **Linking social complexity and vocal complexity: a parid perspective**. *Phil. Trans. R. Soc. B*, **367**(1597):1879–1891, 2012. 93

[169] Hilary B. Moors and John M. Terhune. **Repetition patterns in Weddell seal (Leptonychotes weddellii) underwater multiple element calls**. pages 1261–1270, 2004. 93

[170] Todd M Freeberg. **Complexity in the chick-a-dee call of Carolina chickadees (Poecile carolinensis): associations of context and signaler behavior to call structure**. *The Auk*, **125**(4):896–907, 2008. 93

[171] James R Hurford. *The origins of meaning: Language in the light of evolution*, **1**. Oxford University Press, 2007. 95

[172] Dan Stowell, Lisa Gill, and David Clayton. **Detailed temporal structure of communication networks in groups of songbirds**. *Journal of the Royal Society Interface*, **13**(119):20160296, 2016. 99

[173] Daniel S Reich, Ferenc Mechler, Keith P Purpura, and Jonathan D Victor. **Interspike intervals, receptive fields, and information encoding in primary visual cortex**. *Journal of Neuroscience*, **20**(5):1964–1974, 2000. 99

[174] Tyler M Schulz, Hal Whitehead, Shane Gero, and Luke Rendell. **Overlapping and matching of codas in vocal interactions between sperm whales: insights into communication function**. *Animal Behaviour*, **76**(6):1977–1988, 2008. 101

[175] Vincent M. Janik. **Whistle Matching in Wild Bottlenose Dolphins (Tursiops truncatus)**. *Science*, **289**:1355–1357, 2000. 101

[176] Patrick J. O. Miller, A. D. Shapiro, P. L. Tyack, and A. R. Solow. **Call-type matching in vocal exchanges of free-ranging resident killer whales, Orcinus orca**. *Elservier*, **67**:199–1107, 2004. 101

[177] Dietmar Todt and Marc Naguib. **Vocal interactions in birds: the use of song as a model in communication**. *Advances in the Study of Behavior*, **29**:247–296, 2000. 101

[178] Seán G Roberts, Francisco Torreira, and Stephen C Levinson. **The effects of processing and sequence organization on the timing of turn taking: a corpus study**. *Frontiers in psychology*, **6**:509, 2015. 101

[179] Nigel Ward and Wataru Tsukahara. **Prosodic features which cue back-channel responses in English and Japanese**. *Journal of pragmatics*, **32**(8):1177–1207, 2000. 101

[180] Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. **On temporal aspects of turn taking in conversational dialogues**. *Speech Communication*, **47**(1):80–86, 2005. 101

[181] Mark Steedman. **Phrasal intonation and the acquisition of syntax**. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, pages 331–342, 1996. 101

[182] Elizabeth K Johnson. **Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech**. *The Journal of the Acoustical Society of America*, **123**(6):EL144–EL148, 2008. 101

[183] Claudia Männel, Christine S Schipke, and Angela D Friederici. **The role of pause as a prosodic boundary marker: Language ERP studies in German 3-and 6-year-olds**. *Developmental Cognitive Neuroscience*, **5**:86–94, 2013. 101

[184] Eugene S Morton. **On the occurrence and significance of motivation-structural rules in some bird and mammal sounds**. *The American Naturalist*, **111**(981):855–869, 1977. 102

[185] Tobias Riede, Adam Clark Arcadi, and Michael J Owren. **Nonlinear acoustics in the pant hoots of common chimpanzees (Pan troglodytes): Vocalizing at the edge**. *The Journal of the Acoustical Society of America*, **121**(3):1758–1767, 2007. 102

[186] S. Gero, L. Bejder, Whitehead H., J. Mann, and R. C. Connor. **Behaviourally specific preferred associations in bottlenose dolphins, Tursiops spp.**

[187] Angela S Stoeger, Anton Baotic, Desheng Li, and Benjamin D Charlton. **Acoustic features indicate arousal in infant giant panda vocalisations**. *Ethology*, **118**(9):896–905, 2012. 102

[188] Joseph Soltis, Kirsten Leong, and Anne Savage. **African elephant vocal communication I: antiphonal calling behaviour among affiliated females**. *Animal Behaviour*, **70**(3):579–587, 2005. 102

[189] Angela S Stoeger, Benjamin D Charlton, Helmut Kratochvil, and W Tecumseh Fitch. **Vocal cues indicate level of arousal in infant African elephant roars**. *The Journal of the Acoustical Society of America*, **130**(3):1700–1710, 2011. 102

[190] Michelle L Hall. **A review of vocal duetting in birds**. *Advances in the Study of Behavior*, **40**:67–121, 2009. 102

[191] Hiroki Koda, Alban Lemasson, Chisako Oyakawa, Joko Pamungkas, Nobuo Masataka, et al. **Possible role of mother-daughter vocal interactions on the development of species-specific song in gibbons**. *PLoS One*, **8**(8):e71432, 2013. 102

[192] Claudia Fichtel, Kurt Hammerschmidt, and Uwe Jürgens. **On the vocal expression of emotion. A multi-parametric analysis of different states of aversion in the squirrel monkey**. *Behaviour*, **138**(1):97–116, 2001. 102

[193] Kurt Hammerschmidt, Viveka Ansorge, Julia Fischer, and Dietmar Todt. **Dusk calling in Barbary macaques (Macaca sylvanus): demand for social shelter**. *American Journal of Primatology*, **32**(4):277–289, 1994. 102

[194] J. J. Rousseau. **Essay on the Origin of Languages and Writings Related to Music**. *Hanover: University Press of New England*, 1781. 102

[195] Jespersen Otto. **Language, its Nature, Development and Origin**, 1922. 102

[196] Frank B Livingstone. **Did the Australopithecines sing?** *Current Anthropology*, **14**(1/2):25–29, 1973. 102

[197] Bruce Richman. **On the evolution of speech: Singing as the middle term**, 1993. 102

[198] Steven Brown. *The musilanguage model of music evolution.* na, 2000. 102

[199] Björn Merker. *Synchronous chorusing and human origins.* na, 2000. 102

[200] Anthony J DeCasper and Melanie J Spence. **Prenatal maternal speech influences newborns' perception of speech sounds**. *Infant behavior and Development*, **9**(2):133–150, 1986. 103

[201] Robert Sekuler, Allison B Sekuler, and Renee Lau. **Sound alters visual motion perception.** *Nature*, **385**(6614):308, 1997. 103

[202] Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. **Visual illusion induced by sound**. *Cognitive Brain Research*, **14**(1):147–152, 2002. 103

[203] Harvey Fletcher. **Auditory patterns**. *Reviews of modern physics*, **12**(1):47, 1940. 105

[204] James W Cooley and John W Tukey. **An algorithm for the machine calculation of complex Fourier series**. *Mathematics of computation*, **19**(90):297–301, 1965. 107

[205] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. **A scale for the measurement of the psychological magnitude pitch**. *The Journal of the Acoustical Society of America*, **8**(3):185–190, 1937. 110

[206] Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler. **Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music**. In *Seventh International Conference on Music Information Retrieval (ISMIR)*, 2006. 110

[207] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, **1**. Springer series in statistics Springer, Berlin, 2001. 114

[208] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. **A training algorithm for optimal margin classifiers**. In *Proceedings of the*

fifth annual workshop on Computational learning theory, pages 144–152. ACM, 1992. 115

[209] Corinna Cortes and Vladimir Vapnik. **Support-vector networks**. *Machine learning*, **20**(3):273–297, 1995. 115