UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Automated classification of humpback whale (*Megaptera novaeangliae*) songs using Hidden Markov Models

by

Federica Pace

Thesis for the degree of Doctor of Philosophy

May 2013

ABSTRACT

Humpback whales songs have been widely investigated in the past few decades. This study proposes a new approach for the classification of the calls detected in the songs with the use of Hidden Markov Models (HMMs). HMMs have been used once before for such task but in an unsupervised algorithm with promising results. Here HMMs were trained and two models were employed to classify the calls into their component units and subunits. The results show that classification of humpback whale songs from one year to another is possible even with limited training. The classification is fully automated apart from the labelling of the training set and the input of the initial HMM prototype models. Two different models for the song structure are considered: one based on song units and one based on subunits. The latter model is shown to achieve better recognition results with a reduced need for updating when applied to a variety of recordings from different years and different geographic locations.

Contents

DECLARATION OF AUTHORSHIP

I Federica Pace declare that the thesis entitled Automated classification of humpback whale (*Megaptera novaeangliae*) songs using Hidden Markov Models and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:

Faria M A, DeWeerdt J., Pace F., Mayer F.X. (2011). Observation of a humpback whale (*Megaptera novaeangliae*) birth in the coastal waters of Sainte Marie Island, Madagascar. *Aquatic mammals* (In Review).

Samaran F., Gandilhon N., Prieto Gonzalez R., Pace F., Kennedy A., Adam O. (2012). Passive hydro-acoustics for cetacean census and localization. *CNRS Revue*.

Pace, F., *Benard, F., *Glotin, H., *Olivier, A. and White, P. (2010) Subunit definition and analysis for humpback whale call classification doi:10.1016/j.apacoust.2010.05.016, *Applied Acoustics,* 71(11), 1107-1112

Signed: ……..

Date:…….

Acknowledgements

I would like to thank the association CetaMada, and in particular Fifou Mayer and the volunteers, without whom this work could not have been possible. CetaMada offered logistics and financial support for the field work of this thesis, and its staff made this experience truly memorable. In addition, I would like to thank Dr Salvatore Cerchio and Dr Danielle Cholewiack for sharing their data and experience, and Prof Doug Cato for his advice at early stages of this project. In addition, I thank my supervisor Prof Paul White for his continuous support and passion towards this work, and Prof Olivier Adam for his feedback at all stages of the project and his collaboration on the field.

Abbreviations

FM – Frequency Modulated

HMM – Hidden Markov Models

AM – Amplitude Modulated

FFT – Fast Fourier Transform

DTW – Dynamic time warping

GMM – Gaussian Mixture Model

LPC – Linear Prediction Coefficient

MFCC – Mel-Frequency Cepstrum Coefficient

TE – Threshold of End

TS – Threshold of Start

S/R – Signal to Noise Ratio

HTK – Hidden Markov Model Toolkit

Glossary

Active space	Area surrounding an animal that has the potential to be utilised by that animal
Amplitude modulation	Phenomenon by which a signal is kept constant in frequency whilst it is modified in amplitude through time
Audiogram	Graph showing the hearing range of animals in relation to frequency
Baleen	Keratinous plate that is arranged transversally from the upper jaw of filter feeding whales
Biphonation	Non-linear phenomenon whose output is the presence of two independent fundamental frequencies in the spectrum of a call
Cetaceans	Marine mammal order identifying those species that have a torpedo-shaped body, are nearly hairless and have no hind limbs. The cetacean order includes all dolphins, whales and porpoises. Cetaceans are distinguished into two suborders, the mysticetes and the odontocetes
Conspecifics	Individuals belonging to the same species
Determistic chaos	Mathematical theory applied in various branches of science, that apparently random phenomena have underlying order. For audio signals it means that the signal appears random and unpredictable whilst following deterministic laws
Entropy	Lack of order or predictability; gradual decline into disorder
Evolution	The process by which different kinds of living organisms are thought to have developed and diversified from earlier forms during the history of the earth
Focal animal	During marine mammal observations, which are usually boat-based, the focal animal is the individual whose behaviour and interactions with the rest of the group are observed and recorded for a set period of time.
Frequency jump	Abrupt change in frequency of a signal
Frequency modulation	Phenomenon by which a signal is shifted in frequency over a very short period of time
Heterogeneous	Consisting of dissimilar parts
Intraspecific	Communication amongst members of the same species
Matched filter	In signal processing, a matched filter is obtained by correlating a known signal, or *template*, with an unknown signal to detect the presence of the template in the unknown signal.
Mysticetes	Suborder of cetacean identifying those whales that feed through filtering prey using their baleen structure
Odontocetes	Suborder of cetaceans identifying all toothed whales that feed by chewing or swallowing prey
Spectrogram	Visual representation of the energy content within a time window over the frequency spectrum of a signal
Subharmonic	Component of a periodic wave having a frequency that is an integral submultiple of the fundamental frequency

Tail slapping	Behaviour usually described in cetaceans that consists on swinging the tail at high speed in one motion to hit prey or the sea surface
Tonal vocalisation	Sound that is characterised by its regularity of vibration

1. Introduction

1.1 Motivation

This study aims to develop an algorithm to detect humpback whales' song units and to classify them systematically and objectively. Such an algorithm could be applied also to other animal species, particularly to the vocalisations emitted by marine mammals. The study of intraspecific communication is a very active field of research inspired by the curiosity of humans to understand certain animal behaviours and also to use biological inspiration to enhance manmade systems.

Research has highlighted the difficulties of analysing and classifying the calls emitted by animals in the wild starting from the data acquisition to the clustering task which is often very subjective and time-consuming.

This project is motivated by the need for an objective method for classifying the vocalisations produced by humpback whales, a well-known species to the scientific community, which will allow automatic detection and classification of the sounds within a recording, reducing human input, reducing analysis time and allowing for easier comparison between datasets.

We propose a novel approach based on the definition of sound subunits which is expected to yield a more accurate classification of the sounds emitted by humpback whales. The research carried out so far over the past 30 years was based on sound units as the basic building blocks of a song defined as "continuous sounds between two silences" (Payne and McVay, 1971) but no one questioned before if this is the best method to characterise such vocalisations.

The first chapter of this thesis presents the basic biology of humpback whales and some principles of underwater acoustics needed to understand the terminology used in subsequent chapters.

1.2 Humpback whale biology

1.2.1 Taxonomy

Although marine mammals of various species are commonly considered as a broad group by the common public and non-specialists, they belong to three different orders of mammals, namely these are the Carnivora (polar bears, seals, walruses, sea

lions and otters), the Cetacea (whales and dolphins) and the Sirenia (manatees and dugongs) (*Figure 1.1*).

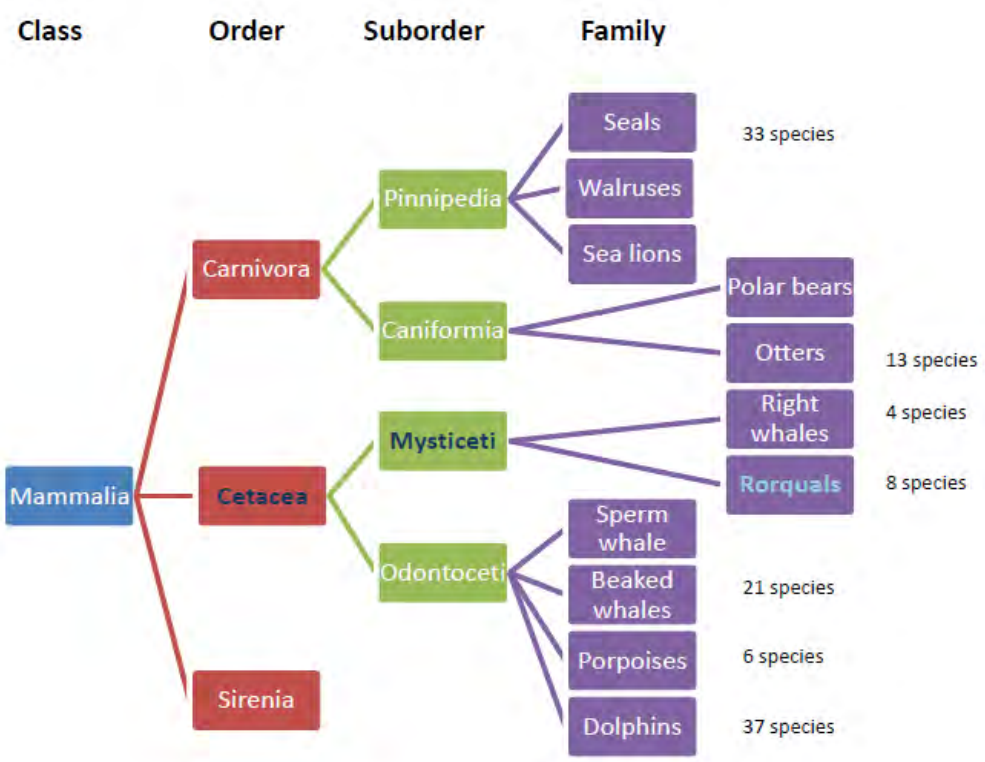


Figure 1.1: Phylogenetic tree of the animals commonly included in the broad category of marine mammals. The phylogeny of the humpback whale is highlighted with the blue font.

The suborder Mysticeti includes most of the species commonly referred to as whales; this suborder is differentiated from the Odontoceti on the basis of their feeding apparatus: all Odontocetes have teeth, which explains why the sperm whale belongs to this class. On the other hand Mysticetes possess baleen plates, structures made of keratin, layered with thick hairs for filter feeding.

Humpback whales belong to the family of rorquals having a peculiar throat structure consisting of longitudinal folds, which allow for huge expansion (Reynolds and Rommel, 1999). This characteristic allows whales to engulf massive amounts of water whilst filtering out minute prey items. Indeed, humpback whales feed extensively on krill and small schooling fish, which are trapped in the fine hairs present along the ridge of their baleen plates (Reynolds and Rommel, 1999).

Humpback whales employ a diverse range of hunting strategies: they can attack fish directly or stun them before consumption and they can also associate with

conspecifics to perform more complex techniques such as bubble net foraging (Reynolds and Rommel, 1999; Leigthon *et al.*, 2004).

1.2.2 Ecology and Behaviour of the humpback whale

Humpback whales (*Megaptera novaeangliae*) are amongst the largest baleen whales (Suborder Mysteceti) as adults can grow up to 16 metres in length (Reynolds and Rommel, 1999). These marine mammals can live up to 70-80 years and weigh around 25-30 tonnes (Reynolds and Rommel, 1999). They are grey in colour except for their characteristic long flippers (Figure 1.2a) and the underside of the fluke (Figure 1.2b) which present white patterns. Such patterns in the tail are very useful for photo identification purposes along with the indentations that may be present along the edges of the fluke. These marks are easy to distinguish and allow one to recognise individuals confidently (Constantine *et al.*, 2007).

Although the English common name "humpback whale" derives from its distinctive characteristic hump on the back before the dorsal fin; another peculiar feature of this animal is its large, elongated, white pectoral fins (

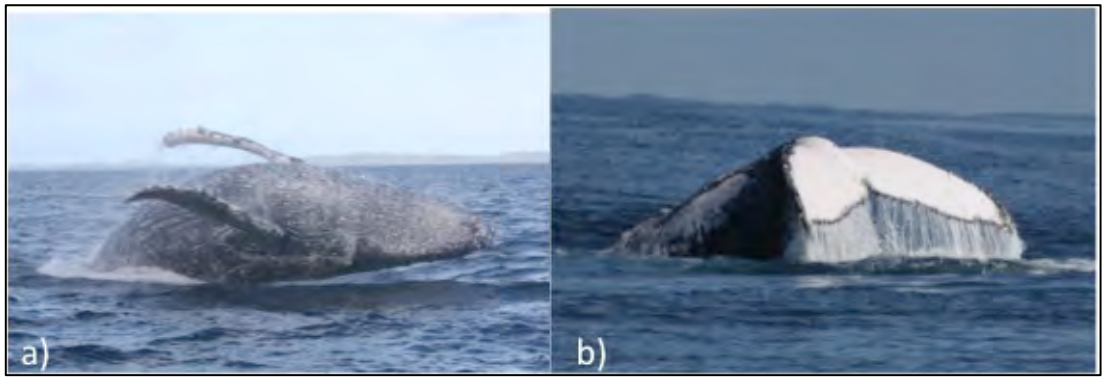


Figure 1.2). Indeed, the Latin name of the humpback whale *Megaptera novaeangliae* refers to this peculiar characteristic, from the greek "*Mega"* means giant and "*ptera"* means wings. The term *novaeangliae* meaning New Englander is thought to refer to the fact that the first scientist to identify the humpback whale sighted this animal regularly in the waters of New England.

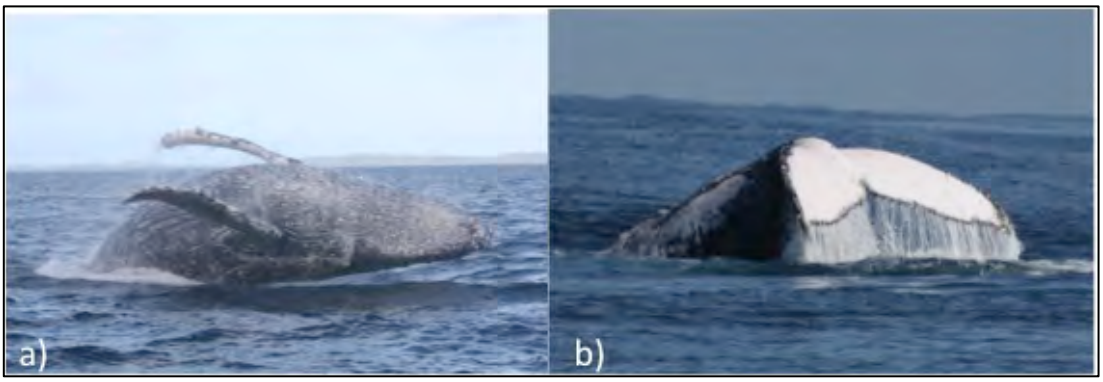


Figure 1.2: Humpback whale jumping out of the water showing its characteristic pectoral fins (a) and fluke of a diving individual (b) (photos taken by the author during the field seasons in Madagascar)

Humpback whales have a worldwide distribution; indeed they can be found anywhere in the oceans between the Antarctic and a latitude of approximately 65° N exceptions include the Mediterranean and the Baltic Seas (Reynolds and Rommel, 1999). Their population size is estimated to be around 80,000 individuals (NOAA, 1991); although this might seem a large number relative to other cetacean species, these marine mammals were endangered until a few years ago because their numbers dropped dramatically due to commercial whaling. Since the ban on whaling, humpback whales numbers have increased and they are now a species of least concern, accordingly to the species Red List (IUCN, 2011) (Reilly *et al.*, 2008). The impact of whaling on humpback whale populations was particularly noticeable in the North Atlantic where by the 20th century there appeared to be only 700 individuals left; for this reason, the USA National Oceanic and Atmospheric Administration (NOAA) implemented a successful conservation plan (NOAA, 1991).

Two major populations live in the North Atlantic Ocean and another two in the Pacific Ocean can be distinguished, although little is known about the migration routes of these animals and the relationship between populations, considering that they are present worldwide and can travel long distances on a seasonal basis. Recent research is trying to shed light on this matter by bio-logging and DNA sampling of individual whales to establish the intra-specific taxonomy.

Humpback whales migrate during the year from their reproductive grounds in warm waters to cold water areas that are rich in food outside the breeding season (Figure 1.3).

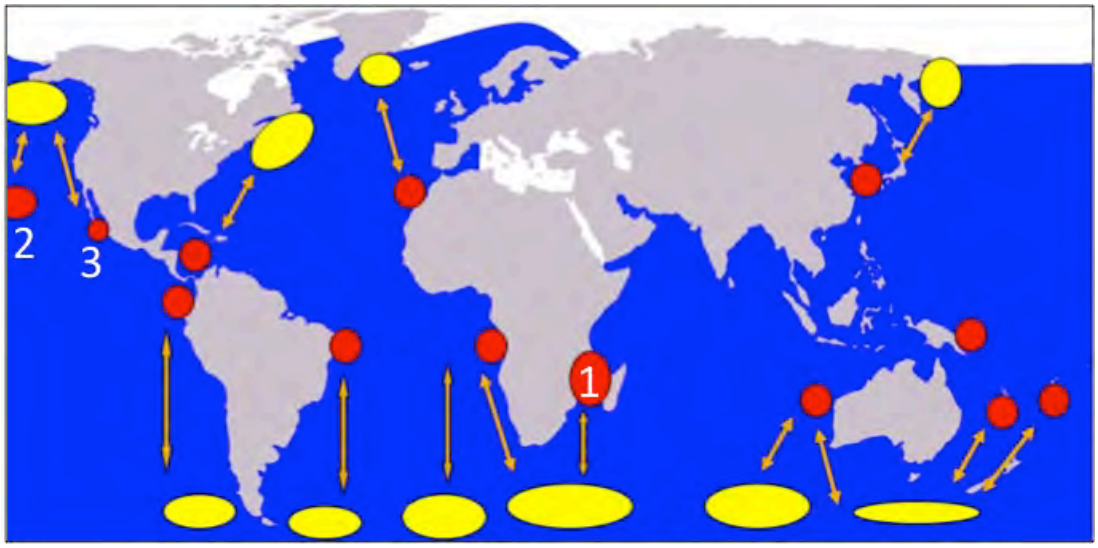


Figure 1.3: Map of humpback whale distribution – blue shaded areas are areas where these whales are found. Yellow circles represent foraging grounds while red circles are areas where humpbacks gather for mating. Possible migration routes are represented by the orange arrows. The location of populations whose songs are analysed in this thesis are numbered in the figure as 1) Ile Ste Marie, Madagascar, 2) Kauai, Hawaii, and 3) Socorro, Mexico

During the breeding season, humpback whales gather in low latitude regions (Figure 1.3) in large numbers and typically the males perform prolonged and complex songs that are believed to attract females. Populations of one hemisphere are segregated from the ones of the opposite hemisphere because their seasons are reversed; in other words, it is unlikely that an individual from the Northern Hemisphere will mix with a population of the Southern Hemisphere and vice versa because whilst the former is at equatorial latitudes to breed, the population of the opposite hemisphere will be foraging in Antarctica.

Females usually breed every 2-3 years and gestation lasts 11.5 months. New born calves are usually the length of their mother's head and they are fed by the mother for six months with a milk which is very rich in fat (ca. 50%), a quality that is important to provide the blubber layer and mass necessary to survive in cold waters (Reynolds and Rommel, 1999). Subsequently - for another period of approximately 6 months – the calves will be partially fed by their mothers but they will also have to start feeding independently. On the other hand, males do not share parental care

(Cerchio *et al.*, 2005; Smith *et al.*, 2008) even though singing males often join mother-calf pairs whilst travelling (Smith *et al.*, 2008) for reasons that are not yet clearly understood. Humpback whales reach sexual maturity between 5 and 7 years, while they still continue growing.

1.3 Principles of underwater acoustics

Sound is a longitudinal pressure wave which travels through an elastic medium: as pressure is applied to the medium particles move locally within the medium forming areas of high pressure (compression regions) and areas of lower pressure (rarefaction region) producing a sound wave. Acoustic pressure is defined as the difference between the local compressions and rarefactions and the surrounding ambient pressure. The local motion of the medium associated with these pressure fluctuations is referred to as particle motion.

Sound can propagate in water relatively easily and for this reason it is used as the main communication tool by many creatures that exploit this environment. Also humans rely on acoustics as a tool for the exploration of seas. However, propagation of sound underwater is affected by a series of environmental factors that make it complicated to predict the behaviour of sound waves in oceans.

In particular, temperature, pressure and salinity affect the sound speed to different extents; indeed, one factor can prevail over another depending on the location of the sound source in terms of depth and latitude (*Figure 1.4*).

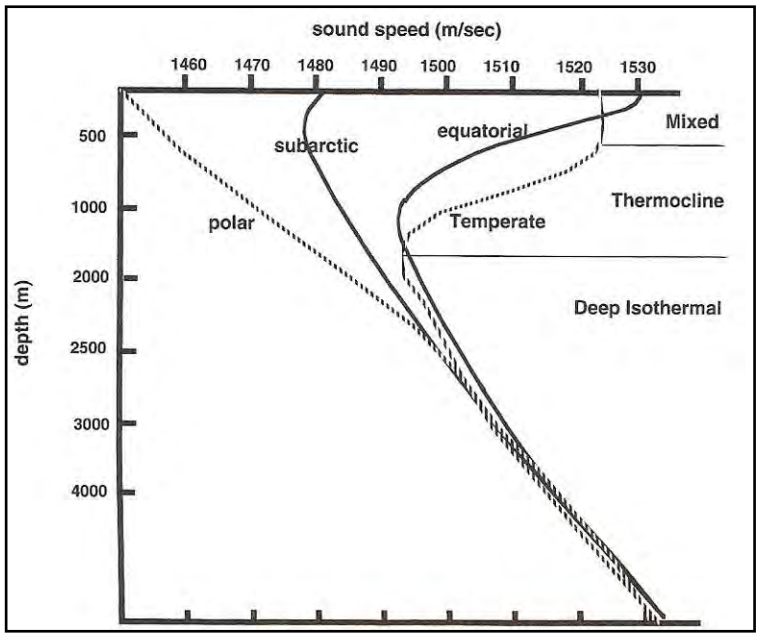


Figure 1.4: Sound speed profile for equatorial, temperate and polar conditions (Leighton, 1998). Note that the major changes occur in the top 1000 meters where temperature has a greater influence on sound speed.

These characteristics exert their influence on sound speed according to Equation 1.1 (Leighton, 1998):

$$c_0 = 1449.2 + 4.6\,T - 0.055\,T^2 + (1.39 - 0.012\,T)(S - 35) + 1.74\times10^{-6}P_h$$

Equation 1.1

Where c_0 is the speed of sound in water in m/s, T is temperature in °C, S is salinity (parts per thousand) and P_h is the hydrostatic pressure (Pa).

The sound speed in water is generally slightly less than 1500 m/s which means that sound in water travels nearly 5 times faster than in air (c = 340 m/s in air). In addition, sound underwater can travel greater distances than in air because the loss mechanisms tend to be smaller. This enables animals to communicate even when they are far apart and also when the visibility is very limited due to poor light conditions or to the water being murky. However, it is hard to predict accurately the distance travelled by a sound underwater because sound speed changes with depth resulting in the acoustic rays bending rather than travelling in straight lines. This phenomenon is known as refraction and it makes it arduous to predict the losses that occur as the sound propagates through the medium.

Two important factors need to be taken into account when a signal needs to be transmitted to a receiver at distance: geometrical spreading and attenuation.

Collectively these two factors contribute to what is called the transmission loss. The loss due to geometrical spreading can be approximated by modelling the sound as spreading spherically in close proximity to the source and cylindrically away from the source.

The intensity of an acoustic wave reflects the rate of flow of energy through a unit area and has the units of Watts/m^2. The instantaneous intensity of an acoustic wave, *I*, is usually expressed as

$$I = \frac{p(t)^2}{\rho c} \qquad \textit{Equation 1.2}$$

where $p(t)$ is the acoustic pressure, ρ is the density of the medium and c is the speed of sound in the medium. The quantity ρc is referred to as the specific acoustic impedance. In practice, intensities are usually averaged over short time periods *using:* $I = \frac{1}{\rho c}\frac{1}{T}\int_0^T p\,(t)^2 dt$ *Equation 1.3*

T is an integration time selected to capture the short-term behaviour of the pressure signal, e.g. if $p(t)$ is periodic then T would be selected to be one period of the signal. The above formulation of the intensity assumes that the incident sound is a plane wave, which is when the measurement point is distant from the source and distant from any reflective boundaries.

The range of acoustic intensities encountered in practice covers more than 10 orders of magnitude. To accommodate such a large dynamic range of values it is normal to express intensity on a logarithmic scale, this measure is referred to as the Sound Pressure Level (SPL) and is usually expressed in decibels (dB) relative to a specified reference intensity.

$$SPL = 10log_{10}(\frac{I}{I_{ref}}) \qquad \textit{Equation 1.4}$$

Where I_{ref} is the intensity of a reference waveform.

It is important to note that the reference pressure in water is different than in air, specifically p_{ref} in air is dB re 20 µPa whereas in water p_{ref} is equal to dB 1 µPa. This means that SPLs of sounds transmitted in air versus water cannot be directly compared but they need to be adjusted to account for the difference in reference pressure (Leighton, 1998). In addition, the reference intensity I_{ref} is different in the two media because sound speed is different. Therefore, from Eq. 1.3 it follows that

$$I_{air} = \frac{p(t)^2}{\rho c} = \frac{p(t)^2}{340(\frac{m}{s})\times 0.0013(\frac{g}{cc})} \qquad \textit{Equation 1.5}$$

$$I_{water} = \frac{p(t)^2}{\rho c} = \frac{p(t)^2}{1500(\frac{m}{s})\times 1.575(\frac{g}{cc})} \qquad \textit{Equation 1.6}$$

We can obtain a numerical value of the pressure difference required for an animal to have an equal sound percept in the two media, i.e. p_{diff}, assuming the reference pressures were identical:

$$p^2{}_{diff} = I_{air} = \frac{p^2}{340(\frac{m}{s})\times 0.0013(\frac{g}{cc})} = I_{water} = \frac{p^2}{1500(\frac{m}{s})\times 1.575(\frac{g}{cc})} \qquad \textit{Equation 1.7}$$

$$p_{diff} = \sqrt{3495} = 59.2$$

Converting the pressure to Decibel scale the equivalent level difference is ~35.5 dB. As a rule of thumb, after considering the level difference due to the use of different reference pressures, the numerical SPL in water can be thought of as being reduced by ~61.5 dB to be comparable to a level reported in air. However, this ensures thatthe two waves are equivalent and does not account for the individual sensitivity that different species have underwater; therefore, it is better to avoid drawing comparisons between sound levels in air and under water.

Attenuation loss includes the effect of absorption and scattering; therefore, its determination is extremely difficult because many factors need to be taken into account. In sea water, absorption cannot be disregarded, especially at higher frequencies, because it assumes quite high values compared to fresh water.

In addition due to the characteristics of water, sound can be deformed by the morphology of the sea floor and the depth of the water column. All of these factors can be modelled using various technologies to estimate, for example, the distance of a singing whale from the hydrophone. Nonetheless, this was not an objective of the current study as the animals whose songs we recorded were relatively close to the hydrophone and absolute measurements of the loudness of the sounds produced were not of interest here.

An active space of a species was defined as the spatial distance between a sender and a receiver as the area in which an animal can detect and perceive a conspecific and therefore communicate with him (Janik and Slater, 1998). Such a definition takes into account both the hearing ability of marine mammal species and the critical ratios for masking sounds, meaning that the effect of background noise is included in the

analysis (Janik, 2000). This is relevant because noise could mask the vocalisations to a level at which these can be detected but the content of the message is lost; this instance is therefore no longer considered as communication.

The active range of humpback whales has been experimentally measured and estimated to be 15 to 160 km at 0.02-8.2 kHz. These results seem to be consistent with the behaviour of the animals but a theoretical calculation for this species has not been produced yet because of the many factors that need to be taken into account when studying underwater sound propagation and no audiograms are currently available for this species. In particular, the hydrophone sensitivity might lower than that of the whales to sounds emitted by their conspecifics. So far, the active space range has been estimated theoretically only for two cetacean species, killer whales (Miller, 2006) and bottlenose dolphins (Janik, 2000).

1.4 Acoustics of Humpback Whales

The mechanisms underlying sound production and reception in cetaceans are still poorly understood, particularly in baleen whales because these large animals cannot be held in captivity for experimental studies. Hence, most of the information about the anatomy of the ear and vocal apparatus in mysticetes comes from dissections of stranded animals.

1.4.1 Sound production

The anatomy of the vocal apparatus in baleen whales was extensively described in a recent publication of (Reindenberg and Laitman, 2007) where they examined 8 humpback whales of various age groups. The major difference with the apparatus of terrestrial mammals is that these whales lack vocal folds; this is the reason why the actual mechanism for sound production is still not fully understood. However, a homologous structure to the vocal folds has been identified, i.e. the U-fold (Figure 1.5: Diagram). The latter is oriented parallel to the airflow rather than perpendicularly, as opposed to terrestrial mammals (Reindenberg and Laitman, 2007) and sounds may be generated by the air flowing between the laryngeal lumen and the laryngeal sac.

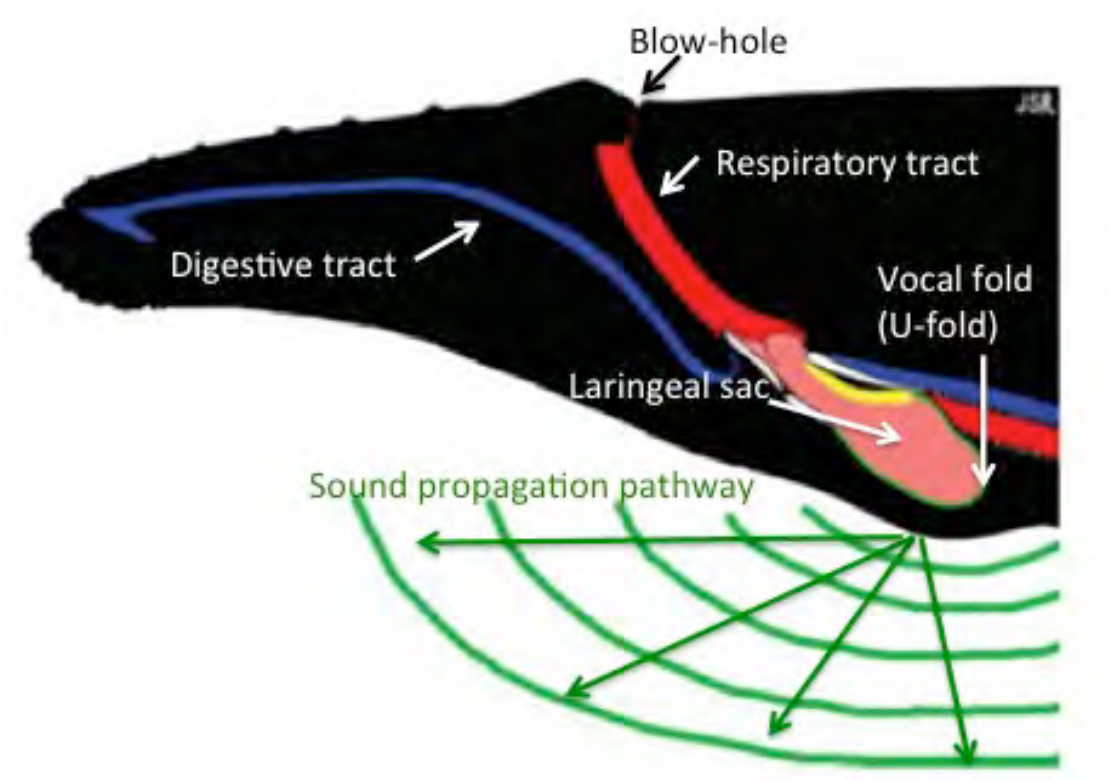


Figure 1.5: Diagram representing the location of the larynx within the body of a baleen whale. The red thick line represents the respiratory tract, the blue line the digestive tract and the white is cartilage. The laryngeal sac (pink) during sound production and the resulting sound pressure radiates from the ventral part of the whale's body (green waves) through the ocean (adapted from Reindenberg and Laitman, 2007).

The U-fold (Fig. 1.5 - green thin line) is very thick compared to other mammals allowing for the production of low frequency sounds and its surface is very flexible and elastic so that it can stretch extensively and move in various planes; such morphology is consistent with the fact that the sounds emitted are very diverse. Therefore, it is believed that by varying the U-fold conformation – i.e. its length, tension and gap between the membranes – the animal can control the sound output.
Another mechanism that might be responsible for sound production is the vibration of the U-fold due to the air flowing from the laryngeal sac and passing through the fold making its edges vibrate. Whilst the U-fold seems responsible for the generation of the fundamental frequency, the laryngeal sac is thought to play a role in further modification of the sound and its transmission (Reindenberg and Laitman, 2007).

Despite the fact that dissections of stranded animals allowed scientists to gain an insight in the anatomy of the vocal apparatus of humpback whales, the exact mechanisms of sound production are still unknown. It is especially intriguing that males can stay underwater for approximately 20 minutes whilst continuously singing and that no air is exhaled during this process (no bubbles are produced). This suggests that humpback whale might have a mechanism to recycle air, similarly to the way in which dolphins and other odontocetes produce sounds pushing air between air sacs.

1.4.2 Sound reception

The study of the mechanism for sound receptions in cetaceans are limited to post-mortem analysis of their auditory system and few audiograms obtained through evoked potentials and behavioural studies carried out on small cetaceans and seals held in captivity. Intuitively, such studies have been possible on smaller species that can be captured and kept in a controlled environment and not on the larger whale species. Research on odontocetes showed that they perceive sounds through their oil-filled lower jaw (Ketten, 1994; Wartzok and Ketten, 1999) and then transmit it through the middle ear, which acts simply as a conductive structure, to the inner ear. However, there is no evidence that the mandible might have such a function in mysticetes.

The ear of baleen whales is still connected to the rest of the skull, unlike toothed whales, suggesting that sound conduction may occur via the bones and also the soft tissue connecting to the skull (Ketten, 1994). The basilar membrane of humpbacks is very well innervated, allowing for an accurate transduction of the sounds into electrical signals. In general, estimation of the hearing abilities of humpback whales and other large whales are based on the assumption that their best hearing should correspond to the frequencies at which they produce most sounds, i.e. between 30 and 21,000 Hz. These are low frequencies compared to the range of hearing of odontocetes, which can reach 200,000 Hz. The range of hearing and sound production in humpback whales is similar to the range of human hearing – indeed most energy in their calls is below 4 kHz; for this reason several methods applied for the feature extraction of their vocalisation are based on logarithmic filters that are applied to human speech (Rickwood and Taylor, 2008).

1.4.3 Sound characteristics and usage

Humpback whales are probably the most vocal baleen whales. They produce sounds, known as vocalisations or calls, in a social context to communicate with their conspecifics during the breeding season or on the feeding grounds. Their complex repertoire has been increasingly studied since the 1970's as it elicited a lot of attention from the general public.

Humpback whales produce numerous and extremely variable vocalizations, used by males and females in social interactions. These are referred to as social sounds as opposed to songs, which are composed of a juxtaposition of vocalisations that are repeated several times in a very specific pattern. Songs are only produced by males and almost exclusively during the breeding season. Songs are the subject of investigation of this thesis and will be discussed more in detail in the next chapter.

Early researchers attempted to classify the sounds emitted by *M. novaeangliae* in Alaska in relation to their behaviour (Thompson *et al.*, 1977; Thompson *et al.*, 1986); these included calls as well as sounds associated with the activity of the animal. The categories of sound thus identified were "moans", "grunts", "pulse-trains", "surface-impacts" and "blow-hole associated sounds".

More recently Dunlop and colleagues (Dunlop *et al.*, 2007b) described the social sounds recorded for the humpback population that migrates to the Australian waters during the Southern winter. They identified 34 separate call types after analysing more than 600 vocalisations based on their frequency characteristics. The majority of the call types identified also appeared in the songs analysed which means that the same vocalisations produced for social purposes are then used by males to compose their complex songs.

1.5 Structure of thesis and contribution to current knowledge

This introductory chapter and the next one, which will review the current knowledge on humpback whale songs and their purpose (Chapter 2), set the scene to understand the complexity of humpback whale song and the challenges faced in developing an automatic classifier for this type of signal, which is the main objective of work presented. The subsequent chapters will shift the focus to the signal processing aspect, starting with a description of the algorithms that are commonly employed in bioacoustics and a background in speech processing in Chapter 3. A key aspect of the work presented is the fact that signal processing tools commonly employed to

classify human speech are applied to humpback whale songs. Specifically, we will investigate how successful Hidden Markov Models are to perform such a task (Chapters 6-8). This is not the first time that Markov chains have been used for classifying complex bioacoustics signals but it is the first time that they are used to classify the building blocks of humpback whale songs. A key aspect of the work is identifying the correct components to be able to classify humpback whale songs correctly using an automatic classification algorithm. For years, songs have been manually classified to investigate how they change through the years within and between different whale populations. However, a lack of clarity exists on the definition the elements on which to base such classification when comparing songs across different research groups. The main original element of the work presented here is formally defining the building blocks of humpback whale songs, i.e. subunits as opposed to units, and using them to train the classification algorithm (Chapter 7). The main recording site in Madagascar for humpback whale songs analysed for this project and the methods used to obtain the data are presented in Chapter 4. The Madagascar population is currently understudied compared to other humpback whale populations (e.g Australian and North American), possibly because of the difficulties in organising the field-work logistics for the infrastructure limitations, and the fact that whales are present there during the rain season, limiting the time windows during which high quality recordings can be collected. During the course of this study, songs of humpback whales have been recorded over the course of 4 years, contributing significantly to the data that is available for this particular population. In Chapter 8, recording site for humpback whale songs analysed for this project the used to obtain the data comparisons are drawn between the performance of the classifier based on songs collected in different geographical areas, and the biological implications of the findings will be discussed. Lastly, the performance of the automatic classifier and its development potential will be evaluated in Chapter 9.

2. Song definition, usage and classification

Whale sounds have been recorded since the 1960s. Initially, this was done infrequently using tapes but started getting more attention when the resources for underwater acoustics research was shifted from the military to the state of marine life, as the Cold War was brought to an end. It is not surprising that the first humpback whale sounds were recorded by the US navy through their hydrophones installed in Hawaii (Johnson and Tyack, 2003), although at that time humpbacks were not confirmed to be the source of such sounds. Shevill and Watkins (1962) first identified the source of sounds as humpback whales, using recordings made by Frank Watlington at a hydrophone station in Bermuda (Payne and Mc Vay, 1971). Payne and McVay recorded the sounds themselves in Bermuda and presented indisputable evidence that humpback whales were the source of the particular sounds recorded in Bermuda and Hawaii in a paper that pioneered research on humpback whale songs.

2.1 Song definition and structure

The structure of Humpback whale songs were first defined by Payne and McVay (1971) noticing that the sounds emitted were long sequences of vocalisations arranged in a very specific pattern. Indeed, such rhythmic structure is comparable to the songs of birds, as observed by the researchers. The term song is intended to refer to one of the three meanings discussed by Bremond (1963), i.e. a series of notes uttered in a succession to form a recognisable sequence or pattern in time. Songs are distinguished from the so called "social sounds" which are calls emitted by whales in an irregular and unpredictable manner not only through their vocal apparatus, but also by slapping their flukes or pectoral fins on the sea surface (Dunlop *et al.*, 2007c).

The building block of a song is called a unit and is defined as a continuous sound between two silences. Units are typically above 1 s long but they can vary, ranging from less than 1 s to 5 s (Miksis-Olds *et al.*, 2008). The length of one type of unit can also change throughout a song, a characteristic, which poses a challenge for automatic classification algorithms, as will be discussed in the next chapter. Two or more units can be repeated by the whale in a specific pattern to form a phrase, as depicted in *Figure 2.1*.

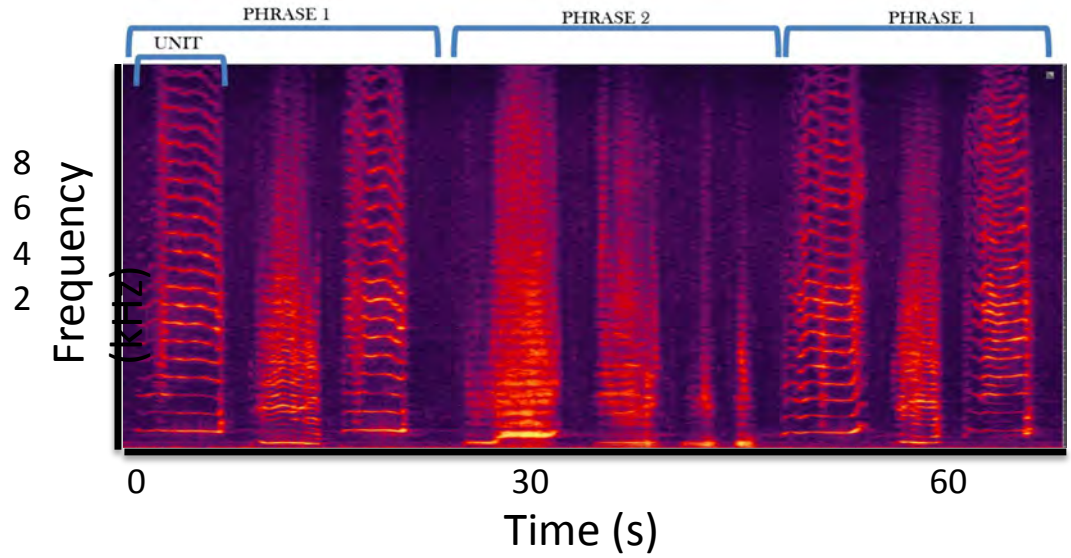


Figure 2.1: Example of two phrases of humpback whale songs recorded in Hawaii in 1989 (recorded by Salvatore Cerchio). The number of units forming a phrase can vary as well as its duration.

Phrases are the combined and repeated several times to form a theme (*Figure 2.2*: Three themes recorded in Australia (Weinrich and Corbelli, 2009).
).

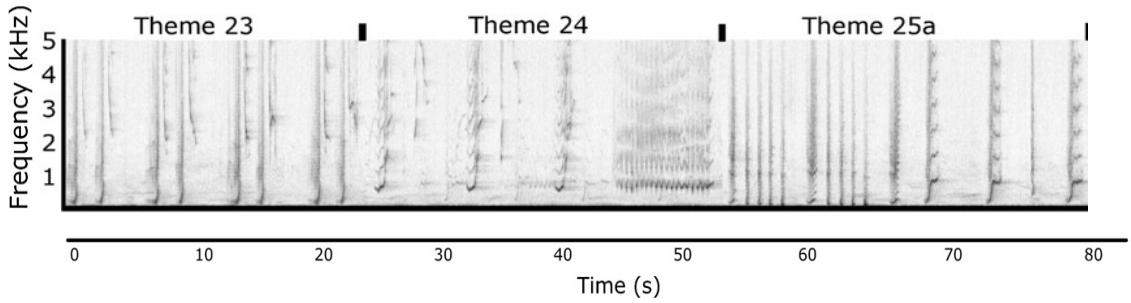


Figure 2.2: Three themes recorded in Australia (Weinrich and Corbelli, 2009).

The silent interval between two units is usually approximately constant within phrases but minor variations can occur within a theme. Silences are usually 2-5 s long (Miksis-Olds *et al.*, 2008). The alternation of units with silences of similar duration confers a song its rhythmical structure.

A song is composed of several themes. The period of a song session corresponds to the interval between surfacings of the whale to breath. The result is a longer pause between units than usually observed within a song.

The hierarchical structure of the song presented by Payne and McVay is depicted in of *Figure 2.3*: Diagram showing the terminology used to describe humpback whale songs as defined by (Payne and McVay, 1971)

.

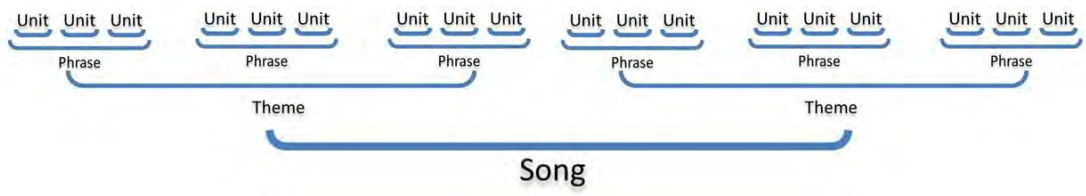


Figure 2.3: Diagram showing the terminology used to describe humpback whale songs as defined by (Payne and McVay, 1971)

songs as defined by A song is typically 16-25 minutes long, corresponding to approximately 100 units. The length of songs varies accordingly to how many times its components are repeated as there is no fixed length for any of the components. In other words, a phrase can be constituted by 2-6 units, and phrases might be present only once or repeated several times within a theme and so on. Most of the energy of the calls is contained between 30-4000 Hz (Mercado III *et al.*, 2008), and source levels vary between 140-170 dB re 1µPa according to sound type (Au *et al.*, 2006). Concatenations of songs are termed song sessions and they can last for several hours and in some cases more than a day. A study of 48 independent song sessions of humpback whales in Australia reported lengths of about 300 to 3100 units per song session (Miksis-Olds *et al.*, 2008). This shows not only how variable song sessions can be but also how much data researchers collect in the field and subsequently analyse to understand the means of communication of these whales, highlighting the need for an automatic classifier.

2.2 Song characteristics

Since songs were discovered in 1971, researchers started recording them and analysing their structure building up a substantial catalogue. Research was particularly active in the Northern hemisphere, specifically in Bermuda and Hawaii where humpback whales were commonly encountered during their breeding season. Payne observed that throughout the season songs changed only slightly and different whales recorded simultaneously sang essentially identical songs. However, songs sang in the same breeding area changed from one season to the next (Winn *et al.*,

1981). These changes affected only a proportion of the themes rather than the entire song so that it was possible to relate songs from one year to the preceding one.

First comparisons across populations of humpback whale were drawn from recordings taken in Hawaii, Mexico, Cape Verde, the West Indies and Tonga, a little Island off the East Coast of Australia (Winn *et al.*, 1981). In that study, three main findings were reported:

1. Songs recorded simultaneously in Hawaii and West Mexico were nearly identical.
2. Songs recorded in the West Indies and Cape Verde were similar to each other but different from the ones recorded in Hawaii and Mexico.
3. Songs recorded during the same year in the Southern hemisphere in Tonga were different.

These results suggested that populations breeding in different ocean basins had a different dialect, which was not surprising for populations of the Northern hemisphere versus the Southern hemisphere that are temporally and geographically isolated because they are extremely unlikely to mix given that their breeding and feeding seasons occur at different times. In addition, the fact that populations of the Atlantic and Pacific basins produced different songs suggests that these populations do not mix when they are both in the Arctic during the foraging season. Lastly, it was striking to discover that songs of distant populations in the same ocean basin were nearly identical. Given its complex structure, it is extremely unlikely that geographically isolated populations developed the same dialect independently which implies that there is some biological phenomenon happening (Winn *et al.*, 1981).

As previously mentioned humpback whales have a worldwide distribution and consequently research groups in other areas started studying *M. novaeangliae* songs extensively in the past 15 years. Australian researchers have been particularly active in studying the changes in songs from year to year and comparing songs recorded in different areas of the continent, aided by the introduction of sono-buoys that allow continuous recording. The fact that different populations of humpback whales migrate each year to the East and West Coast of Australia for mating makes this country an ideal candidate for studying the song repertoires of different populations. Analysis of Eastern Australian recordings from 1995-1998 revealed that humpback whales copy each other's songs (Noad *et al.*, 2000) (*Figure 2.4*).

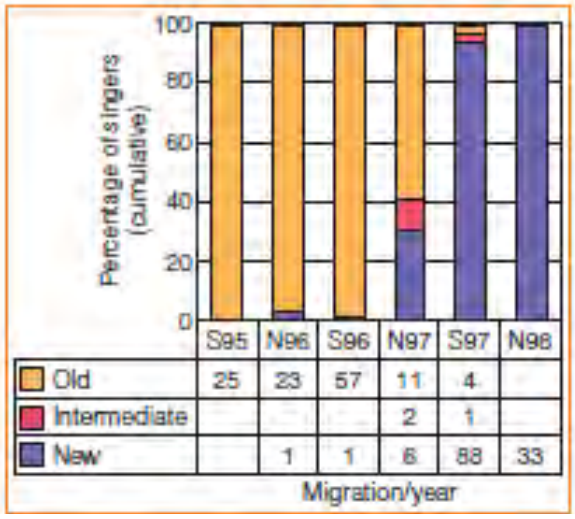


Figure 2.4: Bar graph depicting the relative prevalence of song sang by males in the waters off Queensland, Australia, migrating Southwardly (S) at the end of the breeding season and Northwardly (N) at the start of the breeding season during 1995-1998. Each colour represents the song type sang by individual males during a particular migration pathway, for instance, N95 is the Northward migration of year 1995:the numbers given in the table indicate the number of males singing a song type during that migration (Noad et al., 2000).

In 1995, all singers within the Queensland breeding ground swimming Southwardly sang the same song (S95). During the Southward migration whales leave the reproductive grounds to reach the nutrient rich waters of Antarctica, where different populations have the potential to mix and interact with each other because they are no longer geographically isolated. Whales start migrating to the north to reach low latitudes where they can reproduce during the winter months. The following year when whales return to Australia from the Antarctic feeding grounds, a new song recorded from a single singer was introduced to this location, which was identical to the songs produced by humpback whales breeding off the West coast of Australia in those years (N96 in Fig 2.4). In 1997, the main song sang in Queensland was the one that was rare the previous year and by 1998 the old song had completely disappeared from the repertoire of the whales present on the breeding grounds (S94 in Figure 2.4). This phenomenon is deemed as a cultural revolution because within only 2 years the new song had completely replaced the previous song that was sang by all whales except for one individual. Evolution, on the other hand, acts over much longer timescales and is usually caused by a large influx of immigrants, rather than a

single individual. Cultural evolution has been previously observed in bird song whereby changes in song structure are accumulated over years as birds started learning sections of the new song, eventually leading to the replacement of the old song (Noad *et al.*, 2000). The fact that humpback whales in Australia learned a complex new song in just two years suggests that changes in song structure are driven by novelty.

2.3 Song usage

Sound frequency and intensity are important in the determination of song usage. Mysticetes use primarily low frequency sounds suggesting they are aimed at communicating over long distances (Wartzok and Ketten, 1999). This is because high frequencies are attenuated faster than low frequencies, and, given the same environmental conditions, a lower frequency sound tends to travel further. Such considerations have led scientists to think that humpback whale songs are mainly used to attract females. This view is supported by the fact that vocalizations are performed mainly by male individuals during the mating season and also by the complexity of songs which might have arisen through sexual selection according to the runaway theory[1] (Payne and McVay, 1971).

Since the discovery that humpback whales were capable of singing, researchers have been trying to understand the reasons behind this complex form of communication. To understand why humpback whales sing, one must first investigate the ecology and behaviour of these whales, as well as the characteristics of songs. In the previous section, songs were defined as sequences of sounds that form a specific pattern; this definition suggested a hierarchical structure of songs, which was confirmed in observations of researchers on a global scale and more recently demonstrated quantitatively in two studies on the entropy of humpback whale songs (Suzuki *et al.*, 2006; Miksis-Olds *et al.*, 2008). Comparisons with birds' acoustic behaviour are inevitable because a wide number of bird species are known to

[1] Fisherian Runaway Theory: sexual selection model proposed by R. A. Fisher in 1915. Fisher explained that selection of certain traits would benefit an animal not in terms of survival but in terms of fitness for increased chance of reproduction. The term 'runaway' refers to the fact that this selection process will favour the development of more pronounced traits with time that differentiate individuals.

produce songs and several studies have addressed the motivation for such complex display.

The first studies reported the occurrence of songs in Hawaii and Bermuda, which we know to be breeding grounds of this species. One reason why birds sing is as a sexual display because males sing only during the spring when the hormones levels in their bodies raise in preparation for mating with their female conspecifics (Wada, 2010).

Indeed, researchers agree on the fact that there must be an evolutionary advantage to whales if they invest resources in singing. Such observation results from the fact that singing unequivocally has a cost for the singer because he will have to invest metabolic energy in producing vocalisations and will also attract unwanted attention exposing him to increased risk compared to whales that do not sing.

Several theories proposed to explain why humpback whales sing are reviewed in the following sections. These theories are by no means exclusive; indeed, behavioural patterns can have more than one function. For instance, bird song appears to have the dual function of attracting females and defending the territory from other males. In addition, in canaries, song appears to have a stimulating effect on the ovaries of females which means that they will produce more eggs increasing the chances of successful reproduction Slater (2001).

2.3.1 Sexual display

Sexual display is proposed as the primary explanation for song production by humpback whales for several reasons. Firstly, it is known that only males sing and they do so at the start of the breeding season in winter. Extensive research was conducted on the behaviour and ecology of humpback whales which showed that males arrive earlier than females on the breeding grounds (Erbe, 2002) and that they compete for access to females directly by physically attacking each other, for instance using tail slapping and indirectly through displays to attract females, e.g. breaching and pectoral fin slapping (Figure 2.5).

Figure 2.5: Behavioural displays commonly observed on humpback whales' breeding grounds. Breaching is when a whale leaps out of the water. Whales also may slap their tails and pectoral fins on the surface of the water, action which produces loud sounds that can be heard at some distance from the animal. Pectoral slapping is part of the courtship behaviour: in the picture above a female and a male were performing this display sometimes simultaneously as part of their courtship ritual. The male was also rolling on its side to display its large pectoral fin. All the above photographs were taken by the author during the 2009 field season in Madagascar.

Singing is performed almost exclusively during the breeding season and singers appear to be particularly abundant at the start of the season when males first arrive on site. Indeed, songs may convey information about the sexual fitness of the singer: songs are produced at high amplitudes and for several hours resulting in high energy expenditure for the signaller (Parsons *et al.*, 2008). In other words, the cost of producing the signal is a proxy of the fitness of the individual producing it because a fit individual is likely to have more energy and invest more energy into singing to attract a partner. In addition, scientists argued that the male fitness could be

conveyed by the song structure in terms of the ability of the male to hold its breath for a long time (Southall *et al.*, 2007).

Studies on the cost of singing have been conducted on birds both in terms of the metabolic expenditure of producing songs and the cost of being more susceptible to predation (IWC, 1996). Confirmation that singing is expensive comes from research on swimming speeds of male humpbacks which showed that males that sing travel at a speed of about 2.5 km/hr in their migration from Australia to Antarctica whereas non-singing whales travelled at 4 km/hr (Sousa-Lima and Clark, 2009).

Indeed, to demonstrate that songs are part of the sexual display of males it is necessary to demonstrate that females are attracted to singing males and that singers are more successful in mating than males that do not sing. Although on few occasions females are observed to approach singing males, usually when singers approach a female they also engage in behavioural courtship displays, which supports the idea that songs play an important role as sexual display (Tyack, 1981). So far, no report of humpback whales mating has been published, which means that it is impossible to clearly determine the role that songs play in the courtship ritual or whether successful males produce more complex or louder songs; therefore, other theories need to be taken into account.

2.3.2 Territorial marking

Charles Darwin was the first to note that male courtship behaviour could be directed at individuals of the same gender to warn them about their superiority and prevent physical competition (Cerchio *et al.*, 2008). The signal produced to deter other males from a territory can be visual, auditory, olfactory or a combination of these. Territorial marking is widely spread in the animal kingdom as males want to ensure access to females during the mating season.

When considering the hypothesis that songs play a role as sexual display to attract females, one needs to contemplate also that the acoustic signals may be directed at other males with the purposes of deterring possible competitors from their territory.

Research on humpbacks in feeding versus breeding grounds showed that not all females complete a full migration during the year and that some females may become pregnant prior to the beginning of the breeding season (Rosenbaum *et al.*, 1997; Slater, 2001). Further males willing to mate face a tough competition as they are more numerous than females: the sex ratio calculated in the northern and

southern migrations in East Australia was 2.4 males for each female (Rosenbaum *et al.*, 1997). Although sex ratios might differ slightly in other mating sites, it is evident that males compete vigorously for access to females, as typically, when active groups of whales are encountered, the female leads the group and is followed by up to 6 males who compete physically with each other to swim as close as possible to her (Shapiro *et al.*, 2011).

2.3.3 Detection of conspecifics

As sound is commonly used underwater by whales and dolphins for ranging purposes, it was suggested that humpback whale songs play a role in detecting females (Parsons *et al.*, 2008). This theory originated from the consideration that humpback whales are a migratory species and as such, when the animals return to the breeding site each year, the males need to be able to detect groups of females (Garland *et al.*, 2011). Furthermore, some of the vocalisations that make up song sequences are highly stereotyped and resemble sweeps that bats produce for echolocating. A mathematical model was developed which supported this hypothesis based on the vocalisations emitted by humpback whales (Frazer *et al.*, 2000). However, this sonar hypothesis was disputed by Stevick *et al.* (2011), questioning several of the assumptions that were made in the calculation of the sonar equation and discussing how behavioural aspects of humpback whales are in conflict with this hypothesis, and is therefore not widely accepted. Specifically, songs of humpback whales are constituted by units that vary from year to year and usually over the course of 5 years entire songs are completely replaced by new ones in a given population (Zimmer, 2011). This complete revolution in songs contradicts the sonar hypothesis because one would expect humpback whale calls to converge so that all whales produce a common vocalisation type that is the most suited for echolocation purposes and is preserved through evolutionary pressures. Furthermore, even within a song, the call sequence may vary significantly with fewer or more repetitions of the same call during a phrase, with calls varying in length and silences between successive units of different duration (Stevick *et al.*, 2011).

2.3.4 Male cooperation

A plausible theory about the function of humpback whale song is that it plays a role in male-male cooperation. Indeed, these baleen whales generally live alone during

the foraging season, whilst male associations are often observed on the breeding grounds. Such associations can be both competitive in the form of a series of males chasing after the reproductively active female (also known as cow) and cooperative, as in the case of juvenile males traveling together or escorts accompanying mothers and their calves.

Generally, singers are observed in isolation and most of the time they will be approached by other males (Darling and Berube, 2001), at which point they will stop singing and the animals will start travelling together (Tyack, 1981). On the other hand, females are rarely observed approaching singing males. These observations suggest that song is used to attract males and form pairs or groups that cooperate to get access to females rather than being used with the aim of directly attracting females (Parsons *et al.*, 2008). This hypothesis is supported by research that showed that males joining singers during female escorting behave more cooperatively than when song is not heard, suggesting that songs are involved in male interactions (Williams and Staples, 1992). However, previous reports on humpback whale behaviour report more aggressive displays between males, even in the presence of singers, to prevail over the others and swim as close as possible to the female (Tyack, 1981; Jurafsky and Martin, 2009). In addition, the evidence presented by Noad *et al.* (2000) shows that singers are driven by novelty, which means that they incorporate changes in old songs to quickly learn the new one; which is in conflict with the male cooperation hypothesis. Indeed, males that wish to communicate with each other to cooperate on the breeding grounds would be expected to try to conserve a common song structure rather than expending energy in learning songs from animals that belong to different populations.

2.3.4 Song usage summary

To conclude, several hypotheses have been proposed to explain why humpback whales produce such complex songs that are unlike any other acoustic communication known in marine mammals. There seems to be a general agreement amongst the scientific community that songs play a role in the breeding behaviour of these animals, as confirmed by evidence linking song production and hormonal levels; however, it is still not well understood if songs are mainly a means of communication between males or between a male and a female. In other words, no clear evidence tells us whether songs play a role in male territorial marking and /or

cooperation or if it is purely a sexual display, as songs produced by birds during the mating season.

Understanding the role of song production is certainly a complex task that is likely to be solved only through studies that combine behavioural observations with a detailed knowledge on song evolution. The work presented in this thesis about automated song classification is aimed at helping researchers in analysing songs more efficiently and being able to get further insight about how songs evolve across populations.

2.4 Song classification

In the attempt to understand the function of humpback whale songs and their evolution, scientists have been comparing song sequences across populations for the past thirty years and studying how songs evolve over the years. This type of work is extremely time consuming and consequently research effort has been invested in trying to develop tools to speed up this process and make it as objective as possible. Advances in computer technology and signal processing facilitated such research.

2.4.1 Brief history of bioacoustics and marine mammals

In the 1950's the main focus of behavioural studies shifted from visual communication to acoustic signals; this rapid development of bioacoustics studies is attributable to the development in technology that allowed scientists to easily record sounds that could be analysed in depth afterwards. In addition, instruments such as tape recorders opened the way for playback experiments: researchers could play specific sounds back to the subject animals in captivity to conduct fair tests. The availability of sound spectrographs was a further pull in favour of bioacoustics studies because scientists could inspect the structure of the sounds they heard with an objective tool and could detect changes in the signals even when they were not perceivable by the (human) listener. This meant that small changes in the structure of stereotypical calls could be explored further to understand if they were just caused by 'errors' or if they conveyed useful information from the signaller to the receiver.

As scientists studied started studying marine mammals with increasing interest, they realised that they needed a more powerful tool than visual observation to understand their behaviour and distribution because these animals spend most of their time underwater. In this context, acoustics seemed the most appropriate tool for this

purpose; indeed, marine mammals make extensive use of sound for ranging, communicating and, in some cases, establishing their social status (Reynolds and Rommel, 1999). The technology was available as underwater acoustics was developed during the Cold War for military purposes and could then be employed for environmental purposes. As in the case of military sonar, acoustic observations may be active or passive. Active monitoring consists on emitting a sound and then listening for the echo which is then analysed to detect and locate marine organisms. This method is used primarily for observing microorganisms and fish. On the other hand, passive acoustic monitoring (PAM) is widely used to investigate the presence of marine mammal species because most species of interest emit calls that can be detected passively. The advantage of the latter system is that no noise is added to the background noise reducing, or completely removing, any disturbance to the animals. PAM is also an attractive tool because both mobile and fixed sensors are available allowing scientists to collect data over large time scales, as well as short time acoustic data coupled with 3D movements (Nowak *et al.*, 2000).

Acoustic studies of marine mammals brought about the discovery that they produce a wide variety of vocalisations; for instance, it was found that bottlenose dolphins and killer whales possess a vast repertoire that reflects the complexity of their group dynamics. Notably, the former produce signature whistles that can be used to identify dolphins to the individual level (Janik and Slater, 1998).

The possibility of recording and storing large data sets for subsequent analysis is becoming commonplace in studies of marine mammals, leading to an increased demand for signal processing tools.

2.4.2 Review of humpback whale song classification

As previously mentioned, the songs of humpback whale have long been the subject of scientific and wider public interest (Payne and Mc Vay, 1971; Tyack, 1981). During the last few decades the analysis of these songs has been used to gain insights regarding the population dynamics (Winn *et al.*, 1981; Noad *et al.*, 2000) and sound production mechanism (Mercado III *et al.*, 2010). The increasing amount of data collected for these animals makes an automatic classification method for the systematic and objective analysis of datasets very attractive.

This section reviews the methods that have been used to classify humpback whale calls from 1970s to the present. Before discussing the different methods proposed, it

is worth highlighting that the success of the classification method is dependent on the task one wants to achieve. In other words, if the aim of a study is to investigate the song pattern, then it is essential to include all the calls produced by the animal; missing out one unit will modify the entire sequence. On the other hand, if one wants to analyse the acoustics associated with the behaviour of an animal leaving out a call is not critical on large data sets, i.e. discarding one data point will have a small statistical influence.

The first methods developed to classify sounds were put in place to analyse the song sequence - whose pattern is evident when one listens to it; these are called the "classical" methods and are based on manual spectrographic analysis coupled with listening. Such methods spun from the effort of Roger Payne, who was the first to define humpback whale songs as sequences of themes which are repeated in a cyclical way; each theme constituted by phrases, each phrase being a patterned association of units.

Despite the great development in technology, nowadays manual classification is still the most widely used method for the analysis of humpback whale songs, which are carried out at the phrase level. Indeed, it is relatively simple to manually classify phrases within a high signal to noise ratio (S/N) for a few samples; however, generally scientists need to study the structure of several hours of songs and then compare them across population and/or from one year to the next. This process becomes extremely time consuming and is further complicated by the fact that the units forming certain phrases can change slightly depending on the singer or the year in which the song is sang. Such variations add subjectivity to the analysis, which is a huge downside to a method that is used to conduct comparative studies.

Several methods have been proposed to achieve an objective and less time-consuming classification of humpback whale sound units but there are still issues to be resolved. Specifically, many algorithms require partial input of the user and/or cannot process large datasets, especially in real time, due to memory constraints. Visual inspection of spectrograms is important to estimate the validity of the automatic algorithms that are being developed (as in this study).

Automatic classification methods for humpback whale vocalisations have been borrowed from human speech analysis because they present some important similarities with human speech: most of the energy is contained below 4 kHz, calls

can have either quasi-periodic waveform (i.e. like voiced speech), or noisy appearance (like unvoiced speech), or be a combination of the two (like mixed speech) (Mercado III and Kuh, 1998). In addition, these whales possess an organ analogous to the vocal folds; although the specific pattern for sound production has not yet been discovered.

A variety of methods have been proposed to detect and characterise humpback whale calls as these include highly transient signals and tonal calls, which vary in duration and frequency. These are designed to capture the frequency components of the signal, which is the feature that most animals are more sensitive to (Deecke and Janik, 2006a). Whereas, variations in the call duration may be used as a strategy to ensure that the message emitted by the signaller gets across to the receiver.

Detection algorithms were based on wavelet analysis (Seekings and Potter, 2003) or on energy detectors (Rickwood and Taylor, 2008), where a threshold is set to distinguish calls from noise and silences; the latter method was chosen for this study. Amongst the feature sets employed to characterise humpback whale vocalisations are cepstrum coefficients, discrete Fourier transform, autocorrelation and linear predictors. Such variety of techniques adopted reflects the diversity of the calls as already discussed.

Neural networks, both supervised and unsupervised, are favoured for the classification task over simpler algorithms when analysing large datasets. This choice results from the fact that the occurrence of vocalisations is often uneven, i.e. some calls are repeated far more frequently than others within a song, and that neural networks provide a large degree of flexibility. The main advantages of adopting neural networks for such studies are that there is no need to make an assumption about the input characteristics.

A novel type of neural network, namely the adaptive resonance theory (ART) neural network, was proposed by Deecke and Janik (2006) to account for the logarithmic perception of sound and also to allow for variability in the time domain. This is possible because the system adopts an algorithm, called dynamic time-warping, that was developed for use on human speech to allow for compression and expansion of the signal duration to maximise the frequency overlap with a reference signal. This technique proved very successful in the study of vocalizations emitted by odontocetes and can be employed to predict the acoustic behaviour of animals when

the exact timing of behavioural events is of no or little interest. Such method was applied to humpback whale calls of the South Pacific for identifying the song structure and draw song comparisons between different regions; the correct classification rate of the ART neural network was on average 94% when compared to a classification carried out by trained observers (Helweg *et al.*, 1998).

Results obtained in these studies improved our knowledge on the characteristics of vocalisations of humpback whales, their usage and function across the World. However, these techniques are tuned to each specific study; whereas, given the importance of comparing the acoustic repertoire of different populations, it is valuable to have a common method that can be successfully applied to classify the vocalisations of all humpback whales. This needs to be flexible enough to allow for regional variations and for changing S/NN. Most importantly, neural networks need to be able to capture the biological meaning of the data, which is sometimes lost with automated classification systems.

Statistical techniques have also been considered and shown to produce good results. In 2007, Rebecca Dunlop (2007) was able to identify around 100 different calls that are commonly produced by these whales in a social context and split them into 6 categories based on their aural and spectrographic characteristics, taking into account factors such as maximum and minimum frequencies of the signal and its duration. The six categories were: (1) low-frequency, (2) mid-frequency harmonic, (3) high-frequency harmonic, (4) amplitude-modulated, (5) broadband noisy and complex, and (6) repetitive sounds (Dunlop *et al.*, 2007a).

One approach to the automated classification of bio-acoustic signals is to adopt a matched filter. Such a method can be implemented in a variety of domains, but has proven to be particularly effective for some cetacean vocalisations when implemented in the spectrogram domain (Strager, 1995). Matched filtering is appropriate for calls that are highly stereo-typed and are observed in conditions where propagation effects can be accounted for. The wide repertoire and variability associated with humpback calls renders this method unattractive. Deecke and Janik (2006) highlighted the importance of using a flexible method which allows one to capture the changes in duration in killer whale (*Orcinus orca*) and bottlenose dolphin (*Tursiops truncatus*) calls: a principle which we believe applies equally to humpback calls.

To cope with signals that have a fixed structure but whose components vary in duration one could extend the power of the matched filter by applying the principles of dynamic time warping (DTW) (Deller *et al.*, 1993; Deecke and Janik, 2006a). DTW has been adopted for the classification of killer whale vocalisations (Brown *et al.*, 2006; Brown and Miller, 2007) and whistles of bottlenose dolphins (Deecke and Janik 2006). However, in a wider context, dynamic time warping has proven less effective than the competing methodology of Hidden Markov models (HMMs) for both modelling human speech (Deller *et al.*, 1993) and for bioacoustics signals (Rickwood and Taylor, 2008; Ren *et al.*, 2009).

2.4 Conclusions

For many decades, researchers have studied the structure of humpback whale songs; however, still little progress has been made in comparing recordings on a global scale primarily because the task of manually classifying song components is extremely time consuming and because comparisons across results of different research groups are complicated by subjectivity in defining the components. Understanding how songs compare on a global scale is extremely important because humpback whales are animals that can travel huge distances even within a single season (and that all the oceans are interconnected. Therefore, communication between humpback whale populations and song learning may occur to a greater extent than what we currently think. To aid researchers in their efforts to analyse huge amounts of data obtained from recordings of marine mammals, automatic detectors and classifiers have been developed in recent years. Some of them are extremely successful at classifying species of marine mammals that emit very stereotyped calls, e.g blue whale calls, but no automatic method has yet been shown to classify the components of humpback whale songs with high levels of accuracy. The aim of the work presented here is to try and fill this gap, starting from a comparison of the efficiency of feature sets that have been used to describe humpback whale songs, according to surveys of published literature, to determine which coefficients are best suited to describe the variety of calls that constitute humpback whale songs.

3. Background to speech processing and implications for bioacoustics signal analysis

Before describing the detection and classification issues arising from previous research and outlining the methods adopted in this thesis, it is important to clarify that the goal is to detect single vocalisations and to classify each one into a sound category with the final intent of automatically describing the whole song sequence, in this case for humpback whale songs. Traditionally, detection and classification methods deal with detecting and classifying marine mammal vocalisations to ascribe them to a particular species to indicate the presence of the species in the area, particularly for monitoring purposes. The detection and classification tasks that are addressed in this thesis are not aimed at distinguishing calls amongst species, given that only humpback whales are recorded in the study site, but deal with the detection of vocalisations emitted by a single individual, recognising the call sequences within songs, and classifying them correctly between songs; in other words, being able to correctly classify song components that are produced by different individuals at different times.

Trained observers can classify humpback whale songs with a high level of accuracy even when they are presented with completely new songs because they can identify the subcomponents of each phrase and compare units across songs. However, as previously discussed, manual classification is extremely time consuming and it is necessary to make the song classification process automatic to be able to conduct large scale comparisons. The driving factor for this thesis is to obtain an automatic classifier that is able to mimic the performance of a trained bioacoustician in terms of classification accuracy, whilst reducing the time effort to a minimum. This is one of the reasons why the methods used for characterising and classifying the humpback whale vocalisations are borrowed from tools that are widely used in speech recognition tasks.

The advantage of using techniques that are commonly employed for speech processing is that a variety of tools have been developed and tested extensively for many years, and consequently toolkits are widely available to the public, reducing the programming effort. On the other hand, the mechanisms for sound production in humpback whales are not completely understood yet, and although dissections of the

animal vocal apparatus showed that they possess structures that are analogous to those identified in the human vocal apparatus, one cannot imply that air flows through the vocal apparatus in exactly the same manner, particularly given that the sound generated needs to be transmitted through a different medium and that breathing in whales is a voluntary action (Reynolds and Rommel, 1999).

Speech models are well suited to describing the mechanisms of human vocal apparatus and human hearing, and through the extensive research that has been carried out on the matter, sentences and words can be characterised and classified with high level of accuracy (90% or more accuracy). The fact that we aim to build a classifier for humpback whale song that mimics the accuracy of a trained human listener justifies the adoption of processing tools that have been developed for human speech. The underlying idea is that the model can be tuned to the way humans perceive whale vocalisations, this does assume that we can classify their songs accurately in a biologically significant way, i.e. keeping into account the way whales perceive and understand the meaning of a song or acoustic signal.

1.1 Speech production and modelling

The mechanics of speech production is well known because numerous experiments have been carried out for decades to understand the pathways of sound production in humans and its development (Deller *et al.*, 1993).

A sound is generated in the vocal tract when air is pushed from the lungs through the glottis in the larynx, generating pressure over the vocal folds which vibrate generating a waveform.

Speech sounds are traditionally divided into voiced, if vocal cords vibrate during their production, and unvoiced speech, if the cords do not vibrate (Deller *et al.*, 1993; Mercado III and Kuh, 1998). Voiced sounds are pulsed and the rate of pulses gives the characteristic pitch of a sound, which is perceived by listeners. These are quasi-periodic excitations that cause the glottis to let air through at a regular rate. On the other hand, unvoiced sounds are aperiodic noise bursts that may originate from turbulence. Other types of excitations are sometimes observed in human speech (e.g. plosives), some of which are peculiar to certain languages (Rabiner and Juang, 1993).

The articulated sounds that form human speech are determined by the length of the vocal tract and the manner in which we position the tongue and mouth (including

lips, palate and jaws). The nasal tract is also involved in the transmission process and it can substantially modify the amplitude of the sound radiated from the mouth (Deller *et al.*, 1993). The degree to which the nasal cavity is coupled to the rest of the vocal tract is controlled through the velum (or soft palate). The muscle fibres sheathed by a mucous membrane that can completely seal off the nasal passage from the mouth (*Figure 3.1*).

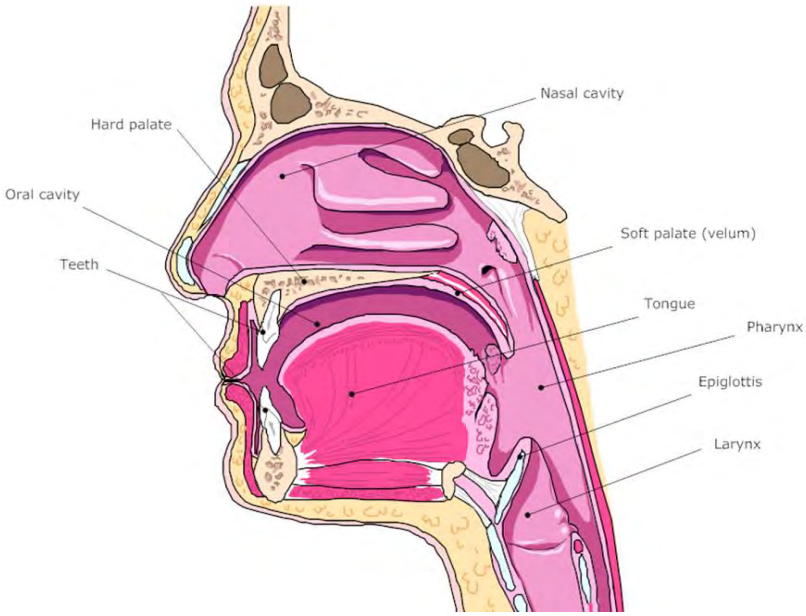


Figure 3.1: Anatomy of the upper vocal apparatus (sagittal plane). The source which are the lungs are not depicted in this diagram (adapted by the author from public template),.

Speech sounds, like all sound waves, can be described both in the time and frequency domains. In the time domain one can observe the amplitude of the signal which gives an indication of the air flow through the vocal tract (Deller *et al.*, 1993; Jurafsky and Martin, 2009). High pressure means that the vocal folds are open and air coming through the lungs flows through them and during regions of low pressure the vocal folds are closed. Thus, when the vocal folds vibrate producing voiced sounds one will observe a sequence of high and low energy peaks. In the frequency domain, this oscillation can be measured to obtain the fundamental frequency of the waveform (*Figure 3.1*). These characteristics can be observed directly from the waveform; however, sound waves are usually quite complex and further analysis is needed to extract their features. The main complication in analysing speech derives

from the fact that its characteristics change through time; therefore, scientists have developed ways to decompose speech into small components to reduce the amount of variability encountered. If sufficiently short segments of speech sounds are chosen, their characteristics will be nearly stationary within a segment; this means that one can decompose the sound into many stationary components, each with its individual characteristics (Rabiner and Shafer, 1978; Deller *et al.*, 1993; Jurafsky and Martin, 2009).

Spectral representations of speech signals were one of the first tools used to analyse them because they allow identifying the frequency peaks that characterise each sound, for instance, vowels can be told apart by identifying their formant frequencies. The Fourier transform of speech signals allows one to create a spectrum of the signal over its entire duration; however, for speech signals, which vary characteristics over time, it is common to describe them through spectrogram representation. Spectrograms allow one to visualise how the energy content of the signal changes over time. This is obtained by calculating the energy of the signal at each frequency point over a user-defined time window and the energy is represented graphically through a colour coding.

The fundamental frequency depends on the rate of vibration of the vocal folds in the larynx and it depends on the size and tension of the vocal folds of the speaker at a given time, and the perceived fundamental frequency is the pitch of that sound (Deller *et al.*, 1993). However, differences in other spectral peaks of speech signals are determined by the cavities in the mouth that act as resonators.

The vocal apparatus depicted in *Figure 3.1* can be schematically represented to highlight the basic mechanics of sound production (*Figure 3.2*); this is known as the source-filter model (Deller *et al.*, 1993; Rabiner and Juang, 1993; Mercado III and Kuh, 1998; Jurafsky and Martin, 2009).

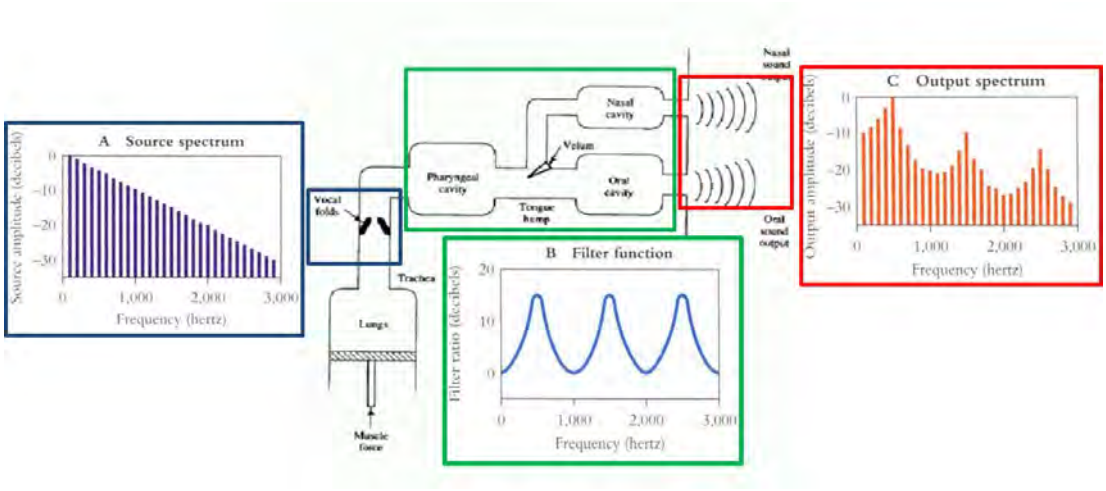


Figure 3.2: Block diagram of human speech (voiced) production model (adapted from Deller et al, 1993), and example of speech sound that is generated passing through the vocal folds (blue box) and further modified by the filter whose frequency response is depicted in the green box to produce the sound output in the red box.

The source-filter model assumes that speech production can be represented through a system with three filters: as air is pushed upwards from the lungs it passes through the larynx where the vocal folds are excited (source), and then through the pharynx which is the first filter. The sound is modified further in the oral and naval cavities or just in the oral cavity if the velum is closed. The cavities act as acoustic resonators and, as such, they enhance certain frequencies and attenuate others depending on the positioning of the structures present in the oral cavity (Rabiner and Shafer, 1978; Deller *et al.*, 1993; Rabiner and Juang, 1993).

In the source-filter model, voiced sounds are usually produced by a pulse train as source, whilst unvoiced sounds are represented by white noise. The sound then is filtered using an all-pole infinite impulse response filter (IIR) (Deller *et al.*, 1993). This model is widely accepted to describe the mechanisms of speech production but it accounts only for the flow of air from the source to one filtering chamber and its propagation out to the environment. In the case of sound production in baleen whales, a different model is required because alternative mechanisms are likely to be involved. Although the specific pathways of sound production and propagation in baleen whales remain to be understood, it is recognised that air is recycled within the vocal tract to allow continuous production of sound underwater as evidenced by the lack of bubbles emitted during sound generation (Reindenberg and Laitman, 2007; Mercado III *et al.*, 2010).

3.2 Parallels between speech and bioacoustics

Speech analysis is concerned both with the mechanisms of sound production and the identification of features that are unique to the speaker for speaker recognition technology. In general, presence or absence of a fundamental frequency indicates the type of sound that is produced by the speaker, giving information on the production mechanism; whilst the length of the vocal tract and the relative amplitude of the harmonics, which are frequency components that are integer multiples of the fundamental frequency, give an insight on the characteristics of the vocal tract of a specific speaker. Therefore, we assume that the spectral features of the signals emitted by large whales will contain information about the production mechanisms and the characteristics of the vocal apparatus. Both sound production mechanisms and speaker recognition in Mysticetes have not received much research attention because scientists are mainly concerned with the influence of propagation on the sounds emitted (Reynolds and Rommel, 1999). However, recently the focus has shifted on how baleen whales produce such powerful sounds underwater to understand the evolution of language through mammalian lineages and to get an insight on the cognition of these animals.

As mentioned in Chapter 1, given the anatomical evidence, it is believed that the main mechanism for sound production in humpback whales is similar to human speech in that the source is the voice box in the larynx where the vocal folds vibrate (i.e. U-folds in humpback whales) (Reindenberg and Laitman, 2007). According to their spectrographic features, humpback whale vocalisations can be classified into broad categories that are comparable to the classification of speech sounds according to the production mechanism involved. Although the terminology used for the description of whale vocalisations is analogous to that used to describe speech sounds, the sounds described by the same term indicate different types of sounds in human speech versus whale sounds. Therefore, below an overview is presented of the terminology used in the literature and in this thesis to describe the vocalisations of humpback whales.

In human speech, examples of voiced sounds are the vowels whose waveform is quasi periodic and one can clearly distinguish the fundamental frequency of the signal which reflects the periodic pulsation of the glottis as air is pushed through the vocal folds (*Figure 3.3*).

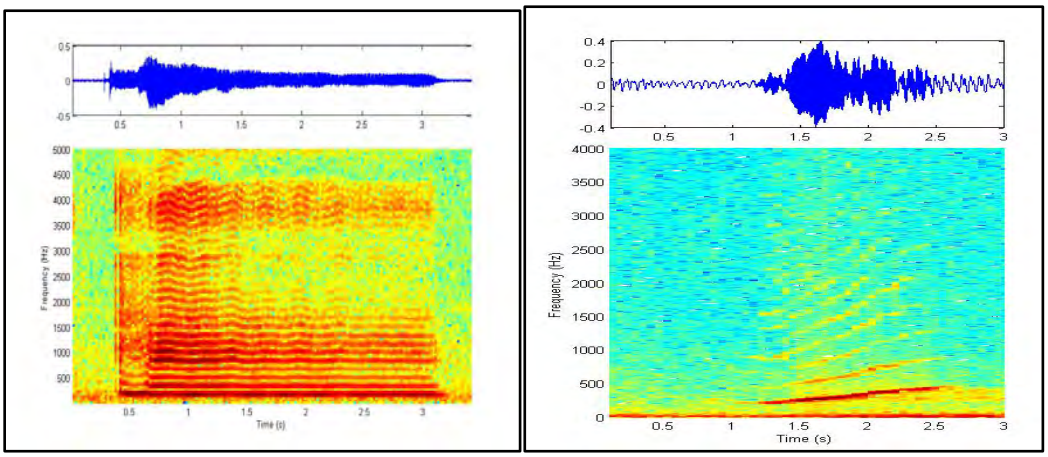


Figure 3.3: Amplitude (top) and spectrogram (bottom) of the vowel /a/ (left) with a Hanning window (frequency resolution 43.06 Hz/bin) and a tonal call (right) of a humpback whale.

In baleen whales, calls that have similar spectrographic features to the ones observed in voiced speech are termed tonal vocalisations as a clear frequency tone can be identified. Tonal calls can be distinguished into further categories based on how frequency changes through time. In some cases frequency remains constant throughout the duration of the call; such sounds in humpback whales are termed moans, i.e. long low frequency calls. Onomatopoeic words are often used to describe whale calls but in this thesis we will try to avoid this terminology because it is very subjective and makes comparisons hard across different parts of the World where onomatopoeic terminology may be different. In other instances, the frequency of a call increases or decreases through time, sometimes very rapidly; calls where this happens are termed upsweeps and downsweeps respectively and overall they are described as frequency modulated (FM) calls (Dunlop *et al.*, 2007c; Dunlop *et al.*, 2008; Mercado III *et al.*, 2010). FM vocalisations are very commonly observed in bat echolocation signals. An additional characteristic of tonal calls is the presence of harmonics that are multiples of the fundamental frequency: some humpback whale vocalisations have very few harmonics (i.e. 2 or 3) and others have many harmonics stretching across the whole frequency spectrum up to 21 kHz.

Humpback whales produce also broadband noise-like sounds where one can observe turbulent flow in the spectrogram, similarly to unvoiced speech (*Figure 3.4*), due to non-linear phenomena.

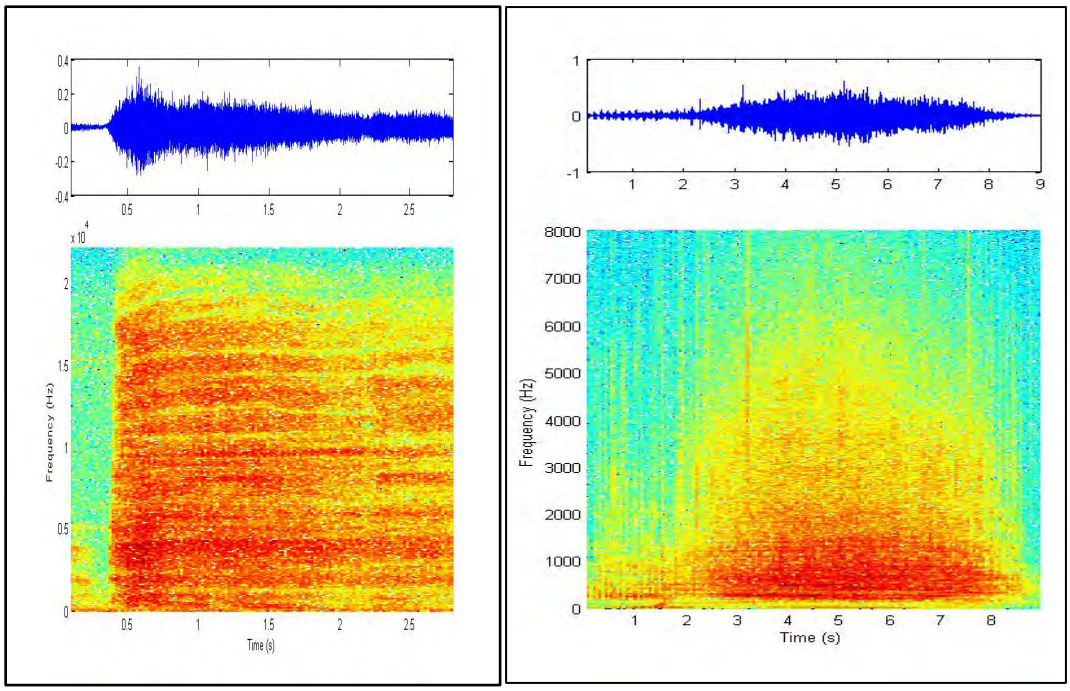


Figure 3.4: Amplitude (top) and spectrogram with Hanning window (frequency resolution 43.06 Hz/bin) bottom) of the unvoiced consonant /s/ (left) and of a "noisy" sound of humpback whale (right).

The last group of vocalisations that are produced by humpback whales are amplitude modulated (AM) calls (*Figure 3.5*). To our knowledge, these types of sounds have no clear analogy with human speech. Although graphically AM and noise-like calls appear quite similar, they sound very different.

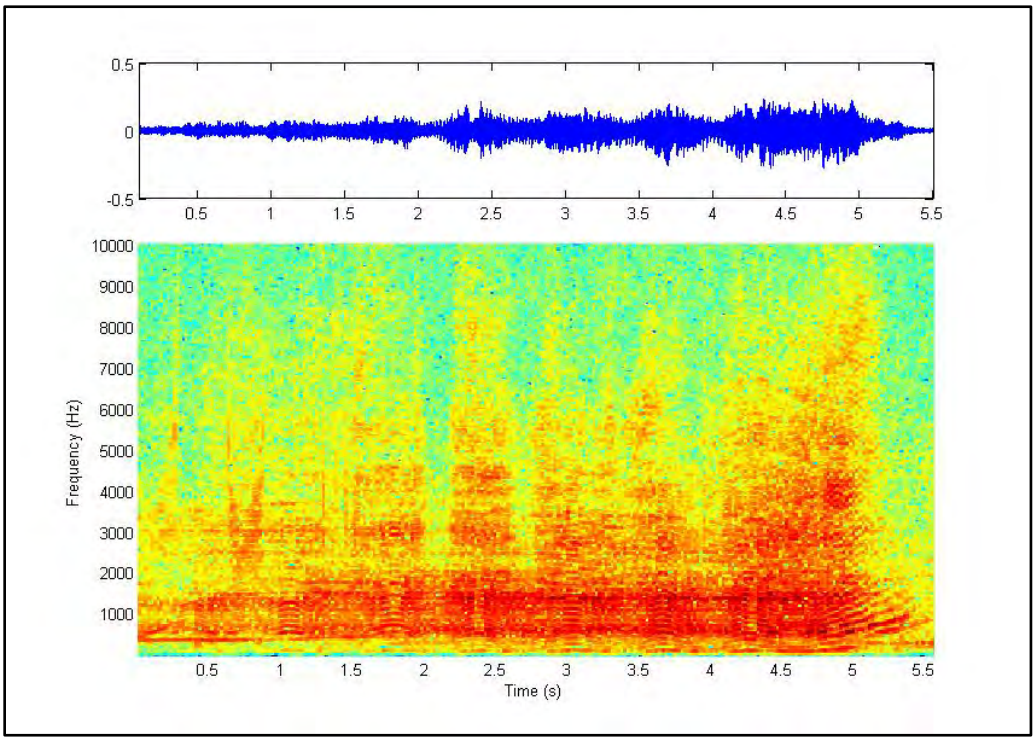


Figure 3.5: Amplitude and spectrogram with Hanning window (frequency resolution 43.06 Hz/bin) of a humpback whale amplitude modulated vocalisation.

All the types of vocalisations discussed here are similar to those used in a detailed study on the social communication of humpback whales by Dunlop *et al.* (2007c), who identified 34 different vocalisation types produced by various groups of humpback whales in Western Australia. The 34 call classes were split into 6 main groups: tonal sounds were split into three subgroups, i.e. low, medium and high frequency harmonic sounds depending on where the fundamental frequency laid in the frequency spectrum, AM sounds, broadband noisy sounds, and repetitive sounds. The broad categories used to describe humpback whale calls in this thesis, as opposed to the Dunlop classification, are summarised in Figure 3.6.

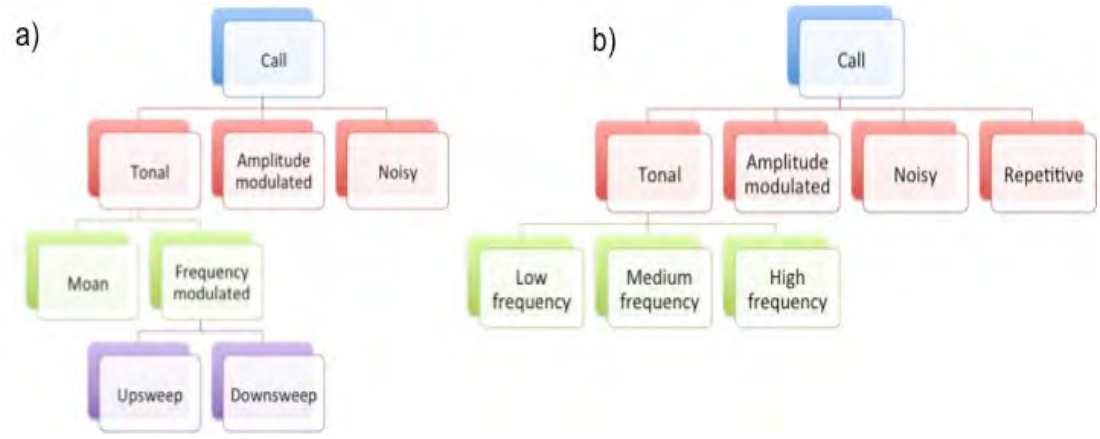


Figure 3.6: diagram showing the hierarchy used to describe calls in this thesis (a) as opposed to the hierarchy used by Dunlop et al. (b)(2007c).

In this thesis, the group of repetitive calls was removed and single calls that would make up a repetitive sequence were analysed individually and grouped under one of the other three categories. Tonal calls were grouped in terms of the trend shape of their fundamental frequency rather than its average value. Whereas, the categories of amplitude-modulated and noisy calls were retained.

3.3 Issues with automatic classification of mammal sounds

In the previous chapter, some of the issues associated with the automatic classification of mammal vocalisations were introduced. Not all marine mammals produce complex sounds but humpback whale calls are extremely variable both at the individual and at the population level. Hence, the clustering algorithm must be flexible to allow for these differences and correctly classify vocalisations that have the same content but vary in length or harmonic components, which depend on the physical characteristics of the animal's vocal apparatus and the environment in which the sound propagates. This problem is exacerbated by the fact that vocalisations may change from year to year and from one population to another (Payne and McVay, 1971; Winn *et al.*, 1981; Helweg, 1996; Helweg *et al.*, 1998; Garland *et al.*, 2011).

Furthermore, detection can be rendered difficult by the presence of several animals, which produce calls in the same area but are not the focus of the analysis. Such vocalisations can be treated as noise and be discarded in the analysis. However, it is usually difficult to identify which sounds were emitted by the focal animal when

analysing underwater recordings since no, or little, information is available about the singer's location relative to the hydrophone.

Apart from problems relating the quality of the recordings, classification issues may arise even in high quality recordings because of the intrinsic characteristics of certain sounds which may not be taken into account in the models used. A problem that has received little attention so far is the occurrence of non-linearities, which are not normally modelled through the source-filter model. Non-linear phenomena in speech production have been known for a while in humans affected by speech pathologies but only recently researchers have started to take them into consideration for studies on mammalian language (Wilden *et al.*, 1998). These phenomena were studied in rhesus monkeys (Fitch *et al.*, 2002), right whales and killer whales (Tyson *et al.*, 2007). The source-filter model for speech production fails to represent non-linearities since it assumes that the source and the filter are independent of each other in the sound generation process.

Non-linearities that have been described in speech are subharmonics, biphonation and determistic chaos (Wilden *et al.*, 1998; Tyack and Miller, 2002; Tyson *et al.*, 2007) and can be often observed in whale vocalisations, although they received little attention.

Subharmonics form when the tension in the two vocal folds is different so that one will observe another spectral component at values that are fractional intervals of the fundamental frequency (Tyson *et al.*, 2007). An example of such subharmonics from humpback whale in our recordings is presented in *Figure 3.7*.

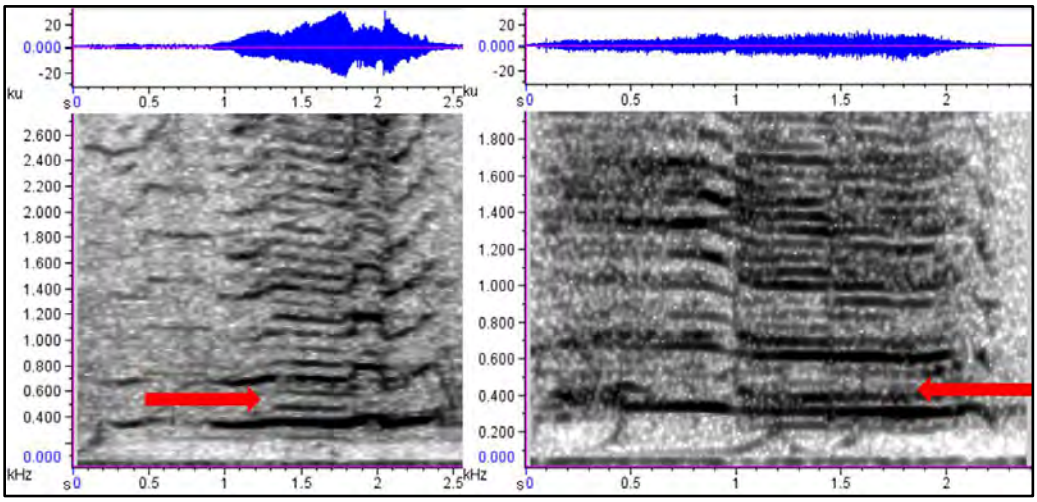


Figure 3.7: Examples of subharmonics, indicated by the red arrows, in a humpback whale song recorded in Madagascar in August 2009.

In addition to subharmonic formation, an extreme case of non-linear phenomenon in sound production is deterministic chaos. This is defined as a period of non-random noise which is produced by desynchronised coupled oscillators. In such cases, the energy is usually distributed across a broad spectrum with some residual periodic energy related to the previous harmonic components (Tyson *et al.*, 2007). An example of chaos during two similar vocalisations recorded for this thesis is shown in *Figure 3.8*. One can notice that during a section of the call a clear fundamental frequency and its harmonics can be identified within a narrow frequency band but in other parts the call becomes noise-like and the energy of the signal is spread over a broader frequency range.

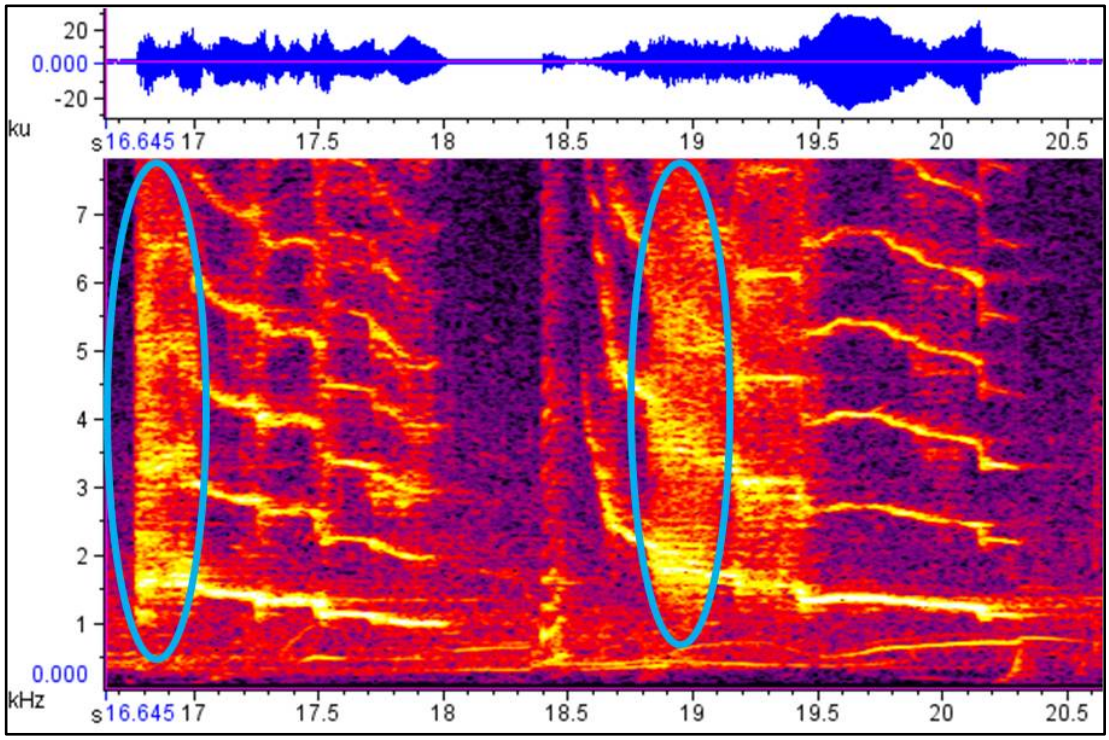


Figure 3.8: Spectrogram of humpback whale vocalisations where deterministic chaos (areas circled in blue) can be observed in the recording of August 2009.

Another feature that presents a challenge in modelling the mechanisms of sound production and their classification is biphonation. Biphonation consists on the production of two independent frequencies simultaneously so that two fundamental frequencies can be identified in the spectrogram (Wilden *et al.*, 1998). This phenomenon results from the weak coupling of the two oscillators in the vocal apparatus or the presence of more than one oscillator. In other words, more than one sound source is needed for biphonation to occur, which is why the source-filter

model is not suitable in this case. This phenomenon has been known for some time to occur in birds, because they have a specialised structure called syrinx whose left and right sides are divided and can generate sounds independently of each other (Suthers, 1990). Amongst the cetacean family, biphonation has been described in bottlenose dolphins (Cranford *et al.*, 1996) and killer whales (Tyson *et al.*, 2007). Both species possess a pair of sound generating structures (called phonic lips) within their nasal cavity providing two simultaneous sound generation mechanisms. Although, only one sound source has been identified in baleen whales, studies of their vocalisations suggest that in some species biphonation may occur (Tervo *et al.*, 2011) because unrelated sound notes have been observed which were produced simultaneously in minke whales, North Atlantic right whales and bowhead whales (Gedamke *et al.*, 2001; Tyson *et al.*, 2007; Tervo *et al.*, 2011). Biphonation occurrences in the latter species has been proven in recent evidence that showed that both sounds were generated by the same individual, although the mechanisms through which this occurs are still unclear (Tervo *et al.*, 2011). Given the varied repertoire of humpback whale songs and personal observations of the sounds emitted, we cannot exclude that this phenomenon occurs in this species too.

Non-linearities introduce a greater level of complexity for the automatic classification of vocalisations in that if a signal is present in the various forms described above then the feature sets used to characterise the sound might not capture the similarities between them so that the number of classes will increase. Such phenomena in the calls are thought to reduce listener habituation and to play an important role in animal communication; they may convey cues about the fitness of the signaller, animal size, and/or function as alarm calls (Fitch *et al.*, 2002). Therefore, it is unsurprising to find non-linear features in the calls of humpback whales songs that are involved in mate attraction and territorial marking given that the signaller may want to inform the receiver about his emotional state.

In addition to non-linearities, frequency jumps are observed within a vocalisation, which are the result of an interaction between the vocal folds' vibrations and the vocal tract's resonant properties, and they pose a problem in terms of the definition of a sound unit which is used as the building block for all analyses carried out so far. Specifically, frequency jumps appear to occur when the pulse rate of the signal gets

close to the resonant frequency of the vocal tract (Titze, 2008). Examples of frequency jumps in humpback whale songs are given below (*Figure 3.9*).

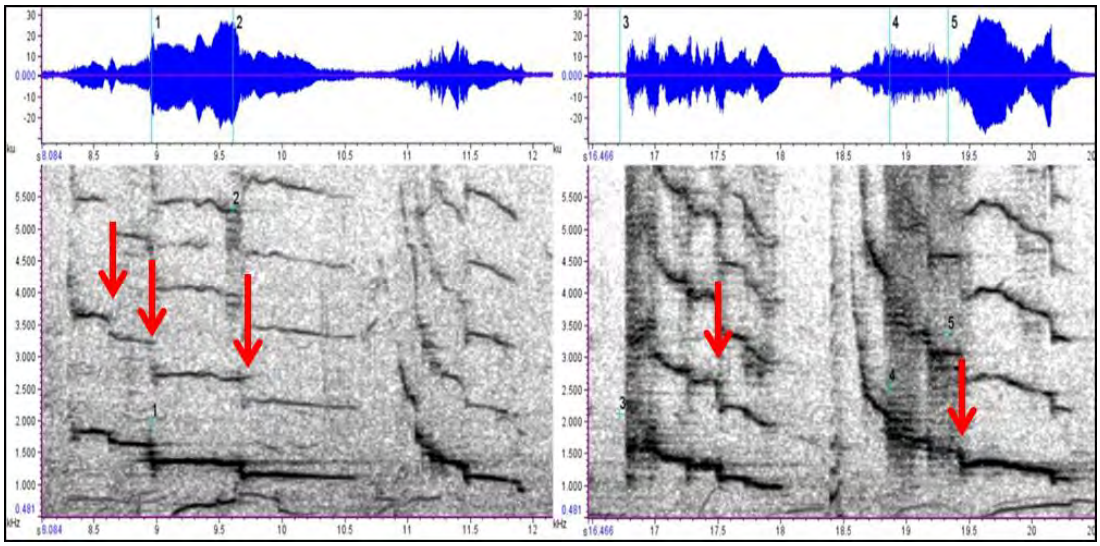


Figure 3.9: Examples of frequency jumps in some units of humpback whale song recorded on the 12th of August 2009. The numbers in the top panel of the figure showing the amplitude of the signal highlight instances in which the jumps occur. Such jumps are easily seen in the spectrogram representation of the signal (bottom panel, red arrows).

The presence of frequency jumps represents a problem at the detection stage because the energy of the signal may drop below the set threshold around the time of the frequency jump, leading to the call getting segmented into several smaller blocks and during the classification stage because they introduce more variation to the structure of the calls which may be classified into different groups of vocalisations, particularly when the length of the signal is affected by these frequency jumps. Presence of frequency jumps highlighted the need for a classification algorithm able to cope with such variability within the calls of humpback whale songs.

For this reason, we propose the definition of subunits as the basic block to classify humpback whale calls. In general, as discussed in the next section, subunits will have shorter duration than sound units and they should be encountered in various parts of the World and in different year, although they might be associated with each other in different combinations to allow for variation in the song sessions.

In terms of signal processing, subunits should be easily detectable and less difficult to characterise compared to song units.

3.4 Subunit definition

In this work, subunits are defined for the first time as the shortest segments of a vocalisation that can be encountered on their own or associated with one or more subunits to form a unit; subunits do not vary their characteristics significantly throughout their duration. Hence, subunits might correspond to continuous sounds between two silences but not necessarily (Pace *et al.*, 2010).

A few examples of subunits are given below to improve the understanding of this novel concept. In the first instance, we examine a fast frequency upsweep that is regularly encountered in our recordings in all datasets, and that was identified in previous analyses of humpbacks' vocalisations in other areas of the world not only as part of the song repertoire but also in a social context on the feeding grounds (Dunlop *et al.*, 2007c). This is usually referred to by the literature as 'wop', based on its aural characteristcs.

In the recording analysed for this study, the 'wop' was repeated several times on its own or associated with other vocalisations without interruption of a silence (*Figure 3.10*).

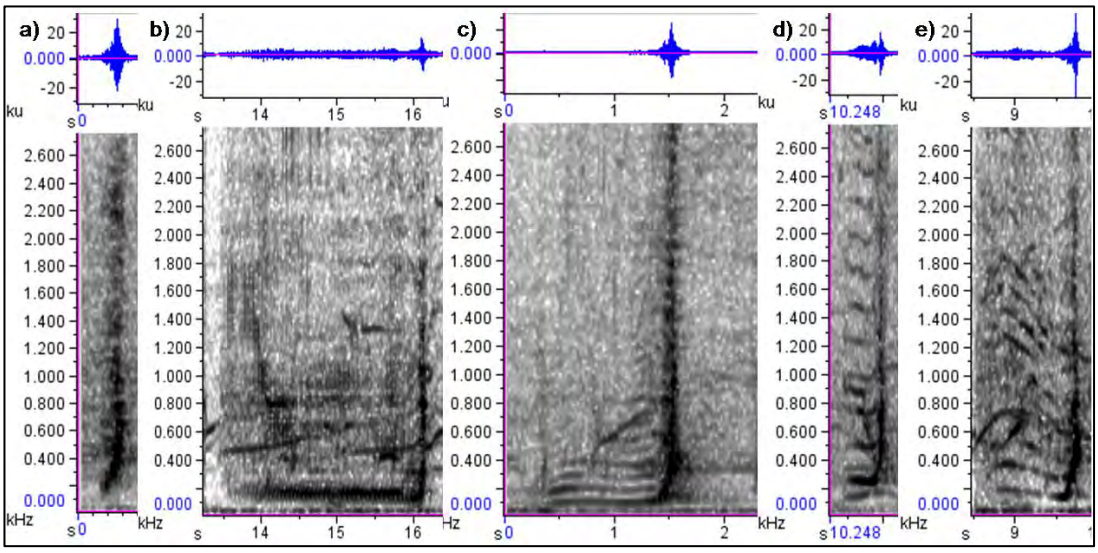


Figure 3.10: Spectrogram of a series of fast upsweeps and the subunits that were found in association with it (frequency,requency rsolution 86.13 Hz/bin, hanning window, 75% overlap). The basic subunit is presented in a); the other graphs represent b) an unvoiced-type call which ends with the sweep, c) a harmonic subunit which terminates with the fast sweep, d) a pulse at a slightly higher frequency that preceeds the sweep, and e) a harmonic call with decending envelope linked to a sweep.

Interestingly, these particular sweeps are encountered very often within a song session in all singers, but they always occur either on their own or right after other subunits (together forming a unit) and never at the start of a sound unit. This suggests there may be physical constraints in the way they are produced.

A second instance of subunits is represented by the association of a subunit with fundamental frequency centred around 900 Hz with other two subunits that were observed in the August 2009 recording represented in *Figure 3.11*.

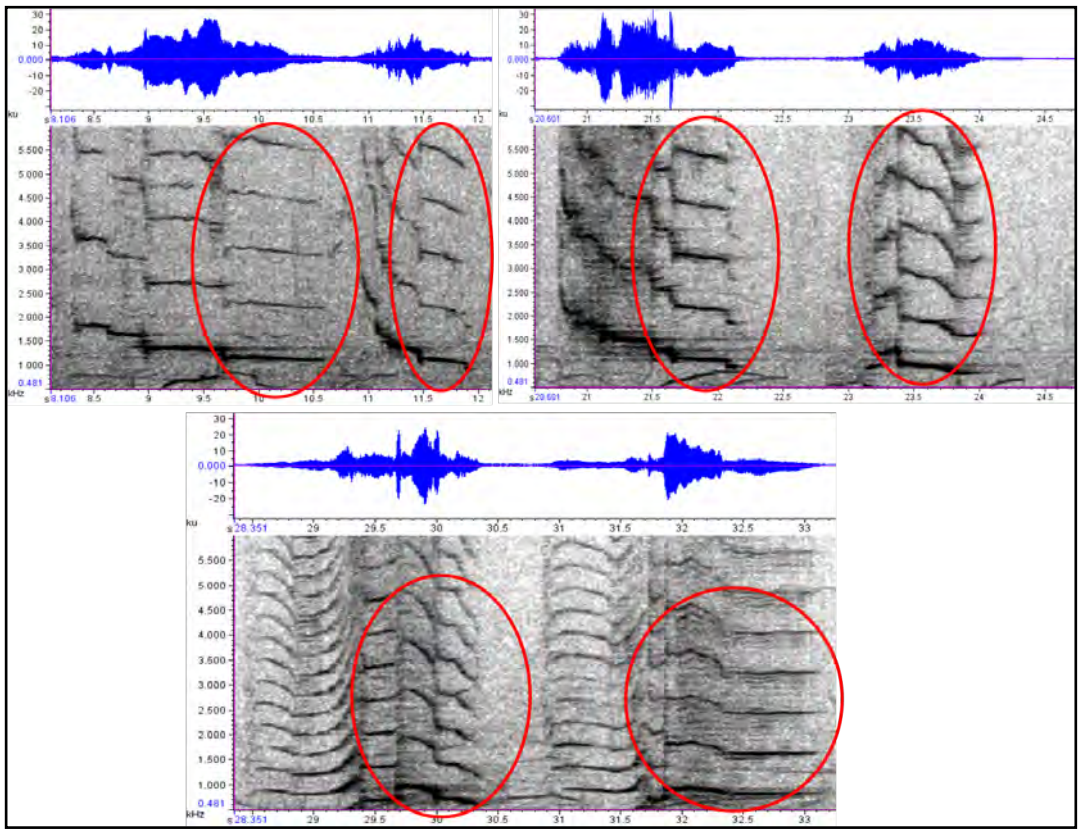


Figure 3.11: The subunit circled in the top right hand side of the figure was encountered on its own or associated with other two subunit types. Hanning window 2048, 512 FFT size and 75% overlap.

A similar concept of subunits was recently introduced for the classification of killer whale (*Orcinus orca*) calls (Shapiro *et al.*, 2011) even though compound calls in this species were first highlighted by Strager (1995). The idea is that compound calls can be split into smaller blocks, termed subunits, where there are marked frequency shifts. Shapiro *et al.* (2011) proposed to automatically segment these call using a pitch tracking algorithm.

In bird song, the need of a flexible approach to classify songs has been recognised for a long time. The extensive research that has been conducted on bird song, particularly on captive zebra finches, showed that songs present a lot of variation within their structure and that units (which are often termed as notes in the context of bird song) can often be subdivided into smaller components called syllables (Williams and Staples, 1992), because of the similarity of this concept with that of syllables encountered in human language. The same two (or more) notes can be juxtaposed in different ways to form different syllables, which are equivalent to units in humpback whale songs. Young birds learn syllables from their adult tutors by copying them and eventually create their own songs which are sequences of the same syllables learnt from the tutors arranged in a particular order and with their own rhythm. Once the song crystallises it will then be repeated by the bird for all its adult life in the same format (

Figure 3.12).

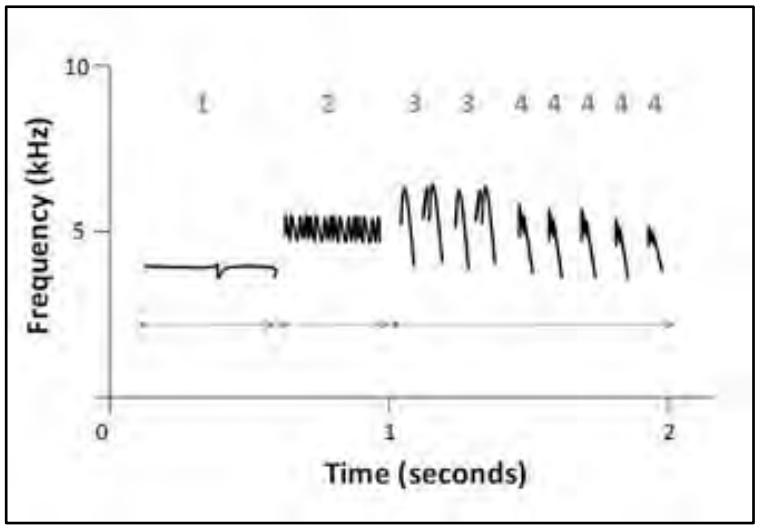


Figure 3.12: Schematic sound spectrogram of a white-crowned sparrow (Zonotrichia leucophrys) song. Arrows indicate phrase (or motif) and numbers indicate syllables which are made up of notes (or elements), the simplest unit of song (Wada, 2010).

Unlike bird song, humpback whale songs evolve continuously within a season and from a season to the next with the effect that within a few years, the song sequences sang by a population will be completely novel, as described in the previous chapter. This represents a major difference and poses a problem in terms of processing large

scale and long term comparisons of humpback whale songs because the vocabulary of units needed to account for all the song compositions will be extremely large.

For this reason, the idea of constructing a dictionary based on subunits is proposed here, because if syllables can be broken down in smaller entities that form all the possible combinations of units that will be present in the song sequence, then one will greatly reduce the vocabulary and, therefore, the computational load and training data needed to obtain good performance.

3.5 Overview of classification methods used in this thesis

This study proposes a new approach for song classification of humpback whales using Hidden Markov Models (HMMs). The power of HMMs derives from their ability to model non-stationary random processes, specifically, they are particularly appropriate when modelling signals whose durations are stochastic. HMMs have become the basis of most (if not all) modern algorithms for the classification of human speech (speech recognition) (Deller *et al.*, 1993). This central role in speech recognition (and their wider use in the field of speech analysis) has meant that considerable research effort has been dedicated to the study of HMMs, one consequence of which is a highly developed, and widely available, tool-set. This makes them attractive tools for application in a wide range of fields, including bioacoustics (Brown and Smaragdis, 2009; Ren *et al.*, 2009).

A HMM is a doubly stochastic process, which comprises a set of (unobservable) states: associated with each of the states is a random process that generates the observed variables (or measurements). For the sake of tractability, the processes within a state are commonly modelled as being stationary. The states are ordered and visited according to a random Markov process (Rabiner, 1989; Rabiner and Juang, 1993). Further, it is normal practice to constrain the transitions between states to only occur in single direction (such models are commonly referred to as left-right models). Many bioacoustic signals are non-stationary and of uncertain length (Ren *et al.*, 2009), making them well-suited to analysis with HMMs.

HMMs were employed in bioacoustics for the unsupervised classification of humpback whale calls (Rickwood and Taylor, 2008); that study demonstrated that these models are robust to varying levels of SNR. However, some level of supervision and previous information about the target signal was needed to reduce the occurrence of false positives, as 9 out of the 19 sound unit classes identified by

the algorithm were found to be related to noise. The methodology applied in the paper includes an initial detection stage, with the objective of identifying data segments that contain vocalisations, removing data segments that are dominated by noise.

No other work is known to date where HMMs have been used to classify humpback whale calls, but they have been employed to study other animal vocalisations (Brown and Smaragdis, 2009; Ren *et al.*, 2009). Specifically, the calls of Asian elephants were analysed in the study by Ren *et al.* (2009) by using cepstral feature sets and a network composed of one HMM for each known call type, a similar method that is used in this study for humpback whale calls. Good results were obtained with an overall classification accuracy of 85% for high SNR data and 60% when noisy and overlapping calls were included in the analysis. Furthermore, the authors applied the same technique for the classification of syllable, song variation and song type of a species of passerine birds with excellent results. The performance of HMM and DTW was previously compared for classifying bird song of zebra finches and indigo bunting (Kogan and Morgan, 1998). The study showed that HMMs were easier to implement with recordings of different quality and where song units changed considerably within songs, and gave more accurate classification results. This suggests that HMMs are a good candidate for our task, particularly in light of the fact that the repertoire of humpback whales presents a wide variety of calls that are similar to the vocalizations emitted by elephants, and fast up- and down-sweeps like the ones observed in bird songs.

Hidden Markov Models were also applied for recognising killer whale calls and, in this instance, compared to Gaussian Mixture Models (GMMs) (Brown and Smaragdis, 2009). Killer whale calls present many similarities to humpback whales', for instance both species produced pulsed calls and bursts and the frequency content of a single call may change considerably through time (Shapiro *et al.*, 2011). Both HMMs and GMMs showed high classification performance of the killer whale calls, with HMMs correctly classifying them in up to 95% of the cases.

In this thesis, the performance of the classification algorithms is based on HMMs and compares the effectiveness of two different structural models for humpback whale songs: one based on the concept of a unit, as defined by Payne and McVay (1971), the alternative model is based on subunits (Pace *et al.*, 2010). Hidden

Markov Modelling and the methods used in this thesis will be presented in greater detail in the next chapters.

4. Data collection and preparation

4.1 Data collection

The data were collected by the author in the Sainte Marie Island Channel by the author; the Channel is located between the Island of Sainte Marie and the North East Coast of Madagascar (*Figure 4.1*).

Figure 4.1: The Island of Ste Marie is located on the North East of Madagascar (red-circled area) and the area surveyed for humpback whale singers is highlighted in red shadowing in the enlarged area (map created using GoogleEarth).

Whales are present in this area during the months of the Southern Winter, i.e. June to October; they come from Antarctica to breed. It is usual to observe solitary animals at the beginning of the season, then active groups that are composed of one female escorted by several males (usually 5 or 6) are seen and, towards the end of the season, mother and calves with or without a male escort can be observed. Singers are found throughout the Winter although they seem to be fewer towards the end of the breeding season; for this reason the field work took place between the end of July and the end of August when singers are still abundant. Although more singers may be found in June since they are meant to attract females to the area and set their

territory, during this period the weather is unfavourable for recordings as the Winter corresponds to the rainy season on this Island; it is only after mid August that the weather starts to improve. In addition, if too many singers are present in close proximity to each other, it is difficult to record the song of an individual singer, since the songs of multiple singers overlap with each other.

The Ste Marie Channel was surveyed between the coral reef in the South of the Island and the Northern part up to the submarine canyon in front of Coco Bay – i.e. where Madagascar and Ste Marie Island are the closest (*Figure 4.1*).

A total of 18 days were spent at sea and 21 hours of recordings were stored. The recordings were taken from a 4 meter long boat, using a COLMAR Italia GP0280 hydrophone (omni-directional, [5 Hz, 90 kHz], sensitivity -170dB re1V/µPa) connected to its amplifier and a HD-P2 TASCAM recorder. The sampling frequency chosen was 44.1 kHz as the harmonics of the vocalisations of humpback whales can reach frequencies up to 20 kHz. The locations of the boat at the start and end of each recording was noted using a Garmin GPS device so that the drift of the boat during the time of the recording could be estimated (*Figure 4.2*).

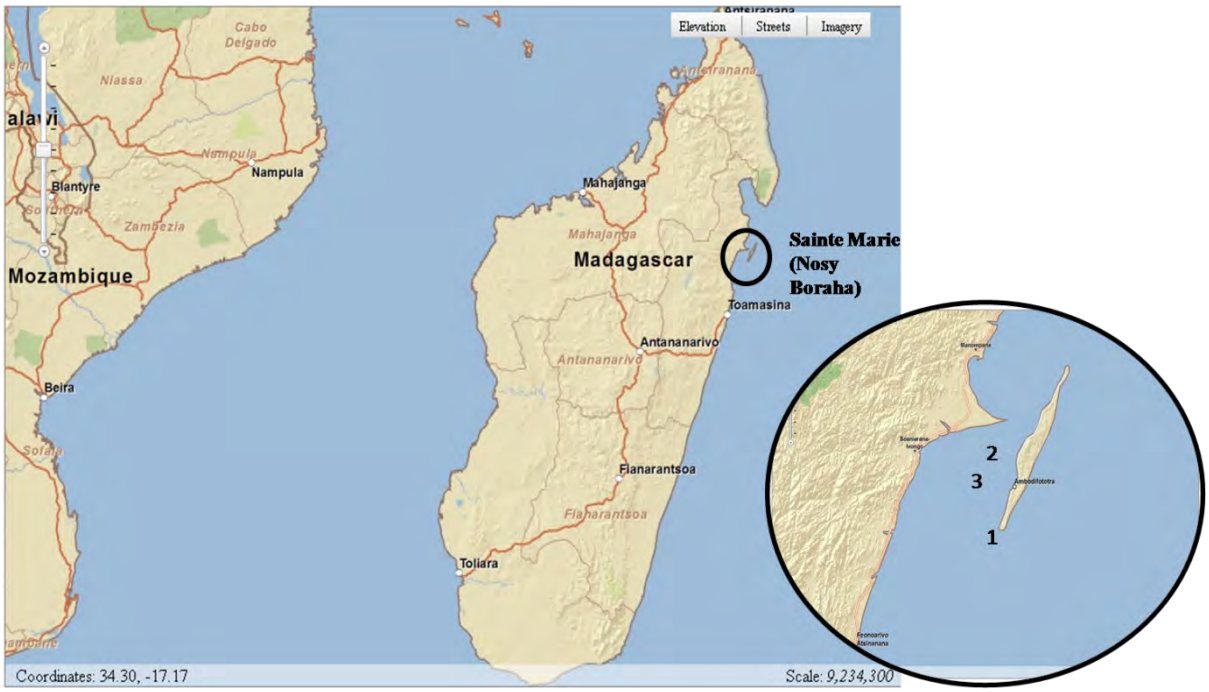


Figure 4.2: Maps showing the location of Sainte Marie Island and a zoomed view of the Ste Marie Channel where the songs were recorded. The recordings presented in this paper were made at the approximate locations depicted by the numbers in the picture; specifically, 1) is the location where the song was recorded in 2007, this location is close to the coral reef at the Southern tip of Ste Marie, 2) shows the site of the 2008 recording, and 3) indicates the area where both songs were recorded in 2009. See Table 2 for further recording details.. Map created using GoogleEarth.

Songs were recorded with variable sea states and weather conditions; however, the ones chosen for the analysis had a high signal-to-noise ratio resulting from the favourable weather conditions, i.e. calm sea and sun, which allowed the boat to be close to the singer for the duration of the recording – the boat was approximately 100 meters from the singer, although the depth of the singer and its relative position to the hydrophone are unknown. Other singers are audible in the recording; nevertheless, the level of their calls was inconspicuous compared to the level of the calls emitted by the focal animal.

Data collection was one of the biggest challenges for this project because the field base was in a very remote location where limited technology was available and because weather conditions were often unfavourable for recording, meaning that most of the time was spent waiting for the right weather window to maximise the quality of the acoustic recordings. The fact that limited technology was available

meant that whenever some issue was discovered with the equipment, "creative" solutions had to be implemented to make sure that at least some songs could be recorded during that season. For instance, after initial tests conducted in the swimming pool, the author found out that the hydrophone had been damaged when it was stored during the summer season and a section of the hydrophone had to be cut out and all the cables re-soldered to the main plug. At the beginning of each season, some problems were encountered possibly because the recording equipment was not stored properly when used by unskilled volunteers and on the spot solutions had to be found. This meant that, despite the fact that the author spent several days on the field, very few recordings with high quality songs were obtained each year.

4.2 Sound detection

An algorithm was developed by the author for the automatic detection of the sound units present in a recording of humpback whale singing (Appendix I).

The algorithm was based on an energy detector with a double threshold, i.e. threshold of start (TS) of the vocalisation and threshold of end (TE) of the vocalisation. The energy of the signal was calculated applying a sliding window of 10 ms. Once the energy within the window exceeded the TS, then all the subsequent samples of the signal were considered a whale vocalisation until the energy fell below the threshold of end.

The TS manually selected was quite high given the high SNR of the recording to ensure that the calls detected by the algorithm were those emitted by the one singer in the proximity of the boat and assuming that the singer emitted calls at an approximately constant source level; whereas, the TE was lower to allow one to capture the majority of the characteristics of the vocalisation (*Figure 4.3*).

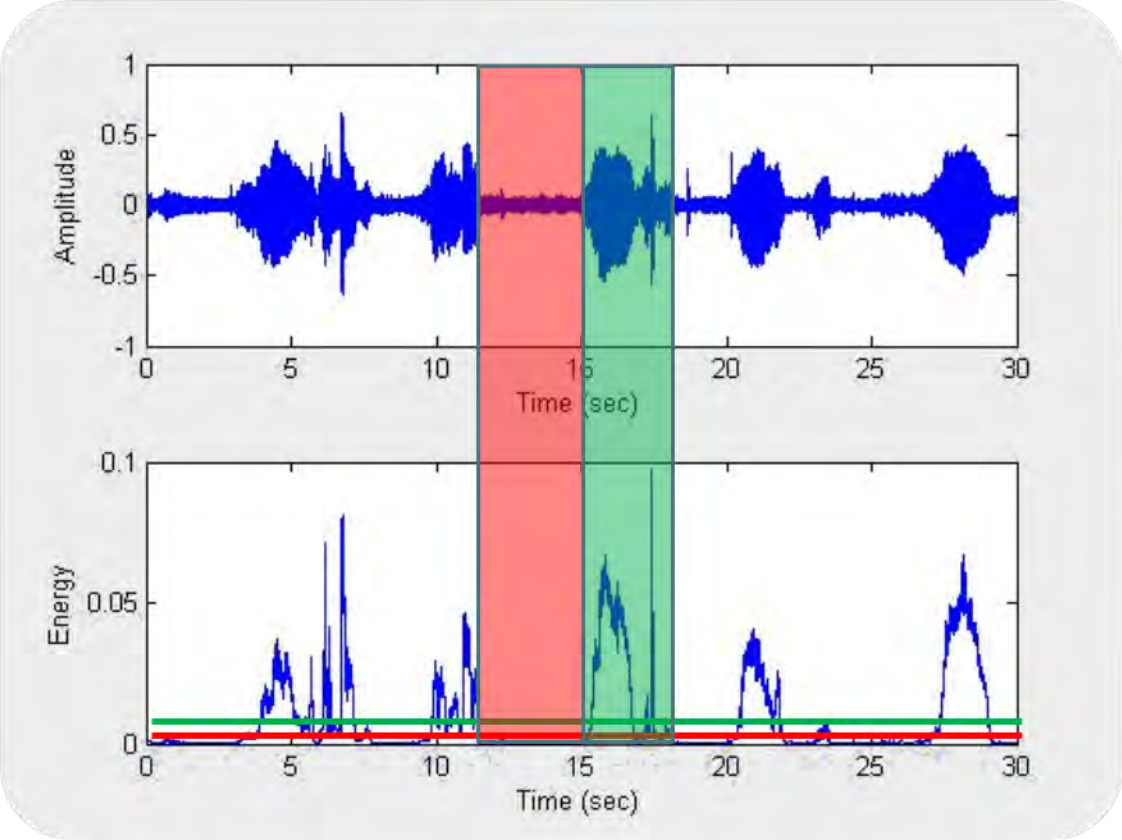


Figure 4.3: Song segment of 30 seconds. The top graph shows the amplitude of the signal (normalised) and the bottom graph is the energy of the signal. The threshold of start (green horizontal line) and the threshold of end (red horizontal line) determine the start and end of a vocalisation (green shaded area) and consequently the silence (red shaded area) in between two calls.

The algorithm detected all the units present in the recording, confirming the validity of our assumption that within a theme the source levels would not change significantly. The validity of the automatic detector was then checked against a manual segmentation by one observer of the units present in the recording. The start and end of most calls was identified accurately by the detector, suggesting that the majority of the calls started off more loudly than they ended (Figure 4.4).

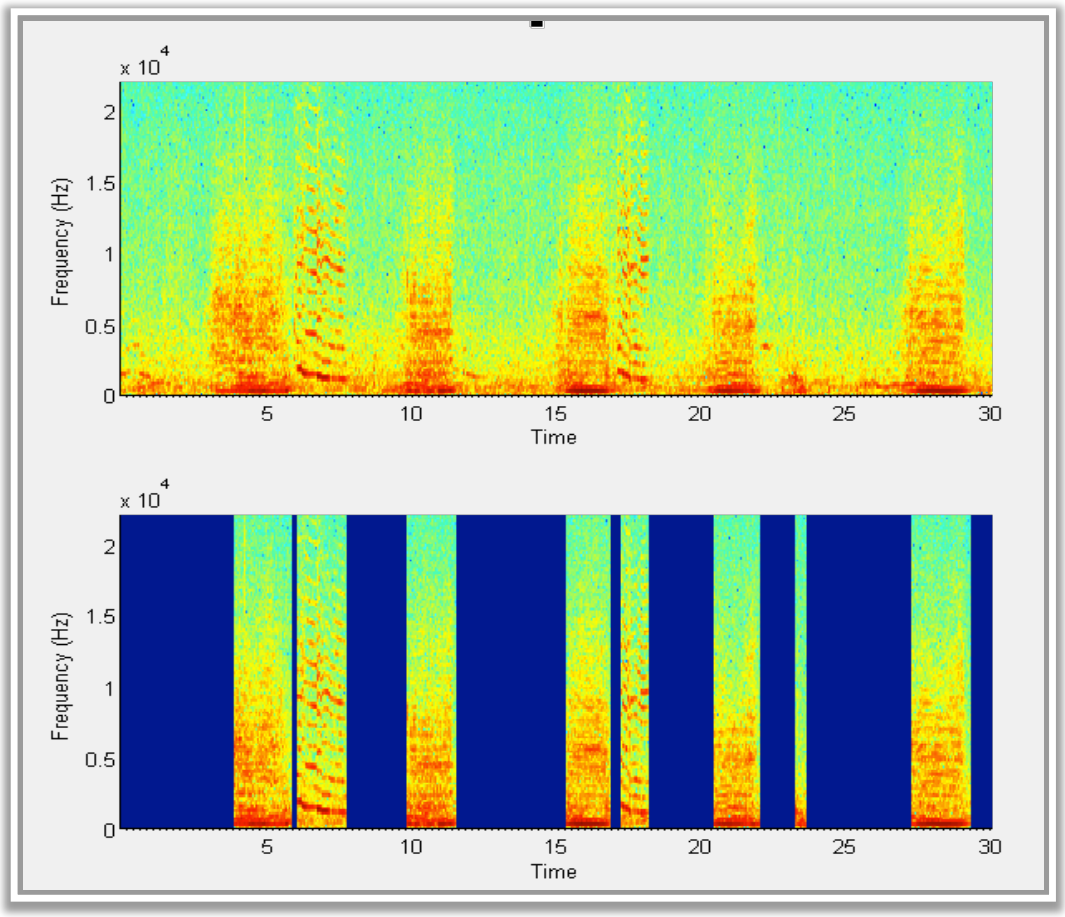


Figure 4.4: Spectrogram (FFT size 256, overlap 50%) of a 30 seconds segment of the song (top) and spectrogram of the units obtained using the automatic detector (bottom). The blue areas represent silences.

Fewer than 10% of the units present in the recording were not detected or only partially detected – the latter meaning that the segment of call obtained was so small that it would impede its correct identification.

Recording labellabel	Date	GPS coordinates	Total number		Duration (min)
			units	subunits	
Mada07Mada07	9th August, 2007	17° 7' 15.99 "S 49° 44' 39.99 "E	141	152	11
Mada08Mada08	16th August, 2008	16° 59' 06.83"S 49° 46' 26.43 "E	227	266	12
Mada09aMada09a	2nd August,2009	16°58'41.33"S 49°45'07.68"E	92	98	4
Mada09bMada09b	12th August, 2009	16°59'14.73"S 49°45'56.88"E	181	219	17

Table 4.1: Details of the recordings from Madagascar analysed in this thesis, their duration and the total number of units and subunits tested.

The units detected automatically were also manually checked by the author to ensure that they belonged to the singer of interest and the start and end points were adjusted to ensure that the full vocalisation was included in further analysis to maximise the performance of the classifiers.

The units obtained from the manual detection of the trained listener were subdivided into their smaller components, i.e. subunits, when necessary to form a catalogue of subunits. The manual detection of units was carried out by visual and aural inspection of the spectrograms of the song using the software Adobe Audition (licenced by the University of Southampton). Although the process was time-consuming, units were detected with a high degree of confidence because the structure of humpback whale songs is very repetitive and it was fairly easy to recognise calls that were emitted by the same singer, by consideration of the sound level, the duration of the silent intervals and the sequence of the song components, such as phrases. The spectrogram in *Figure 4.5* shows an example of a song segment that could be classified with 100% confidence and that represented the gold standard of recording.

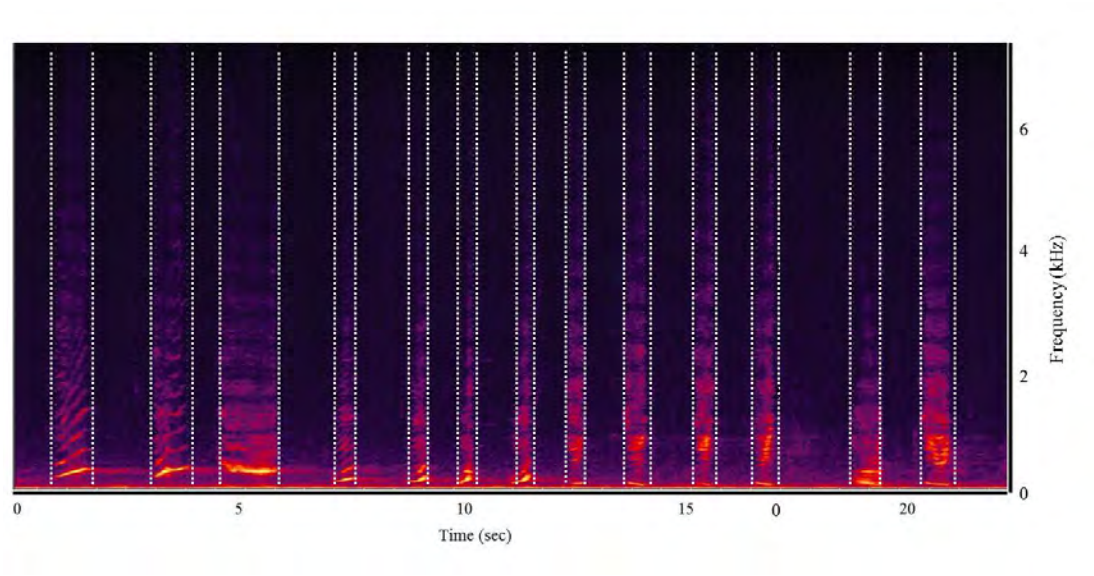


Figure 4.5: Spectrogram (Hanning window, frequency resolution 21.5 Hz/bin) of a song segment with high SNR and its manual classification (white vertical dotted lines). The high S/N is apparent from the fact that there is a huge contrast between the dark background and the yellow fundamental frequency of the signals, meaning that when song was not heard, the background noise was extremely quiet.

Manual classification of the songs was first carried out on the best recordings to ensure that the song structure and all the calls present could be classified correctly with a high degree of confidence. Therefore, prior to initiating the segmentation and classification process, all the songs recorded during a year were subjectively scored according to the signal to noise ratio and ranked, taking into consideration the contrast between signal and noise, the duration of the recordings, the type of noise present, and the overall quality of the song. In some cases, only parts of long recordings were used in the analysis when the singer was in the proximity of the boat since at other times the signal deteriorated. Several possible causes of signal deterioration could be identified: the boat could drift away from the singer due to strong winds and currents, the singer could swim in the opposite direction to the boat or position himself at a different depth, a whale watching boat could approach adding to the background noise, or additional singers could join the focal animal to produce a chorus, resulting in overlapping calls.

If multiple singers, i.e. more than 3, were heard that emitted vocalisations at similar levels, the recording was discarded because too many overlapping signals made the song sequence unintelligible. If such recordings were included, not only would they

have issues in the automatic detection and classification, but also they would have not been confidently detected by the trained listener (*Figure 4.6*).

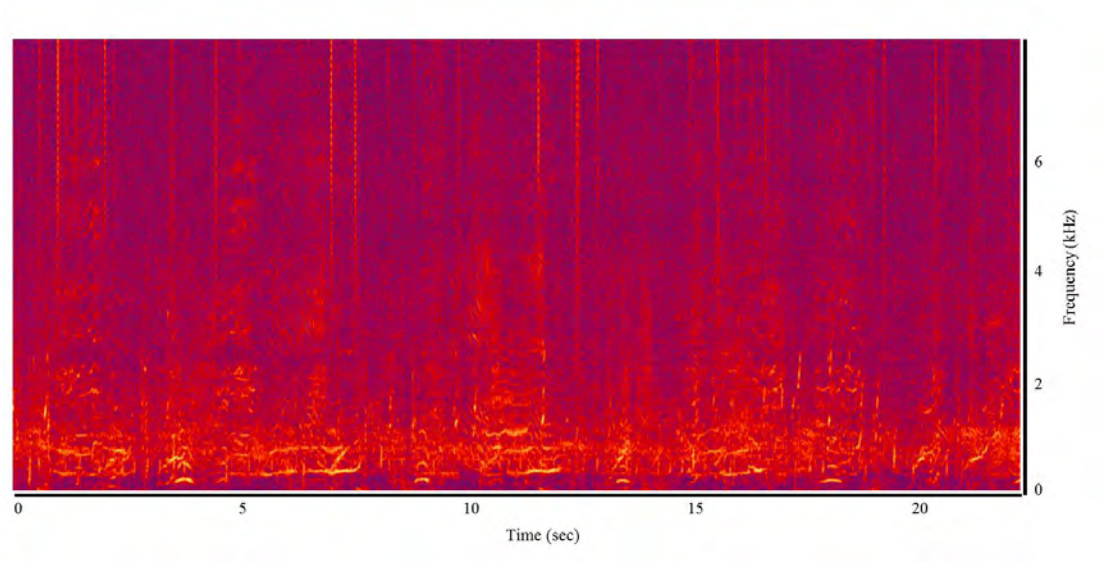


Figure 4.6: Example of song that was discarded because too many overlapping calls made it impossible to identify the full song sequence of a single singer.

All the units and subunits present in the selected recordings were manually classified before the automatic recognition was carried out, one recording at a time. The same detection process was executed for all recordings analysed in this thesis. Automatic detection of the subunits was not developed prior to the classification task because it was embedded in the HMM model, as will be explained in the following chapters. A detailed description of the manual classification is given in the next section.

4.3 Manual classification

4.3.1 Background

As mentioned previously, scientists are trying to reduce the manual classification effort to a minimum because the process is extremely time consuming and it does not allow comparing huge amounts of acoustic data across populations and from year to year quickly. In addition, manual classification is regarded as being subjective and therefore difficult to repeat. However, all the research aimed at developing automatic detection and/or classification algorithms requires one to carry out a manual detection/classification task as baseline to compare the outcome of the automatic algorithm and assess its performance.

The standard procedure is that this manual classification task is carried out by trained observers who, totally independently of each other, group the signals into a number of sound classes that is an unknown at the start of the classification process, totally independently of each other. The outcome of these independent classifications is then compared to reach a consensus. Nevertheless, in most cases, this process is carried out by a single trained observer, or when possible two, because it is extremely time consuming (Mercado III *et al.*, 2003; Suzuki *et al.*, 2006; Seekings and Potter, 2008; Shapiro *et al.*, 2011). Although the methodology used for the manual classification task is often lacking detail in the published literature, a common denominator amongst all the papers is that the classification is carried out combining visual inspection of the spectrograms of the signals with listening to capture the aural features too: it is sometimes the case that two sounds look very similar in the spectrogram but may sound very different. Again, as mentioned in the previous chapter, we are using our human perception of sound to shape the classification algorithms of bioacoustics signals. There are several studies available in peer reviewed journals where there is no mention of how many observers carried out the manual classification used to corroborate the performance of the automatic classifier or describe the repertoire of the species of interest (Helweg, 1996; Helweg *et al.*, 1998; Miller and Bain, 2000; Arraut and Vielliard, 2004; Dunlop *et al.*, 2007a). It appears to be generally accepted that acoustic and visual inspection of bioacoustic signals performed by one or two trained observers is sufficient to obtain a highly accurate classification. This idea was tested on killer whale vocalisations in two studies, where the authors conducted experiments involving trained and untrained observers to ensure that the manual classification task was as objective and accurate as possible (Deecke *et al.*, 1999; Yurk *et al.*, 2002).

Deecke *et al.* (1999) conducted a similarity test to check the performance of their manual classification and then compare this with their neural network approach, and they noticed that the similarity scores assigned by human subjects with no previous training matched the similarity scores obtained for the same vocalisations classified using neural networks. The performance of classification was more accurate when the subject had musical training, which demonstrates that training plays an important role in the manual classification task. In any case, they proved that similarities and differences in call types could be identified by human listeners with no previous

training in categorising calls. The study conducted by Yurk *et al.* (2002) was more detailed and comprehensive than the previous one because they tested a larger number of observers, 17 of which had never been exposed to killer whale calls before and 7 of which had had previous experience in classifying killer whale calls or at least cetacean sounds. Again, a similarity test was performed and the subjects were presented with three calls at a time and had to find the most similar call to the one under examination. Inexperienced listeners agreed in 71% of the cases, whilst experienced observers were in agreement 88% of the time (Yurk *et al.*, 2002). Specifically, the latter group had 100% agreement when presented with 8 out of the 12 call types tested, and had a larger disagreement margin in the remaining 4 call types. This demonstrates that the characteristics of certain vocalisations may be more difficult to discern for human listeners. Because killer whale calls are not dissimilar to humpback whale calls in their general characteristics, one can assume that human subjects will perform similarly when having to classify their vocalisations.

Furthermore, the manual classification performance is bound to improve when previous information about the repertoire is taken into consideration. For humpback whales this is very important because we have extensive information about their song structure. Indeed, the analysis of humpback whale songs is mostly concerned with comparing songs across different populations and to this extent researchers usually assess the similarities based on the phrase or theme sequence (Cerchio *et al.*, 2001b; Garland *et al.*, 2011; Murray *et al.*, 2012). Such level of classification requires one to analyse the sequence of units that make up such themes; however, minor differences in the structure of a phrase, such as the repetition of a unit for three times rather than just two or the addition or removal of the initial segment of a unit, are overlooked to maintain the integrity of the song sequence and reduce the variability (*Figure 4.7*).

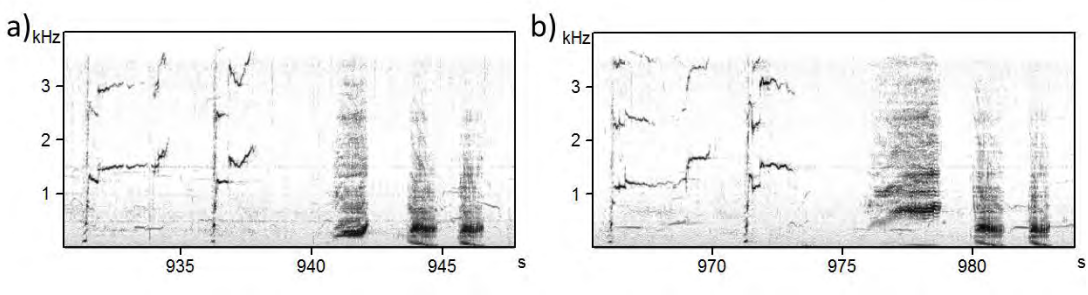


Figure 4.7: Spectrograms of two phrases that were classified as phrase D by Dr. Salvatore Cerchio in a recording from Hawaii of 1989. Whilst the similarity between the two phrases is evident, the first 3 units present slight differences between the two phrases. In particular, the third unit has a different fundamental frequency which is higher in the second phrase (b). (Spectrograms from personal notes of Dr Cerchio).

Whilst for a human observer it is easy to look at the overall structure of the phrase and pick up the similarities, a machine needs a flexible algorithm to be able to perform an automatic classification task of the type described here. However, the focus of automatic classification is at the unit level of classification. Therefore, in this thesis the manual classification focus is at the unit level. The following section describes the steps followed by the author for manually classifying the units present in the songs analysed.

4.3.2 Manual classification of the sound units

The manual classification task to prepare the data for the training stage of the automatic classification algorithm represented at least half the time spent on this project as a whole. Manual classification of the calls identified during the detection stage was obtained through several steps using Adobe Audition.

Initially all the sound units were visually and aurally inspected using spectrogram representation with the same resolution throughout the recording and assigned a category. The categories were labelled following the order of the English alphabet, according to the sequence in which new units were found. In other words, the first unit of the first song analysed was classified as “a” and the subsequent new unit encountered was called “b” and so on. To be classified under the same category, two calls had to sound the same and have the same basic structure, with only slight differences in frequency. Small variations in the fundamental frequency were allowed if the overall shape and sound of the signal was similar because sounds are

never identical. The calls were also scored according to how confident the author was that the sounds belonged to the same category.

Once the entire song of interest was classified, a table was created with the start and end time of each call, the category and the confidence score, adapting a programme to read the Audition markers with Matlab (copyright of Ricardo Antunes, University of St Andrews).

Subsequently, all the calls that had been allocated in the same category were opened in a new file within the same sequence and visually and aurally inspected for a second time. This was to confirm that all the units in the same category were indeed similar. This process was repeated for each sound category. If any error was found, it was rectified to ensure that all calls were ascribed to the correct category prior to training the automatic classifier.

Lastly, the song structure was taken into account to review the classification of the calls that were given a low confidence score. Hence, the latter were put into the context of the phrase they were encountered in and compared to phrases that could match the sequence of units present in the phrase with the unit of interest. In this case, if the unit was similar to the one present in the matching phrase and all the other units followed the same sequence in both phrases, then the unit was confidently classified in the same sound category, and its confidence score was increased.

The same procedure was followed for all the recordings analysed with the only difference being that the labelling of the sound categories started sequentially after the last label of the previous recording.

4.3.3. Manual classification of subunits

Once the unit classification was complete, further steps were taken to perform the manual classification into subunits. As anticipated in the previous chapter, a subunit is a smaller building block of the song. The concept of subunit is introduced to reduce the number of sound categories that make up the repertoire of humpback whales. Not all the units present in the songs were broken down into smaller components; the decision to subdivide a unit was based on two criteria. Firstly, only the units where frequency characteristics changed abruptly were considered as candidates for segmentation into smaller blocks. If a shift in the frequency or a frequency jump was observed (as described in Chapter 3), then the section was

isolated and compared against all the sound categories obtained from the unit classification and the other 'subunit candidates'. A subunit category was created if the 'subunit candidate' could be found also on its own, in which case it would correspond to a unit category, or associated with another 'subunit candidate' in a different order and/or in a different pair. In the latter case, we would obtain two new subunit categories, which could give 4 different combinations (subunit 1 plus subunit 2 with no silence in between the two, and the opposite, and each subunit on its own with silence before and after its start and end, i.e. appearing as isolated units). The decision process involved with segmenting units into subunits is summarised in the diagram below (*Figure 4.8*).

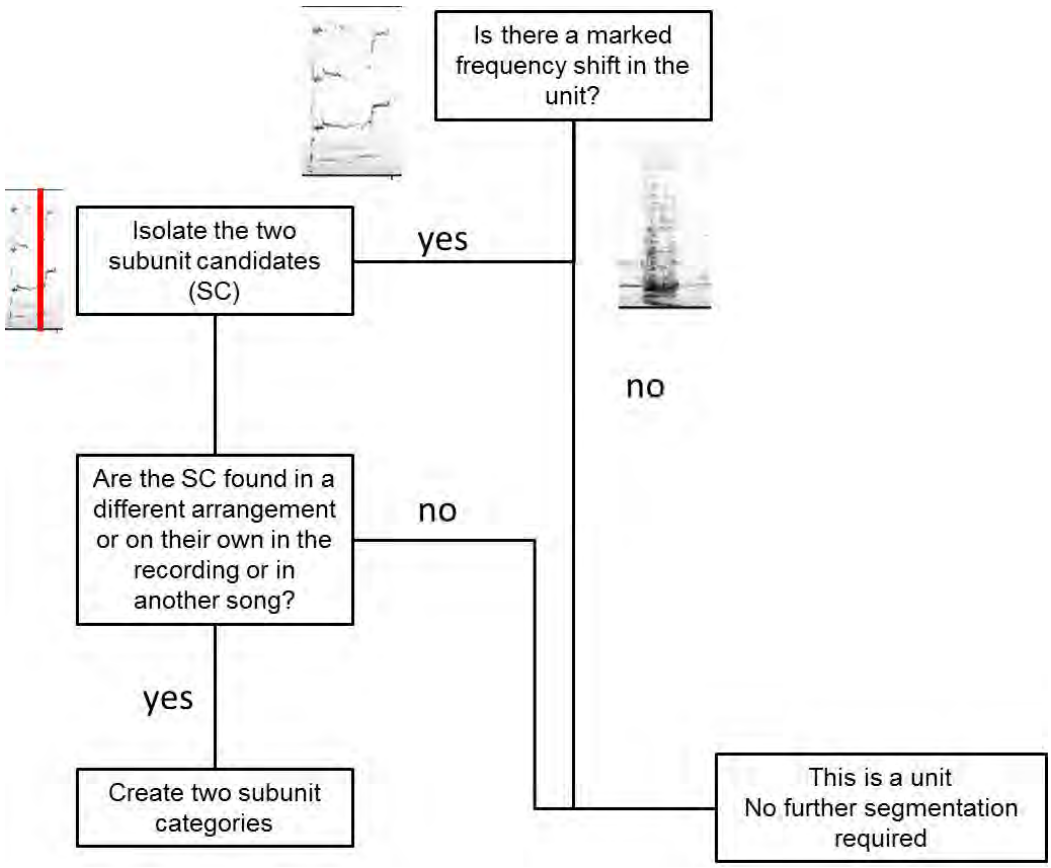


Figure 4.8: Diagram showing the process followed to determine if a unit could be segmented into subunits.

A frequency shift or jump was not immediately deemed to produce a new subunit category because we wanted to ensure that the vocabulary of subunits would be smaller than the one of units, making it worthwhile to add this layer to the analysis. It is envisaged that the larger the dataset analysed, the more and more one will encounter units that are composed of a recombination of previously observed

subunits rather than entirely new units. The process through which subunits and units were modelled using the automatic classification algorithm will be described in Chapter 6, which describes Hidden Markov Modelling and how it is implemented in this thesis.

4.4 Detection algorithm improvements

As mentioned in the previous chapter, the detection algorithm developed was simply aimed at picking up vocalisations within humpback whale songs, as opposed to periods of silence. This is different from what traditional detection algorithms are set out to do because they are usually aimed at identifying a section of sound that is likely to contain a vocalisation, which may then be used for, say, species classification for mitigation purposes. The simple energy detector used here was quite successful at detecting individual calls within songs: only a small percentage (5%) of the calls were missed or presented as a single unit instead of two because the end threshold was not triggered..

Most of the errors in the detection occurred when there were frequency jumps within a sound unit which is not surprising because the energy at points where frequency jumps occur becomes very small for a very short time period, setting off the threshold of end of the signal. Adjustment of the time window helped prevent this issue.

The algorithm for the unit segmentation was not developed further because the objective of the thesis was primarily to obtain a reliable classification method for the calls identified, rather than perfecting the detection stage. Although 99% of the calls were detected in the recordings analysed, there is margin for improvement in the algorithm in that one could fine tune it to detect more accurately the exact start and end of each vocalisation and avoid joining two units together. The calls that started quietly were problematic for the detector because they did not set off the starting threshold until some time into the unit, leading to the loss of the attack frequency if there was a slight increase or decrease in the fundamental as time went on. In addition, those calls that are made up of a series of short bursts were often split up into several sections because the energy keeps rising and dropping over the signal duration. Furthermore, if one's aim is to classify phrases to compare song repertoires across populations, it is crucial to maintain, intact, the whole sequence of a song. The level of detection accuracy obtained here though is enough to be able to identify the

phrases and themes present in songs with good SNRSNR given that the sequences are repeated several times within a song cycle. In fact, as seen in Chapter 2, humpback whale songs are extremely redundant in that the same phrases are repeated over and over, which means that if a phrase is omitted from the analysis one is likely to encounter the same sequence shortly after or before, or in the following repetition of that theme. The problem with the missed detection of a unit though is that the automatic classifier might then create a string of units whose sequence is completely shifted by one or more places (depending on the number of missed units). The result of this error would be that the entire sequence after the missed unit is modified leading to the misclassification of phrases. This can be avoided by introducing some flexibility to the system, such as an index of similarity, or by adding a level of supervision to the algorithm.

Much research effort has been dedicated to developing detection algorithms for bioacoustics signals, but humpback whale songs have proved to be particularly problematic because of their variety and because they change through time (Cerchio *et al.*, 2001b). Indeed, it is easier to detect signals that are highly stereotyped because one can train the detector to look for that particular signal, training the algorithm on its specific characteristics. Such detectors usually operate in the frequency domain. Examples that have been implemented in marine mammal vocalisation detection are matched filters, spectrogram correlation (Mellinger and Clark, 2000) and frequency contour edge detection (Gillespie, 2004; Mellinger *et al.*, 2011). The spectrogram correlation method has been adapted to detect humpback whale calls but simply to determine whether the species was present and vocalising in the study area rather than for segmenting songs (Abbot *et al.*, 2010). This was possible because some of the calls of humpback whales are relatively stereotyped and can be observed in their repertoire for several years. However, such method is not well suited for isolating all the units present within a song because no sound template can be generated that would fit all the possible units encountered and that would be transferable across populations.

5. Signal characterisation

Signal characterisation is a critical part of developing a successful algorithm for classifying sound emissions. This consists of converting the original digitised signal into a few numbers that capture the essence of the signal; in other words, through the features chosen one should be able to uniquely describe the sound, capturing the characteristics necessary to differentiate it from all those sounds that belong to a different sound category.

This chapter will present a brief overview of the main feature sets used for the characterisation of bioacoustics signals, with particular emphasis on whale vocalisations. A comparative study of the performance of different coefficients applied to the characterisation of humpback whale sounds was carried out as part of the MSc dissertation of the author (Pace *et al.*, 2009); therefore, the choice of features used in this thesis will not be discussed in great detail here. Results on the comparison of different feature sets, which was carried out in previous work by the author are presented in Section 5.2, alongside further tests that were conducted specifically on MFCCs, which are of interest for the work presented in this thesis. The chapter will focus on the feature set chosen for this project and specify the steps taken into reducing the original vocalisations into a few Mel Frequency cepstrum coefficients (MFCCs), presenting the results obtained applying different numbers of features to represent the calls of humpback whales detected in Madagascar songs. The methods used to calculate the MFCCs will also be presented, after giving a background to this set of features.

5.1 Overview of feature sets used in bioacoustics

A survey of scientific literature showed that several methods have been used for characterising humpback whale calls in the past few decades; this diversity is attributable to the variety of signals that humpback whales can produce. As mentioned in previous chapters, vocalisations range from narrowband pulses to sounds whose frequency ranges from 100 Hz up to 21 kHz (Au *et al.*, 2006); some calls are composed of numerous harmonics, whilst others have only one or two. Moreover, the mechanism for sound production in baleen whales (Mysticetes) has yet to be completely understood.

The feature sets surveyed in Pace *et al.* (2009) were specifically Linear Predictor Coefficients (LPCs), Cepstrum Coefficients, and Mel-Frequency Cepstrum Coefficients (MFCCs), because these three feature sets are more frequently used in previous studies on bioacoustic signals and they all present some characteristics which suggest that they may be suited for the task.

LPCs are the least complex of the feature sets named above in computational terms. Whilst the latter two methods are based on the Fourier transform of the signal and are well suited to characterise harmonic signals. Although MFCCs were used in previous studies of humpback whale calls (Mazhar *et al.*, 2007; Mazhar *et al.*, 2008a), they were included in our comparison of feature sets performance with some scepticism because the frequency filters of this model are tuned to human hearing.

5.1.1 Linear prediction coefficients (LPCs)

Speech can be described in terms of some characteristics of the signal that are perceptually important. Some models used to describe speech are based on the idea that these type of signals can be modelled as being produced by a periodic or random source that is driving a heterogeneous tube (Gold and Morgan, 2000), i.e. the vocal apparatus. This type of approach assumes that it is possible to clearly distinguish between the sound generation process and the filtering process that occurs in the oral/nasal cavities. The resonant frequencies of the vocal tract tube are called formant frequencies or formants (Rabiner and Shafner, 2011). The formants depend upon the shape and dimensions of the vocal tract: different sounds are produced by varying the shape of the vocal tract so that the spectral properties of a speech signal vary with the shape of the vocal tract that varies with time.

As described in chapter 3, speech sounds are produced by 3 main excitation mechanisms: i) air flow from the lungs through the throat where the flow is modulated by the vibration of the vocal cords resulting in quasi-periodic pulses, ii) air passing through a constriction in the vocal tract so that the flow becomes turbulent giving rise to noise-like excitation, or iii) air is trapped behind a point of total closure of the vocal tract so that the pressure builds up and when this is rapidly released a transient excitation is produced. The vocal tract imposes its resonances upon the excitation spectrum so as to produce the various speech signals (Holmes, 1988). In a spectrogram, voiced signals are characterised by a striated appearance

due to the periodicity of the waveform, whereas unvoiced speech signals are more solidly filled in (Gold and Morgan, 2000).

The excitation is the input of a dynamic filter system that models the combined effects of the spectral trend of the original sound source and the frequency response of the vocal tract. The transfer function of the filter is chosen to give the least-squared error in waveform prediction (Holmes, 1988).

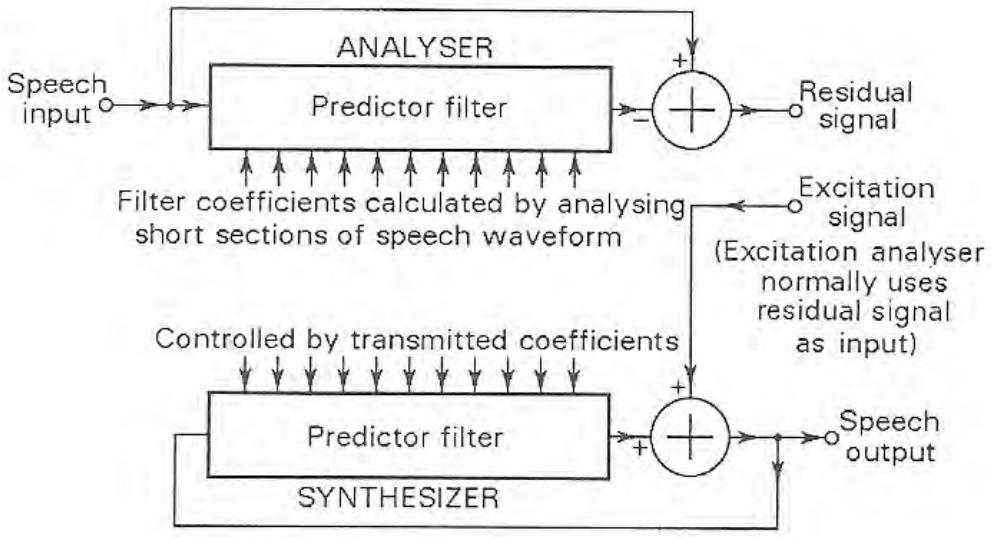


Figure 5.1: Diagram of the predictor filter of the type used for linear prediction systems (Holmes, 1988).

The excitation sequence convolved with the impulse response of the vocal tract gives the speech output (S(n)) which is the only information about the system available to us.

$S(n) = e(n) * \sigma(n)$ *Equation 5.1*

where $S(n)$ is the speech output, $e(n)$ is the excitation mechanism and $\sigma(n)$ is the impulse response.

Linear prediction coefficients are used to represent speech signals based on the assumption that a speech sample can be approximated as a linear combination of past speech samples (Rabiner and Shafer, 1978). They are calculated by analysing short sections of the speech waveform. This is feasible because in resonant systems the modes continue ringing after the excitation that caused them has ceased. Speech signals do not exactly reflect these ideal conditions;

The linear prediction filter output represents the difference between the input speech and the predictor output (i.e. the residual); the predictor coefficients are the weighting coefficients used in the linear combination based on the idea that speech can be modelled as a linear, time-varying system which can be excited by random noise or quasi-periodic pulses. The former case represents unvoiced speech signals whereas the latter corresponds to voiced speech. All the regions of the spectrum are treated equally with respect to frequency. Consequently, the variations in frequency resolution that are easily perceived by the human auditory system are not taken into account. Using coefficients that are independent of the way humans perceive sounds may be better to represent whale calls because our hearing range is much more limited than the frequencies over which marine mammals' vocalisations span, and currently we do not know how humpback whales perceive different frequencies.

The resonant properties of the synthesis filter produce a fairly accurate approximation to the spectral shapes of the formants. On some occasions the approximation is not very good due to inherent characteristics of the vocal apparatus that are reflected in the signal properties; as a result, the LPC synthesis will produce spectral peaks rather than a roughly flat spectrum.

The linear prediction method has the advantage of producing an estimate of the smoothed spectrum of a signal even when much of the influence of the excitation is removed (Rabiner and Shafer, 1978). This means that the spectrum obtained from the LPC coefficient will give us the formant peaks of the vocalisations independently from the source that originated the signal, which is useful for analysing calls of a species whose sound source is not well understood. Because we do not currently know how whales recycle air through the vocal tract to keep producing vocalisations during songs, it would make sense to use a feature set that is independent of the source characteristics. In addition, LPCs are calculated extremely easily, which is advantage to having a tool that is used by biologist.

5.1.2 Real cepstrum

Cepstral analysis was originally created to characterise seismic echoes but it is now largely employed to describe the features of human speech and musical signals. The cepstrum is the Fourier transform of the logarithmic spectrum of a signal (*Equation 5.2*) - the peculiar terminology adopted for this kind of analysis is in fact derived from the word spectrum (Deller *et al.*, 1993).

$$C(t) = FT^{-1}\{\log |FT\{s(t)\}|\}$$ *Equation 5.2*

Where $C(t)$ is the cepstrum, *FT* is the Fourier transform of the signal, and *s(t)* is the signal analysed.

Cepstrum analysis is designed for problems centred on voiced speech and is particularly good at separating the excitation and vocal system components in the frequency domain so that the formant of a signal can be identified. This characteristic can be usefully applied for pitch estimation and voice recognition (Deller *et al.*, 1993).

5.1.3 Mel-frequency cepstrum coefficients (MFCCs)

The mel-cepstrum was developed in the field of psychoacoustics to process speech signals in a way that was partly matched to human auditory perception (Deller *et al.*, 1993). A 'Mel' is the unit used to measure the perceived pitch or frequency of a tone (Deller *et al.*, 1993) and it does so using a non-linear scale to approximate the human perception of frequency. In 1940, Stevens and Volkman were able to determine a Mel scale (Deller *et al.*, 1993) of the effective frequency (Hz) versus true frequency perceived by human subjects based on experimental data (*Figure 5.2*: Mel-scale produced by Stevens & Volkman from (Deller et al., 1993).).

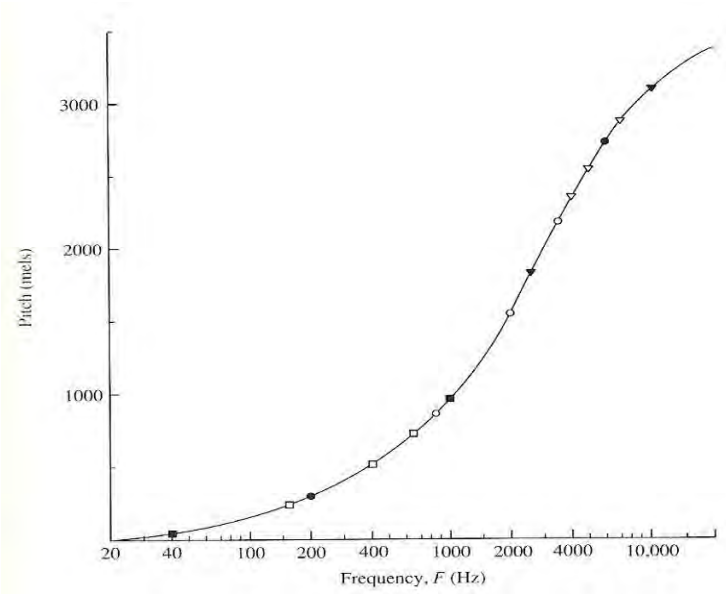


Figure 5.2: Mel-scale produced by Stevens & Volkman from (Deller et al., 1993). The x-axis represents the true frequency of the sounds played, whilst the y-axis represents the frequency perceived by the human listeners (pitch). The difference in colouration and shape of the points (squares, circles and triangles) refers to the

method through which the authors interpreted the data obtained during their experiment applying a psychological technique, known as equisection (Fagot, 1961).

The Mel scale is approximately linear below 1 kHz; whilst it becomes logarithmic above this threshold. This means that the way human listeners perceive sounds that are above 1 kHz is logarithmic. In order to obtain the coefficients to describe the salient characteristics of a signal, a series of filters is computed. However, in this case, the filters applied are not uniform but their bandwidth varies according to the centre frequency; in particular as frequency increases the filter becomes broader to reflect the shift in the way humans perceive sounds below versus above 1 kHz (Deller *et al.*, 1993; Gold and Morgan, 2000) (*Figure 5.3*: Mel-scale filter bank (Gold and Morgan, 2000).
).

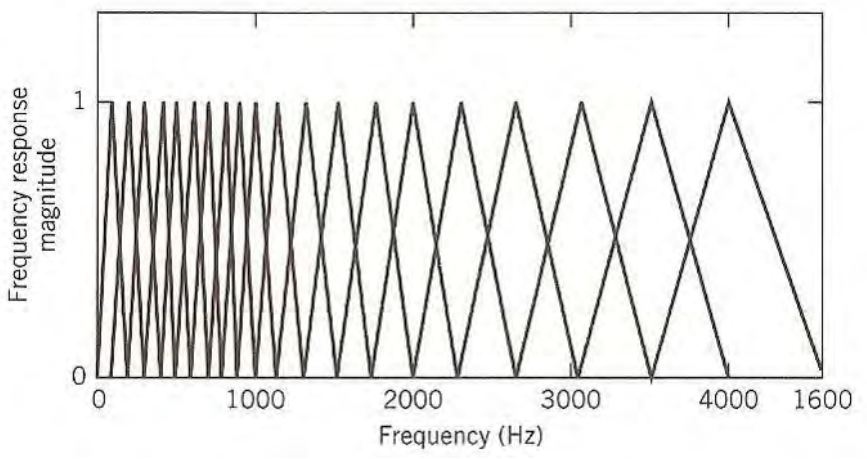


Figure 5.3: Mel-scale filter bank (Gold and Morgan, 2000).

The Mel-frequency cepstrum coefficients are derived by computing the Fourier transform of the signal and taking the power spectrum of the FT to map it on the Mel scale using the triangular overlapping windows described above. The Mel scale is calculated based on the way human listeners perceive the pitch of sounds that are judged by them to be equidistant from each other. The frequency of 1 kHz1kHz is used as a reference point for the scale, where 1000 Mels correspond to 1 kHz1kHz for a tone that is 40 dB re 20 μPa above the threshold of hearing of the listener. TheThe magnitude squared of each Mel-frequency is put on a log scale and its transform is calculated. The amplitudes of such a spectrum correspond to the MFCCs (Gold and Morgan, 2000).

The magnitude of Fourier Transform of the signal is weighted by a triangular filter frequency response of centre frequencies, f_0, specified according to the Mel scale (Deller, *et al.*, 1993) shown in *Figure 5.3*: Mel-scale filter bank (Gold and Morgan, 2000).

The Mel frequency scale is defined by:

$$Mel\left(f_0\right) = 2595\log_{10}\left(1+\frac{f_0}{700}\right) \qquad \textit{Equation 5.3}$$

The Mel cepstral coefficients for the n^{th} frame, denoted $\mathbf{c}(n)$, are computed as a Fourier transform of the filter bank outputs expressed on a logarithmic scale. The raw MFCCs can also be augmented with the so-called Delta MFCCs (ΔMFCCs), which measure the temporal rates of change of the MFCCs. The ΔMFCCs, $\mathbf{d}(n)$, are calculated as the slope of a regression line, fitted using least squares, through window of coefficients centred on, but excluding, $\mathbf{c}(n)$. This computation can be simply realised using

$$\mathbf{d}\left(n\right) = \frac{\sum_{\tau=1}^{K} \tau\left(\mathbf{c}\left(n+\tau\right)-\mathbf{c}\left(n-\tau\right)\right)}{2\sum_{\tau=1}^{K}\tau^2} \qquad \textit{Equation 5.4}$$

where K defined a user-specified window size ΔMFCCs (Young, *et al.*, 1995). In the implementation report here K=2 is used.

The use of ΔMFCCs allows, for instance, to distinguish between upwardly chirping vocalisations and those vocalisations that contain downwards frequency sweeps.

The choice of using MFCCs to describe whale calls may seem unusual given that one of the few things we known about their sound perception is that their hearing is different from ours in that it is more sensitive at lower frequencies and able to detect calls over a wider frequency range; however, what we are trying to establish is whether the automatic classifier is as accurate as a human listener at grouping similar whale calls. Therefore it makes sense to use MFCCs to mimic the way a trained listener perceives the characteristics contained in the variety of vocalisations produced by humpback whales.

5.2 Feature sets performance comparison

As previously mentioned, all the feature sets described in the above sections were used in studies involving characterisation of humpback whale signals and MFCCs are the most common. The results showed that calls characterised using MFCCs were classified more accurately than those described using the other two feature sets for nearly all call types despite the fact that they are based on an anthropomorphic perception of sound. This apparent contradiction can be reconciled by expressing the objective of this work as attempting to mimic human perception of humpback whale calls, i.e. refining the objective to be that of developing a system with a classification capability that approximates that of the human listener. Similar objectives in other machine learning tasks typically prove ambitious and it is certainly true that, were the final automated system to perform at level broadly equivalent to a human listener, then such as system would be extremely useful. The success of the MFCCs relative to the other feature sets is, in part, a consequence of the fact that the MFCCs are more robust with respect to ambient noise (Deller, *et al.*, 2003) and typically ambient noise levels in underwater recordings are comparatively high.

This section is divided into two sub-sections: the first one gives a brief overview of the classification performance obtained as part of the MSc project of the author which was tested on recordings of 2008. Whilst the second section presents the results obtained at the start of the PhD project, when the performance of the feature sets was tested again on more recent recordings for two reasons. Firstly we wanted to ensure the validity of the results obtained the previous study, and secondly we wanted to compare the performance of the feature sets when categorising songs based on their unit components versus their subunit components that were described in Chapter 3. This was to ensure that MFCCs would perform better than the other feature sets independently of the model chosen because the aim of this study is to determine which of the two building blocks (units versus subunits) is more suited for comparing songs of humpback whales across the world.

5.2.1 Feature sets performance comparison with *k*-means algorithm

A song recorded in August 2008 was initially analysed as part of the MSc project of the author to test the performance of the three feature sets analysed. The automatic classification was perfomed using a *k*-means algorithm, which is unsupervised but

requires the user to enter the number of clusters (*k*) that the signals need to be subvided into.

To evaluate the performance of the feature sets, a manual classification of the calls present within a song was carried out by the author as described in the previous chapter, whilst the automatic classification was performed using the *k*-means algorithm. The feature sets were calculated respectively by using the standard MATLAB functions 'lpc' and 'rceps' , and a script developed by Y. Andrianakis at ISVR was used for computing the MFCCs. The model order used for these set of tests was 12, value which recurred in the literature and that also corresponds to the model order chosen for speech recognition tasks. Further tests are described in the next chapter to estimate the optimal model order. The tests were conducted classifying the sounds using a *k*-means algorithm which is simple to apply.

The results were compared against a manual classification which was carried out following the method described in Chapter 4. The sound classes of the results presented are expressed by numbers in chronological order of the appearance of each sound class in the song. This is different from the nomenclature that was described in Chapter 4 to avoid confusion; the nomenclature described in the previous chapter will only be adopted for tests conducted applying the HMM classification which is the main focus of this thesis.

The frequency of occurrence of each vocalisation identified through the manual classification is presented in *Figure 5.4* to give an overview of the variability in the song repertoire.

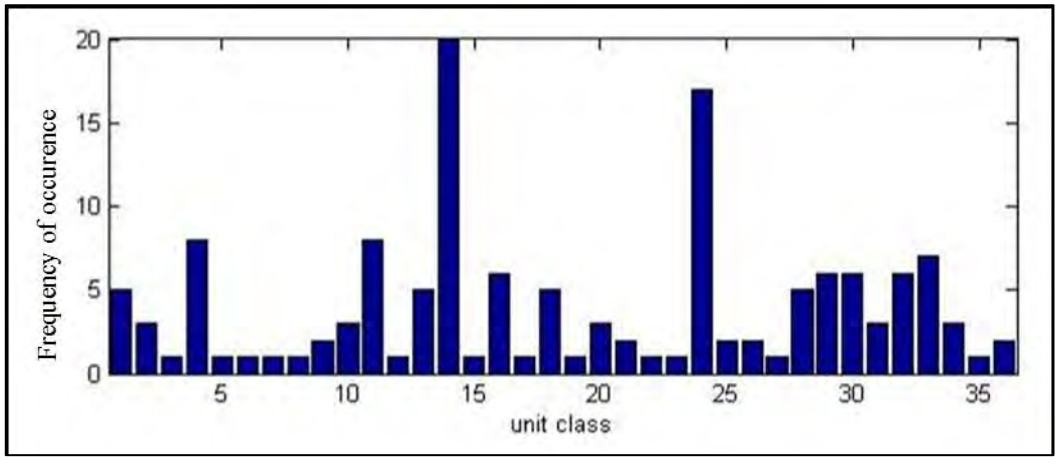


Figure 5.4: Histogram of occurrence of the 36 sound classes identified manually.

The performance of the feature sets was estimated by calculating the mean between all the units of a group that were correctly identified and are presented here as a percentage of the total number of units of each group (Figure 5.5). Note that in Figure 5.5, the vocalisations which correspond to sound units are grouped into the broad sound categories described by Dunlop et al. (2007a) for social sounds. This is the only instance where this grouping is used in this thesis because it was subsequently decided that the calls identified in the songs of Madagascar could not be well represented by such categories and we preferred to use categories that did not require setting thresholds to define what constituted low, mid and high frequency sounds.

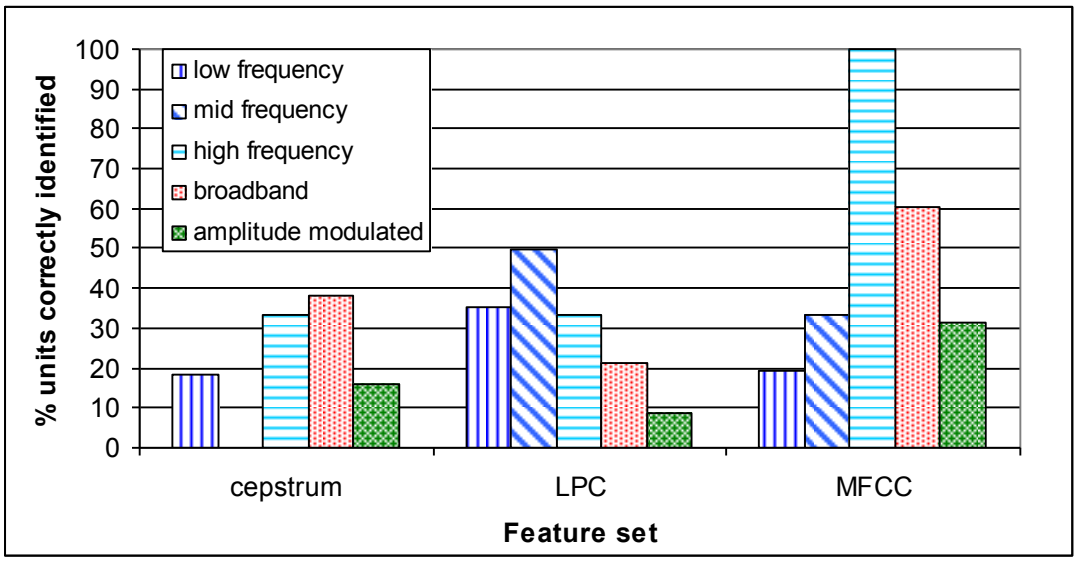


Figure 5.5: Performance of the three feature sets used in the study. Here, the percentages of vocalisations correctly classified using the k-means algorithm are presented according to the five major groups within which all units are clustered. Note that in this figure, frequency refers to the sound pitch rather than the frequency of occurrence.

Overall, the Mel-frequency cepstrum coefficients classified correctly the same vocalisations within a group of sounds more often than the other two feature sets. The performance was excellent for high frequency vocalisations although a very limited number of examples was present in this instance. On the other hand, LPC were the best predictors for low and mid frequency vocalisations although the signals characterised with these coefficients were clustered together correctly only in 35% and 50% of the cases, respectively. MFCCs were particularly good at characterising

broadband and amplitude modulated vocalisations, which were the most frequent sounds encountered in the recording section analysed. The real cepstrum coefficients were the worst for classification purposes and, in particular, no mid-frequency vocalisations were clustered running a *k*-means algorithm.

A detailed list of the performance of each feature set for the classification of the various sound units not divided by group is given in *Table 5.1*. Vocalisations that were observed only once during the recording were omitted from the table as it would be impossible in this case to assess the performance of the automatic classification.

Manual classification		**Automatic classification (% correctly classified)**			**Sound type**
Unit class	***Frequency***	**Cepstrum**	**LPC**	**MFCC**	
1	5	0	40	60	am
2	3	67	0	67	am
4	8	25	29	25	am
9	2	0	0	100	bb
10	3	67	0	100	bb
11	5	40	40	40	low
13	5	40	0	0	am
14	20	13	16	23	low
16	6	0	33	33	mid
18	5	0	0	0	low
20	3	0	67	0	low
21	2	0	0	100	am
24	17	22	18	16	low
25	2	0	0	0	low
26	2	0	0	0	am
28	5	40	40	60	bb
29	6	0	67	33	mid
30	6	33	33	42	bb
31	3	0	0	0	am
32	6	50	33	0	bb
33	10	20	30	23	am
34	3	67	67	100	high
36	2	0	0	100	high

Table 5.1: List of the units classified manually, their frequency and type – where am represents amplitude modulated, 'bb' broadband, and 'high', 'mid' and 'low' refer to the frequency as in the groups described in the previous section. The performance is given by the percentage of vocalisations classified together for each group. If

some units were recognised correctly as being the same but they were split into two classes, then a mean value was calculated and listed in the table. Such instances are highlighted in red.

The MFCC coefficients performed extremely well at classifying some particular vocalisations, namely classes 9, 10, 21, 34 and 36. The latter two are high frequency sounds, hence as observed previously this feature set seems particularly useful to characterise such sound types. Nonetheless, it is important to note that the number of such vocalisations present in the section of song analysed is very small, i.e. 5 in total combining both unit classes. Therefore, a more extended analysis is required to confirm such results.

Both the LPC and cepstrum coefficients recognised one type of high frequency vocalisation but failed to cluster it together with the other class of high frequency units. The LPCs performed better in the case of pulsed vocalisations rather than broadband and amplitude modulated calls. They were especially consistent in characterising unit types 20, 29 and 34 which were respectively low, medium and high frequency sounds. On the other hand, the performance of cepstrum coefficients was better for broadband and amplitude modulated sounds than for the other types. Although good performance was noted for high frequency vocalisations as well as before, this might only be a consequence of a limited sample size.

Furthermore, the percentages presented in *Table 5.1* do not take into consideration the fact that even though some vocalisations were clustered into different classes by the *k*-means algorithm according to the features obtained with the three sets, they might be the only calls present within those classes. This means that the classification method split the same vocalisations into two or more groups according to minor changes in their characteristics. However, such results could only be a consequence of the fact that the order of the clustering algorithm must be fed by the user and can be very subjective and also that there is no feedback system allowing the classification to adapt to the signals encountered – as can be the case for classification based on neural networks.

5.2.2 Feature sets performance for unit versus subunit model using *k*-means algorithm

A song recorded in Madagascar in August 2009, which is used for further analysis, was manually classified and the *k*-means algorithm was applied to test the feature sets performance (*Figure 5.6*). The results showed below are based on the comparison against manual classification of the subunit components (described in Chapter 3) of the song, i.e. the smaller building blocks than the ones normally used for humpback whale classification tasks.

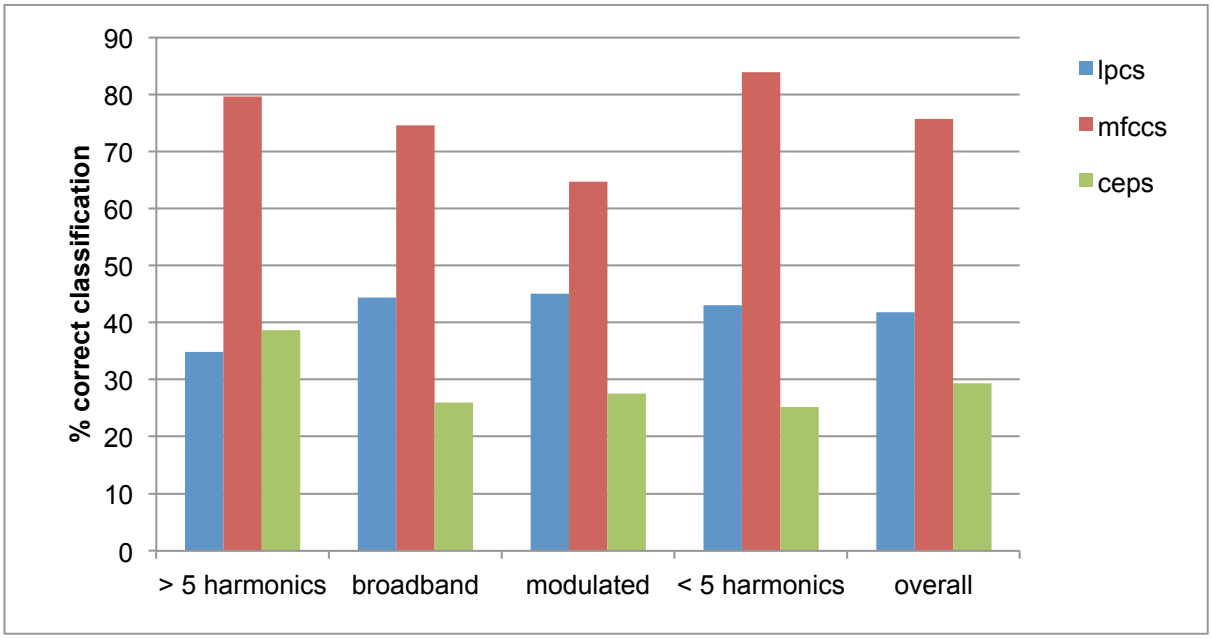


Figure 5.6: Comparison of the performance of subunit classification obtained using the three feature sets. For each test, 12 features were used to describe all the call categories, based on results obtained during MSc work (Pace et al., 2009). The different classes of calls were grouped into broad categories based on their frequency characteristics.

The results show that MFCCs classified all the types of vocalisations emitted by humpback whale better than the other two feature sets. The performance of LPCs and cepstrum coefficients was extremely poor as they classified vocalisations correctly in less than 50% of the cases. Cepstrum coefficients did slightly better than LPCs in classifying songs with many harmonics as one would expect given that they are based on the Fourier transform of the signal; however, the performance was worse than one would expect if the calls were randomly classified (<50%).

Subsequently, we compared the classification the classification performance based on subunits versus units, again using the *k*-means algorithm (*Figure 5.7*).

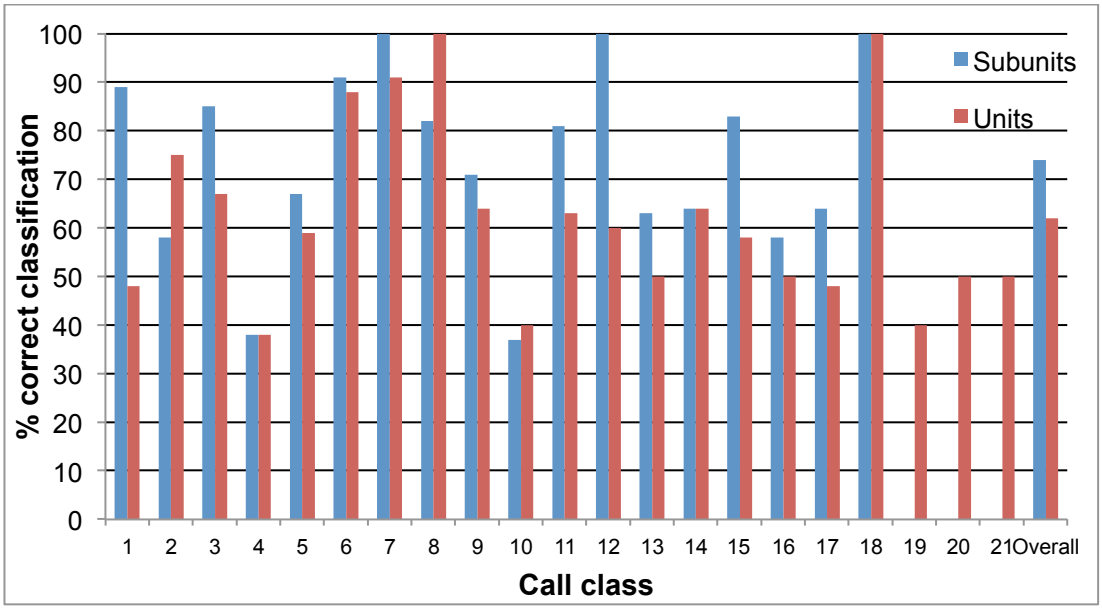


Figure 5.7: Classification performance of units versus subunits obtained comparing a manual classification carried out by the main author and automatic clustering where MFCCs features were applied in the k-means algorithm (model order dictated by the number of classes manually identified). 18 subunit classes and 21 unit classes were identified through the manual classification.

From visual inspection, thet results show that the subunit model performed better at classifying the calls emitted by humpback whales overall and for mostmost call types. Further results and considerations about the performance of models based on unit versus subunit classification are describedin detail in chapters 7 and 8.

5.3 Conclusions about coefficients choice

This chapter presented a brief overview of the coefficients that have been used to characterise the calls emitted by humpback whales in songs or in a social context, given that there is no consensus on the most appropriate method to model them. We have shown that using Mel-frequency cepstral coefficients as input features for the classification algorithm achieved more accurate results both in terms of the overall classification performance and in terms of the number of calls that were classified correctly compared to the cepstrum and linear prediction coeffcients.classification. It may be that MFCCs gave higher classification results because they mimic the way human listeners perceive sounds; if so, they are suitable to be employed for carrying

out large scale automatic classification of humpback whale songs with the goal of reducing the computational load for researchers that wish to compare data from different populations.

The results peresented also showed that the *k*-means algorithm is unsuitable for the task of humpabck whale'swhale song classification because the user needs to specify the number of call classes to run the algorithm and, most importantly, because the algorithm tends to spread the unit number of calls evenly amongst classes whislt we know that a song contains vocalisations that are repeated a few times and others that are very common, as shown in Figure 5.4. This suggested a different classification algorithm is needed to improve the classification performance, which provided motivation to consider the use of Hidden Markov Models, that are commonly used in speech recognition tasks.

6. Hidden Markov Models

Hidden Markov Models (HMMs) are widely adopted in speech recognition tasks because they are extremely flexible and allow one to classify sounds from a series of given observations of the signal. The main reason why Hidden Markov Models were adopted for this project is because of their ability to model signals that are variable in duration, which is a key element to consider when classifying bioacoustic signals such as humpback whale calls. In addition, HMMs can cope with the fact that some calls may be extremely common whilst others rare, and will not try to distribute the observations into even categories. The latter needs to be considered because during a song some vocalisations are extremely common and they will be repeated over and over again throughout the duration of the song, whilst others may be present only at transition stages between themes or phrases, making them quite rare. One could discard rare calls from the analysis but this would be detrimental for the objective of this project, which is to build an automatic classifier that can be used to compare calls across populations where the rare calls might be used differently within a song and become common calls. Hence, we want the vocabulary of vocalisations to be as comprehensive as possible.

The first part of this chapter presents an overview of Hidden Markov Models and of how they are used in speech recognition. Whereas, the rest of the chapter details how HMMs have been adapted to model humpback whale vocalisations in this project.

6.1 Overview of Hidden Markov Models

A Hidden Markov Model is a stochastic process that allows the prediction of the statistical properties of the signal, which, in several cases, are sufficient to characterise the signal (Rabiner, 1989)..

A HMM is an extension of the Discrete Markov Process or Markov Chain, i.e. a discrete random process whose conditional probability state at the next step depends solely, in a stochastic manner, on the current state. An HMM extends the Markov process to include an observation which is a probabilistic function of the state. This means that there is an underlying stochastic process which is not observable or "hidden", hence the nomenclature of Hidden Markov Model.

Given a set of N distinct states $S_1, S_2, ..., S_N$, and a series of time instants t=1, 2,...t_n a discrete, single state Markov Chain is defined by:

$$P(A,B) = P(A)P(B|A)$$

$$P[q_t|q_{t-1}|,q_{t-2}) = P[q_t|q_{t-1}) \qquad q_{t-1}) \quad \textit{Equation 6.1}$$

wherew q_t is the state at time t.

The state transition probabilities a_{ij}, defined as the probability of moving from state i to state j, have the following properties:

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \textit{Equation 6.2}$$

$$a_{ij} \geq 0 \qquad \textit{Equation 6.3}$$

Extending this definition to include the case in which the observation is a probabilistic function of the state, we obtain the Hidden Markov Model. Intuitively, for each observable sequence there will generally be more than one model that can explain the observable sequence, but one needs to find the one that will maximise this probability. The state observations can be modelled using a variety of statistical distributions; in this case, they were modelled through a Gaussian distribution, which means that the variance was equal to 1 and the mean to zero.

When designing such model one is faced with three major problems:

1) Given the observation sequence and a model, how do we calculate the probability of the observation sequence, given the model?
2) Given the observation sequence and the model, how do we choose a corresponding optimal state sequence that best explains the observations?
3) How do we adjust the model parameters to maximise the probability of the observation sequence, given the model?

Rabiner (1989) covers all the aspects involved with formulating an HMM model including suggestions to tackle these three issues. The next section will detail the implementation of the HMMs used in this project and showing the stages at which these problems are solved.

6.2 HMM in speech recognition

As previously mentioned, Hidden Markov Models are the most widespread tool used nowadays for automated speech recognition. They can be used to model speech at various levels from the basic phone to the word or sentence recognition tasks,

depending on what the user sets as hidden state of the Markov models. Usually, the hidden layers chosen are phones so that words can be represented as a sequence of phone units and the models are left-to-right meaning that they do not allow a state to transition back to its previous one. Phones are the basic building blocks of speech, and they are the actual sounds that are produced during speaking, which can be ascribed a class of phoneme. Phonemes are the theoretical units to describe how speech conveys a meaning. Whilst the phoneme is the ideal model that corresponds to the full set of articulatory movements needed to produce a sound, the phone represents the actual utterance. Such distinction is necessary because different people may articulate a sound that conveys the same meaning (phoneme) in different ways (phones) due to his dialect, gender, age and other effects that affect sound generation (Deller *et al.*, 1993). Phones that represent variations of the same phoneme are collectively termed allophones. Below is an example of how a word is broken down into its phonetic components, which are modelled through a left-to-right HMM for speech recognition tasks (*Figure 6.1*).

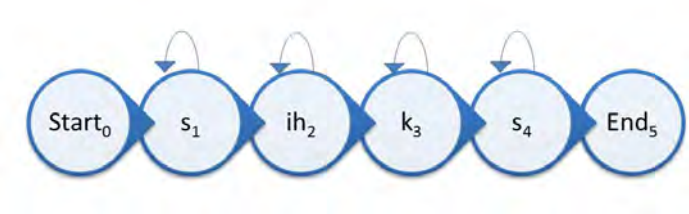


Figure 6.1: Bakis diagram of a left–to-right HMM of the word 'six', whose components are represented through the phones that make up this word. Note that the word has been described phonetically, instead of transcribing each state following the orthographic representation of the sounds. The sequence of states is expressed through the subscript numbering. Two additional states are added at the start and at the end of the word to inform the model that the utterance is about to start and finish respectively.

One can represent the totality of sounds that convey a specific meaning in a language by using a few phonemes. The juxtaposition of this limited number of phonemes will generate all the words present in the vocabulary of that particular language. Intuitively, the benefits of this system are huge in terms of automatic recognition tasks.

For modelling purposes, phonemes are distinguished into two categories: continuant and noncontinuant. These two categories simply distinguish between sounds that are

produced by a steady state configuration of the vocal tract (continuant phonemes) and sounds during which the vocal tract changes configuration (noncontinuant). An instance of the former is vowels, whilst the latter include diphthongs (Deller *et al.*, 1993).

Given the flexibility of Hidden Markov Modelling, several systems were developed for automatic speech recognition depending on the tasks one wants to achieve and the dataset available to train the models. The principal choice that one is faced with when building a speech model using HMMs is deciding the level at which one wants to carry out the analysis. Specifically, given that we can analyse the language at different levels (e.g. sentence, word, phoneme etc.), one needs to choose whether to build a phoneme model, a word model, and so on. Once this is set, then one can fine tune numerous settings to adapt the model to reflect the language syntax (Deller *et al.*, 1993; Young *et al.*, 2000). For instance, the likelihood of finding a particular phoneme after a given one can be fed in the system to improve the performance of the recogniser. This is possible because we know exactly how the language is constructed and that particular sounds never occur before or after others. An example is presented below of the type of grammar that can be constructed for modelling speech using HMMs (Figure 6.2).

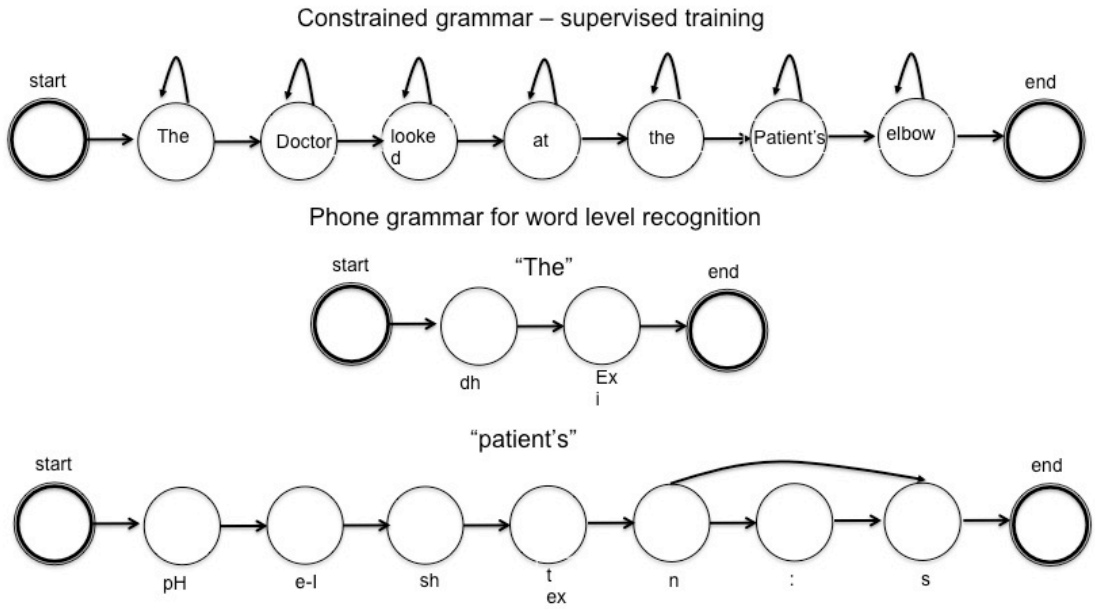


Figure 6.2: HMM model example for the sentence "the doctor looked at the patient's elbow" (top) where each word is modelled through one HMM. The model of the sentence is left-to-right so that the sequence of words must be respected to model this particular sentence. The second example (bottom) shows a model, which is trained on phones so that each word is broken down into smaller components that are unique. Note that the model of the word "patient's" allows skipping one state from 'n' to 's' to account for different pronunciations that may occur.

Whilst the word based model is faster to implement because training is required for each word. The phonetic unit training is much more flexible because the HMM is not constrained by the word grammar. The latter type of model can more easily cope with utterances from different speakers, which is a very attractive characteristic for comparative studies (Picone, 1990; Deller *et al.*, 1993).

Considering that little is known about the way animals formulate their language, one cannot assume that they arrange sound sequences in a particular order to convey a specific meaning. What is known about how animals structure their signals is based on observations and some tests carried out in captivity to understand more about their cognition. In the case of humpback whales we know how they structure their songs but do not have any information about the meaning associated with each song sequence or their smaller building blocks. Therefore, the HMMs that were developed for speech need to be adapted to the structure of the signals that we wish to analyse in light of the goal of our study. The rest of this chapter will describe how HMMs

were used in previous bioacoustic studies and, in particular, how the models are adapted in this thesis.

6.3 HMM for bioacoustics signals

Hidden Markov Models are very flexible and useful to classify sounds whose states are not directly observable. Research suggests that HMMs are a useful tool for analysing bioacoustic signals because they allow for the change in spectral characteristics over time, unlike more common and straightforward classifiers (Ren *et al.*, 2009). The importance of using a tool which takes into account the time-varying component of the signal was highlighted in the context of studies of cetacean acoustics, particularly for those species like bottlenose dolphins and killer whales that have a complex repertoire (Deecke and Janik, 2006b); and they suggested the implementation of Dynamic Time Warping (DTW), which is a common tool in the recognition of isolated word recognition for small vocabularies in human speech. However, HMMs were shown to be more robust to noise and to vocalisation variability than DTW, at least for bird songs (Weisburn *et al.*, 1993). This, and the fact that the feature extraction process of HMMs does not require to measure the frequency contours of the signals in the pre-processing stage – which is very time consuming – makes them more attractive than DTW for the task.

6.4 Implementation of HMMs for humpback whale song classification

The HMMs were implemented using the HMM Toolkit (HTK) (Young *et al.*, 2000). This toolbox has been used by other authors in the context of bioacoustics (Kogan and Morgan, 1998). Building an HMM consists of four main stages:

i) Definition of the model structure.

ii) Feature extraction which consists of dividing the data into frames and computing summary features for each frame.

iii) Model training: a portion of the dataset for which the manual labels are provided are used during this phase. The training data is used to estimate the parameters of the HMM that maximises the likelihood of the training sequence.

iv) Automatic classification: a set of data, generally different to that used for training, can then be applied to the HMM algorithm to identify which of the units are most likely to have generated the measured sound.

HTK is available for free download on the web, alongside with the manual for the implementation of the models. This program was chosen over others for several reasons. Firstly, the toolkit has been extensively used in scientific literature for several applications, including speech recognition and bioacoustics; it is a reliable tool and it is easily replicable.

Furthermore, HTK is made up of several toolboxes which make it a very flexible tool so that one can tailor the model to your needs and the dataset available. This section is concerned just with the tools employed to build the model of this project; for a comprehensive description of the toolkit one is referred to the user manual (Young *et al.*, 2000). The toolbox used to train the HMM requires five inputs:

1) The training data
2) The labels for each training data
3) A list of the sound classes, that is to say a list of HMMs each of which corresponds to a sound class
4) A prototype Hidden Markov Model for each sound class
5) A "grammar" that specifies the language structure

A summary of the steps involved in the full HMM development is presented below (*Figure 6.3*)

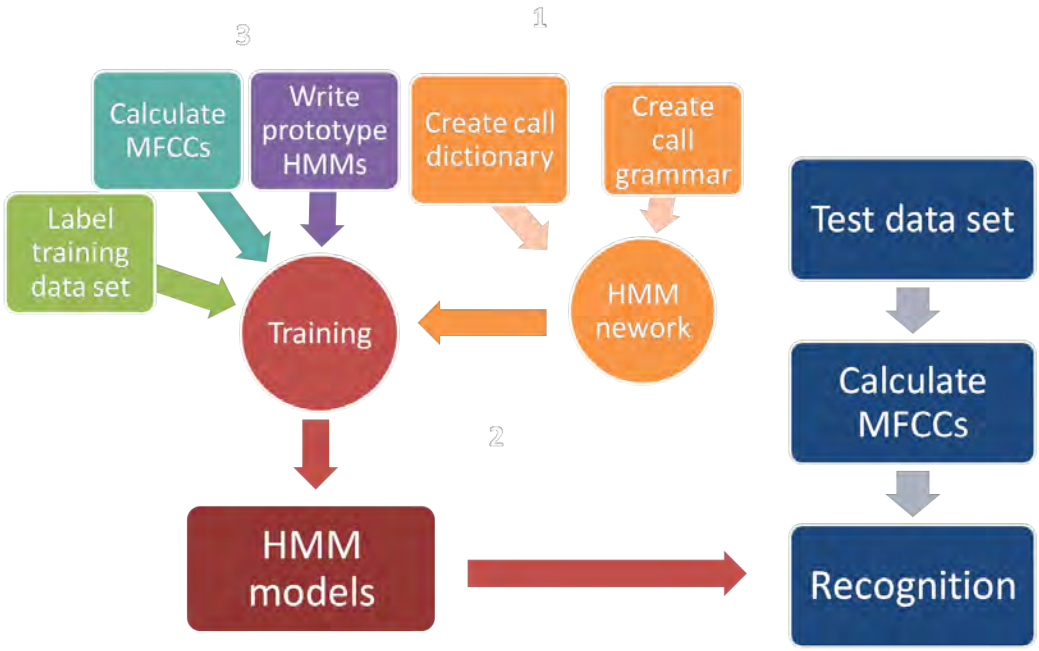


Figure 6.3: Diagram of the method followed for the training and recognition processes.

The training stage of Hidden Markov Modelling is represented by the steps on the left-hand side of *Figure 6.3*. Setting up the model requires the user to input several variables that define the structure of the calls one wants to classify during the recognition stage. Although, this can be time consuming it allows tailoring the model to the bioacoustics signals to be analysed. In addition, once the model is set up it requires very little (or no) modification for analysing additional data because one could just use the HMMs trained on a different dataset to recognise new vocalisations. As shown in *Figure 6.3* once the system is trained, the only necessary step required to run the recognition is the calculation on the features (MFCCs) of the test dataset. The calculation of the MFCCs using HTK and the other steps involved in the training and testing are described in the next sections.

6.4.1 Feature extraction for Hidden Markov Modelling

The individual calls identified during the manual classification stage described in Chapter 4 were segmented into individual '.wav' files, each containing one vocalisation that may (or may not) have some silent section at the start and/or the end of the call (note that a silent section will contain ambient noise). The incoming data stream from each file is segmented into overlapping frames of 25 ms duration. The data in each frame is characterised through a set of M features (MFCCs); a suitable value for M is identified through testing. The choice of frame size has to take account of the fact that some vocalisations are very short (0.2 s) and that others change very rapidly in terms of frequency; a longer window would fail to capture such rapid sweeps. The choice of the number of features (M) is important since it, not only, controls the computational load of the system but also the amount of data required to train the system.

The way in which the HMM toolkit calculates MFCC is described below, where details are given of the parameters chosen for the analysis carried out in this thesis. A configuration file needs to be specified which includes several inputs to inform the toolkit about the characteristics of the files used in the analysis, as described in *Figure 6.4.*

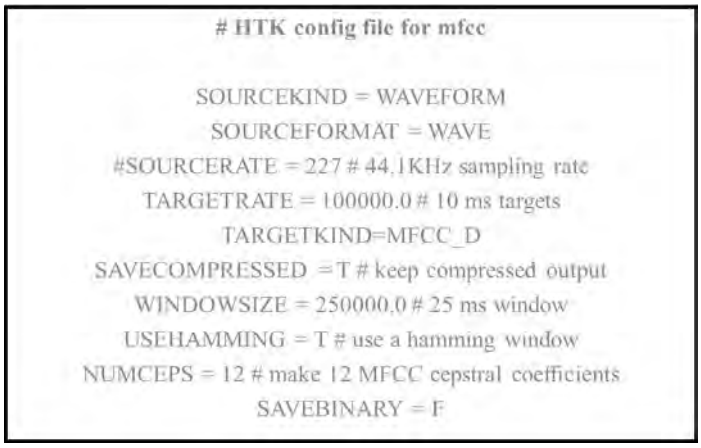


Figure 6.4: Box showing the configuration used for calculating the MFCCs using HTK. The format of the source file in the input needs to be specified; in this case, all our input files were in '.wav' format, with sampling frequency of 44.1 kHz. The source rate depends on the sampling frequency of the recording, and specifies the data points in each window frame. The target rate and kind refer to the output file, in our case we used MFFCs and ΔMFCCs (expressed by the added command _D in the target kind). This means that the output file we obtain will be a file in '.mfc' format (i.e. an HTK format) containing the number of features specified in the NUMCEPS setting. Note that the number refers to the number of features that one wants to obtain for each of the coefficient types specified; in other words, with the configuration depicted in this figure one will obtain an output containing 12 MFCCs and 12 ΔMFCCs.

The type of feature sets used were chosen accordingly to the results of the tests presented in the previous chapter but further tests were conducted to choose the number of features to be used to describe the vocalisations, to maximise the classification performance, as discussed in Section 6.4.2.

6.4.2 Determination of the dimension of the feature vector

Although MFFCs and other feature sets have been used in a variety of studies on animal bioacoustics including classification of humpback whale calls, the number of features chosen to describe the calls is often omitted or when specified not justified (Potter *et al.*, 1994; Mercado III and Kuh, 1998; Mercado III *et al.*, 2003; Mazhar *et al.*, 2007; Mazhar *et al.*, 2008b). Therefore, a step in determining the classification performance using HMMs consisted of testing the number of features described to

understand how changing the number of coefficients included in the analysis changed the outcome of the classification performance.

The number of MFCCs commonly used in the literature is 12, perhaps because this is the number chosen for representing speech and scientists have not tuned it for bioacoustics signals. Initial tests were conducted in order to identify an appropriate size of feature vector, *M*. The number of feature sets used was chosen to maximise the recognition performance: the computational load of the system not being a high priority. One anticipates that for a finite training set, tested on a separate testing data set that for a small number of features recognition rates will be low because of the inability of the features to adequately represent the data. Conversely, a large feature vector, the tendency of the system to over-fit the data will lead to poorer performance. The value of *M* identified through this testing will define a suitable value which can be trained with data sets of the size available to us. Larger data set may allow successful training of systems with larger feature vectors.

Data from a single recording from August 2009 (i.e. Mada09a in Table 4.1) were used to determine the *M* with the highest classification performance of units present in the song (*Table 6.1*). The training set comprised 119 units, which corresponded to 50% of calls of each class for 14 classes of units, whereas the test set included 181 units.

M	**MFCCs**	**MFCCs plus Delta MFCCs**
8	72.0%	87.3%
10	83.2%	91.7%
12	87.4%	94.5%
16	84.%1	90.1%

Table 6.1: Correct classification rates, expressed as percentages, for a range of feature dimension M for the MFCCs alone and MFCCs plus their Δs. Note that when the ΔMFFCs are used the total feature dimension is 2M.

Table 6.1 demonstrates that maximum classification performance for this data set was obtained with 12 features, a value which is typically used to represent human speech too (Grimm and Kroschel, 2007). The largest feature set number tested was 16 because at this point the performance dropped meaning that there was no advantage in further increasing the computational load.

Taking into account the advantages of using a smaller feature vector one might reasonably suggest that sacrificing some performance to realise these benefits could be justified. For instance, the reduction in correct classification rate using 10 MFCCs as opposed to 12 MFCCs in only 2.8%. However, in the system developed here computation time is not a critical feature: the computational efficiency of the algorithms means that even with 16 features the overall computational times for processing a data set of 300 vocalisations are typically in the order of 2 seconds when run on a laptop Dell Latitude E6400. When analysing a very large data set reducing the performance by a small percentage could mean misclassifying a significant number of calls, considering that in 1 hour a humpback whale is likely to produce around 1,000 sound units. Furthermore, in the dataset analysed there were two sound classes whose classification performance was particularly affected by the reduction in number of features, one of which is the most common call throughout the recording. This means that choosing the wrong number of feature sets can affect disproportionately different call categories.

The next chapter will present Hidden Markov Models and how they are implemented in this project for the classification of humpback whale calls.

6.4.3 Model structure

Each call class (i.e. either unit or subunit depending on the model chosen) was represented by one left to right HMM with one state if the call frequency was stable throughout its duration or two to three states if the frequency was varying, e.g. in the case of an upsweep or down-sweep, plus two "non-emitting" states at the start and at the end of each model (*Figure 6.5*).

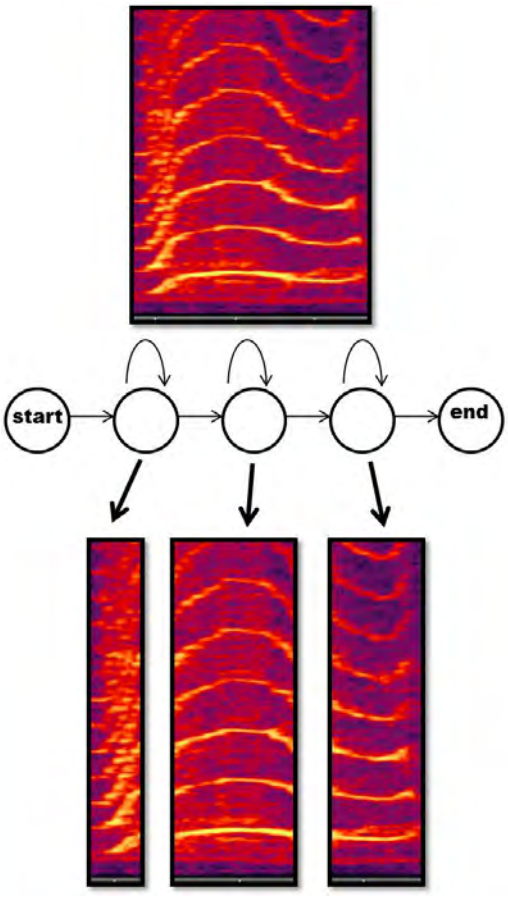


Figure 6.5: Example of humpback whale unit vocalisation which is used for training an HMM. This unit is broken down into 3 states that correspond to three changes in direction within the call. The first segment (a) is a quick upsweep, the second segmented is an upside down arch (b) and the last part of the call is almost flat but with a slight upward curvature (c). Two states at each end mark the start and end of the unit.

A definition file was therefore created for each HMM which includes a general description of the features, i.e. type and vector size, and the number of states. The number of states of each HMM determines the size of the transition matrix: a three state HMM will have a 3×3 transition matrix. Prior to training, each state is initially defined by a Gaussian distribution with 0 mean and variance equal to 1. The transition between observations was modelled by a Gaussian mixture where, by definition, each state transition probability was a real number from 0 to 1 and sum to unity. Given that each recording segment containing a call could include a silent part at the start and/or at the end of the sound clip, one HMM was created with the same characteristics described above to model the silences.

The sound classes are simply listed in a text file; their number was determined by the manual classification. The number and names of the classes was different for units and subunits, given that in some instances a unit corresponded to two subunits.

The last part of the model definition deals with creating a definition file which describes the relationship (or 'network') between the HMMs created. For the purpose of this study we compared two model structures, as depicted in *Figure 6.6*:

a) Unit model based on their definition which states that a unit is a continuous sound between two silences (Payne and McVay, 1971);
b) Subunit model based on the idea that one unit can be divided into smaller components where marked frequency changes can be observed (Pace *et al.*, 2010; Pace *et al.*, 2011; Shapiro *et al.*, 2011).

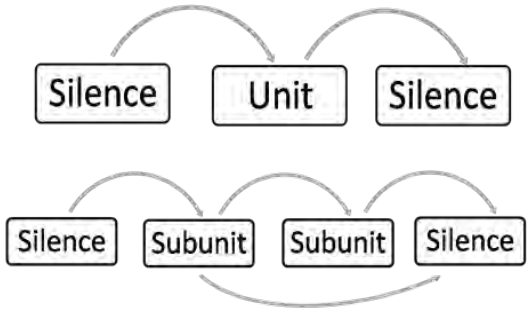


Figure 6.6: Diagram showing the two model structures used for classyfing the vocalisations present in the humpback whale songs analysed in this thesis. The first model is based on recognition of units (top) where in each segmented recording one could find a unit with silent portions before or after the call. The alternative model (bottom) is based on subunit recognition, which means that each call between two silences could be represented by one or more subportions (in the model depicted above the maximum number of subunits per unit was 2). Because the second model allows skipping one state, the recogniser could go from the first subunit to silence directly, in which case the subunit model is equivalent to a unit (Figure 6.7).

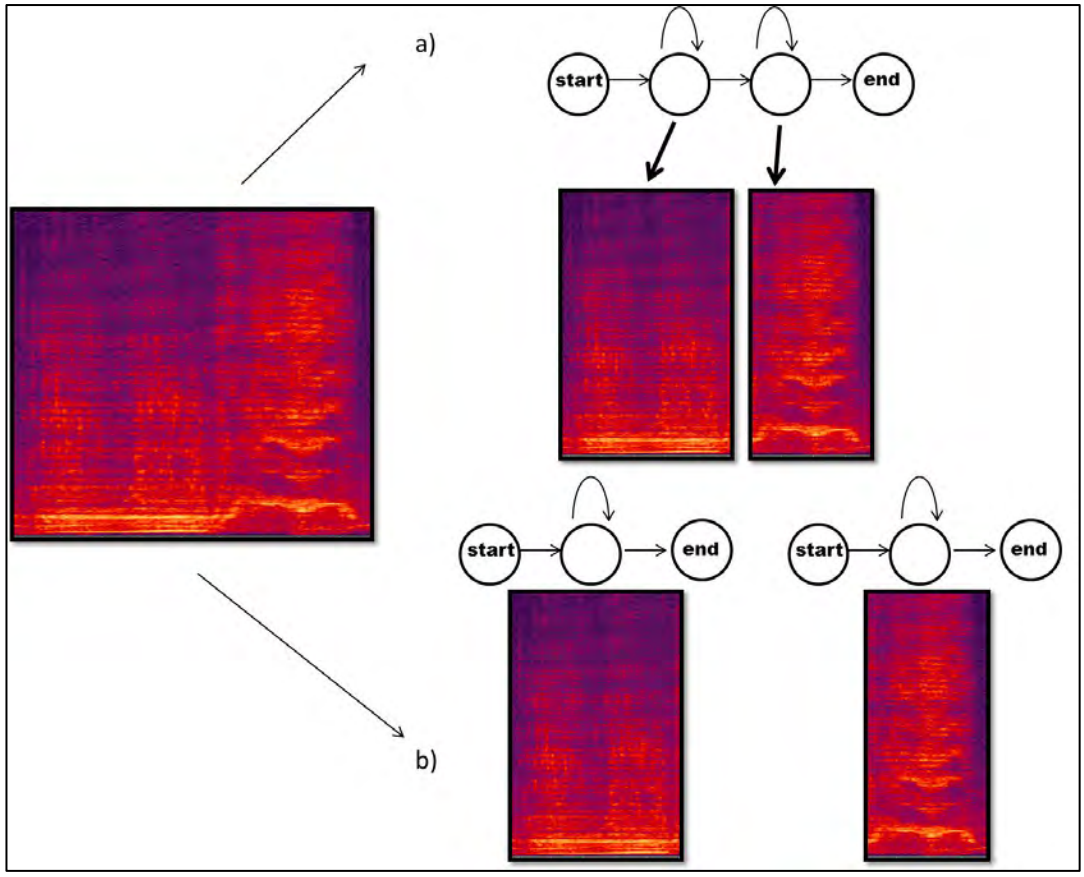


Figure 6.7: Spectrogram of a sample humpback whale vocalisation encountered in a recording of a 2009 Madagascar song. The diagrams show the two different hidden Markov Models that were employed for the classification task. The vocalisation could be modelled as a unit between two silences (a) where changes in the call's characteristics are captured by shifting from one state to the next in the left-to-right HMM. Alternatively, the call was split into two subunits (b) if there was a marked shift in frequency and the conditions explained in previous chapters were encountered. Each subunit was modelled through one single state HMM.

The network of the unit model was therefore, the network was SILENCE-UNIT-SILENCE; whilst the network for the subunit model was SILENCE-SUBUNIT (and) SUBUNIT–SILENCE. The model allowed for a maximum of two subunits per unit because in the recordings analysed here units made up of three or more subunits were not encountered. However each unit and subunits could contain multiple (up to three) states.

The models described here for humpback whale calls are analogous to Hidden Markov Models based on phonetic representation of speech. Whereas the representation of humpback whale songs through unit classification can be thought

of as constructing word models for speech recognition. Indeed, using the subunit approach one can represent more combinations of calls using fewer symbols, reducing the computational load and training effort. A catalogue of the units and subunits of the recording Mada09b can be found in Appendix 2, which will detail the number of states chosen for each Hidden Markov Model.

6.4.4 Model training

During the training phase a data base of labelled (manually classified) data was employed. The manual classification was performed by thethe author. In the case of the unit model, the units were automatically segmented and the manual classification was performed by listening to the acoustic data and spectrographic analysis. For a subunit model, the segmentation into subunits was also performed manually, as described in the previous chapter.

The labels for each training segment were written using the software Wavesurfer, developed by the KTH Royal Institute of Technology, Stockholm (Beskow and Sjolander, 2000) This software allows one to manually enter the labels for a sound segment and to save them in the HTK compatible format ".lab".

The training process is carried out by feeding the inputs described above in the HInit toolbox; this is designed to work out the transition probabilities of each HMM from one state to the next starting from the HMM prototypes and based on the training dataset. The successive step is a second estimation of the transition probabilities, which is done through the HRest toolbox; the resulting HMMs are then copied in a single file which is used in the recognition process.

The training stage deals with calculating the maximum likelihood estimates (MLE) of the transition probabilities matrix of the states. In practice, this means that starting from a prototype HMM after the training process one obtains a model whose mean, variance and transition probabilities are calculated based on the statistical properties of the data present in the training set. This is achieved in two steps:

i) The Viterbi algorithm (Forney, 1978) is used to find the most likely state sequence corresponding to each training sample;

ii) A Baum-Welch (Baum *et al.*, 1970) re-estimation is performed to find the probability of being in each state at each time frame using the *Forward-Backward* algorithm. This probability is then used to form weighted averages for the HMM parameters. An excellent review of the use of HMMs is provided by Rabiner (1989).

6.4.5 Recognition

The recognition stage requires one to specify the network of the model, i.e. the structure of the elements contained in the files are fed into the system for recognition to allow silences at the start and/or end of the file, and a list of the HMMs obtained after the training stage. This study evaluates the performance of HMMs for the classification of four humpback whale songs and compares the classification performance of the unit model and the subunit model, which were described in the previous sections.

Each vocalisation to be recognised is stored in a single file and converted into MFCCs. A Viterbi alignment (Viterbi, 1967) was performed to match each call of the testing dataset the best matching HMM. The output of the HMM recognition is a text file listing class to which each MFCC file is allocated. The results obtained were compared to the manual classification and the correct classification rate computed as a percentage.

The next chapters will present the results of the recognition carried out on several recordings and discussing the HMM performance with different amounts of training.

6.5 Conclusions

Whilst Hidden Markov Models have been extensively used for speech recognition purposes, they have been applied to signals produced by animals on limited occasions. This is partly because HMMs are more complex than other signal processing tools that are more accessible to animal biologists who may want to compare and/or interpret sounds produced by different species and partly because animal species often produce either few sounds or very stereotyped ones that can be easily classified using other automatic classifiers that do not require as much variability built into the model.

In this chapter, steps for designing HMMs were described, which are suitable to deal with the hierarchical structure of humpback whale songs. HMMs can be adapted to recognise song elements from the basic building blocks to more complex sequences, such as phrases, because they are very flexible. They could be adapted to classify sounds emitted also by other marine mammal species, for example for dolphin whistles and killer whale vocalisation, which are variable in duration. Results on the

application of HMMs to classify the building blocks of humpback whale songs recorded in Madagascar are presented in Chapter 7.

7. Madagascar song analysis

The songs of humpback whales that breed in Madagascar waters have been poorly described so far in comparison with those emitted by conspecifics that winter around the coasts of Australia and in waters around Hawaii and Mexico.

A first description of Madagascar song was published by researchers of IWC (Razafindrakoto, 2001) who recorded songs in Antongil Bay in 1996, a bay in the North East of Madagascar, not far from the Island of Ste Marie where data were collected for this thesis. Since that study, no other description of humpback whale songs around Madagascar has been published. Considering that humpback whale songs evolve considerably from one season to another, a record of songs recorded 15 years ago during two weeks of one field season cannot be considered representative of the dynamics of the population that comes to the area to reproduce on an yearly basis.

In this chapter we describe songs of humpback whales recorded in the Ste Marie channel, as described in Chapter 4, over four years. The evolution of songs over this time will be presented, as well as an in depth analysis of the units and subunits that constitute the songs in question. Finally, the performance of the automatic classification using Hidden Markov Models (HMMs) will be assessed and compared to the manual classification conducted.

7.1 Madagascar song description

All the songs recorded in Madagascar between 2007 and 2011 were visually and acoustically inspected to determine the song structure each year. According to the literature, all humpback whales on the same breeding ground sing the same song during the same period of the breeding season. Therefore, we expected all recordings to contain the same song with slight variations of the themes that could be due to individual variability and minor evolution of the song within the breeding season. It has been shown that song structure progresses throughout the season resulting in the song sang at the start of the season being slightly different from the one sang at the end of the breeding season (Winn *et al.*, 1981; Cerchio *et al.*, 2001b).

The themes present in the songs were identified each year and the song sequence determined using more than one recording for each year to corroborate the analysis.

An overview of the songs used for understanding the song structure is given in the table below (*Table 7.1*).

Year	**Number of singers**	**Number of song cycles**	**Themes in each song**
2007	1	3	6
2008	2	5	5
2009	5	6	4
2010	4	9	4
2011	2	6	4

Table 7.1: Number of themes encountered in the songs recorded each year from 2007 to 2011.

The number of singers corresponds to the different days the recordings were taken because it is assumed that each day that in the field a different is recorded. Given the number of animals present in the Ste Marie Channel during the winter, it is unlikely that we recorded the same individual on different days. The number of song cycles refers to the numbers of entire repetitions of the theme sequence, according to the humpback whale song definition described in Chapter 2. In some recordings, a song was sang more than once and as a result the number of songs analysed is greater than the number of singers. Studying more than one song cycles is necessary to identify individual variation within themes and to understand where the song started. Indeed, very rarely does the start of the recording match the start of the actual song cycle.

The structure of the songs recorded in the channel of Ste Marie between 2007 and 2009 - which are used for further analysis and the automatic classification task - is given in *Figures 7.1-7.3*.

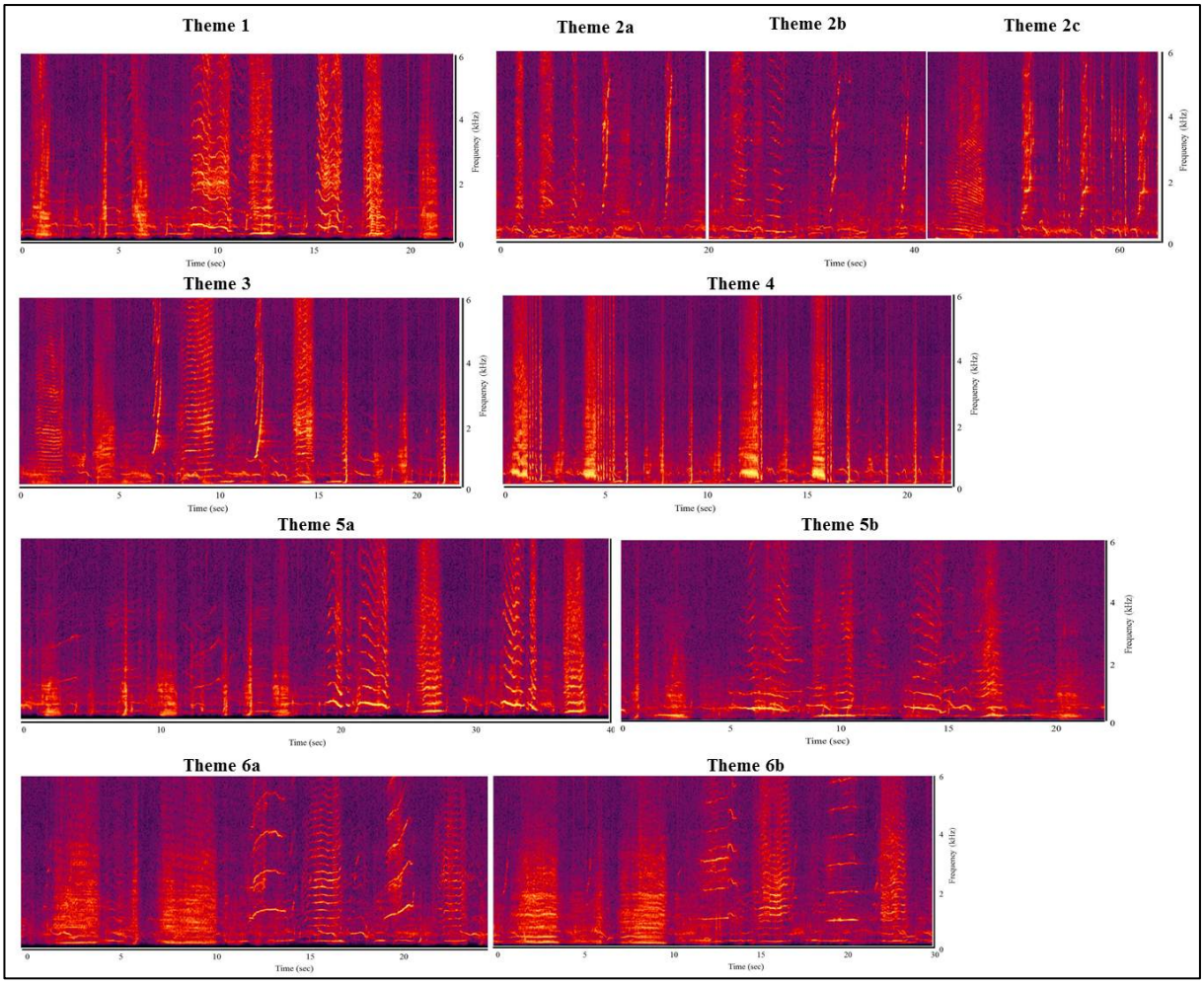


Figure 7.1: Spectrograms showing the sequence of themes in the 2007 song in the channel of Ste Marie in Madagascar at the end of July 2007. Six distinct themes were identified, 3 of which presented slight variations in their unit components or in the number of repetitions of a particular vocalisation throughout the song sequence. The variations observed in the phrases that compose the same theme are progressions of the song sang by a single singer rather than differences deriving from inter-individual variations.

Only the song cycles sung by one singer were analysed for 2007 because only this one recording was made available for analysis. Nonetheless, this was sufficient to determine the full song structure for that year because the recording length was approximately one hour during which the same song was repeated three times. It is possible that more variation of themes occurred during that year and that the themes represented in *Figure 7.1*.These themes may not represent the entire repertoire of the population for that year. In addition, given our knowledge of the singing behaviour of humpback whales and the data of the following year we can assume that the song

sequence presented above is representative of the song cycles sung during July 2007 by all the singers present in the Ste Marie channel at that time.

Of the 12 days of recording carried out during August 2008, only the songs recorded on two separate days were used to determine the song sequence for the 2008 season because most of the remaining dataset included songs where there was extensive overlap with other singers, making it impossible to confidently determine the song sequence. The 2008 song was different from that of the previous year, as expected, considering that humpback whale songs evolve through time. Two themes were shared between 2007 and 2008 (*Figure 7.2*).

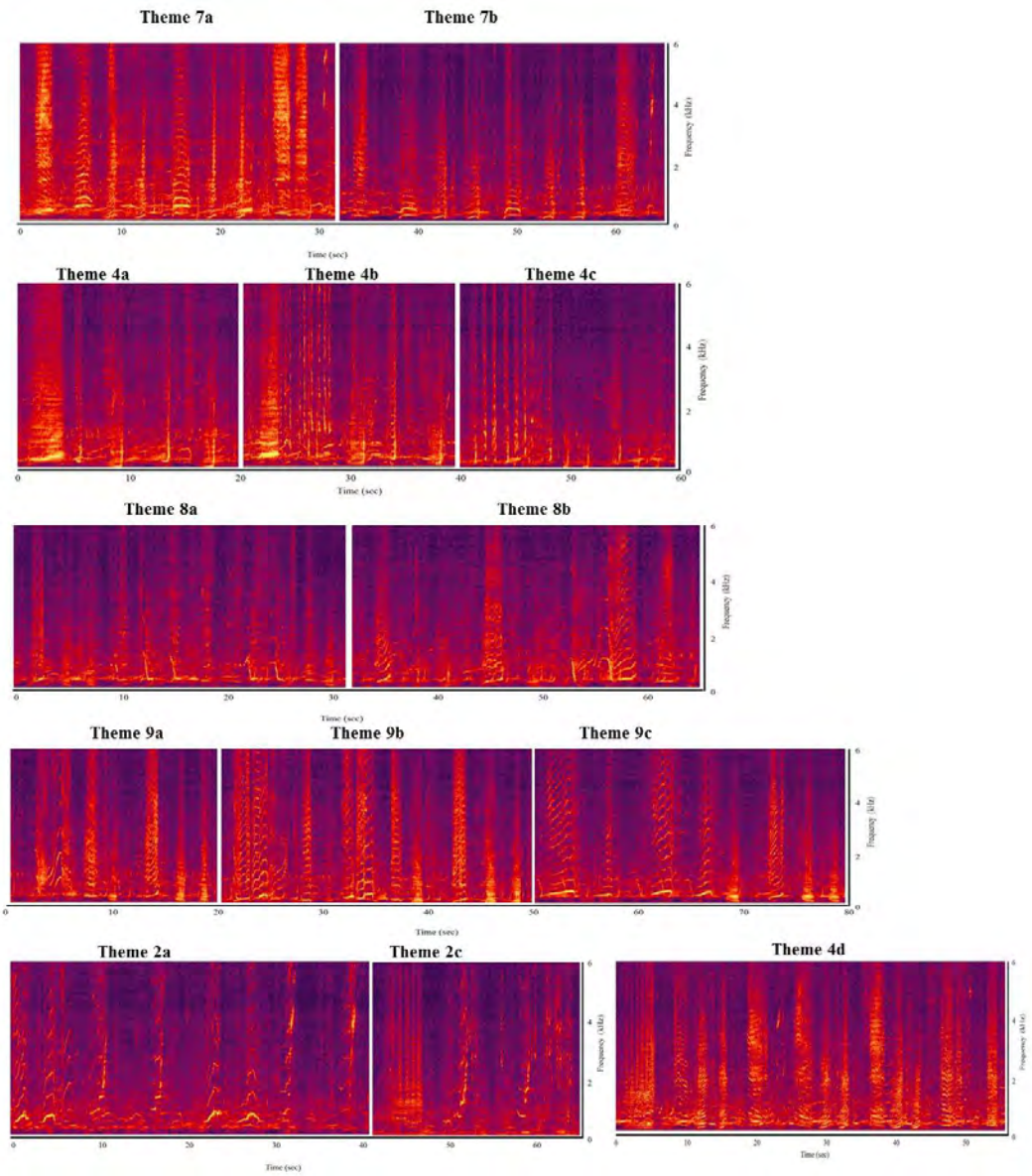


Figure 7.2: Spectrograms showing the sequence of themes in the 2008 song sung by humpback whales in the channel of Ste Marie in Madagascar in August 2008.

The song of 2008 was constituted by 5 themes, with one theme (i.e. theme 4) evolving through the song to assume different forms and being repeated at the end of the song cycle. This theme was present in the 2007 song but only in one form. The fact that only one form of this theme was observed in the recording of the previous year may be attributable to the paucity of data available or to the learning ability of humpback whales that might have taken the basic theme of the previous year and started making it more complex. Evolution of themes from songs of different years have been observed in other parts of the world, for instance in Australia (Garland *et al.*, 2011). The song of 2009 was shorter than the previous two and was composed of only 4 themes, 2 of which were shared with the song of 2008 (*Figure 7.3*).

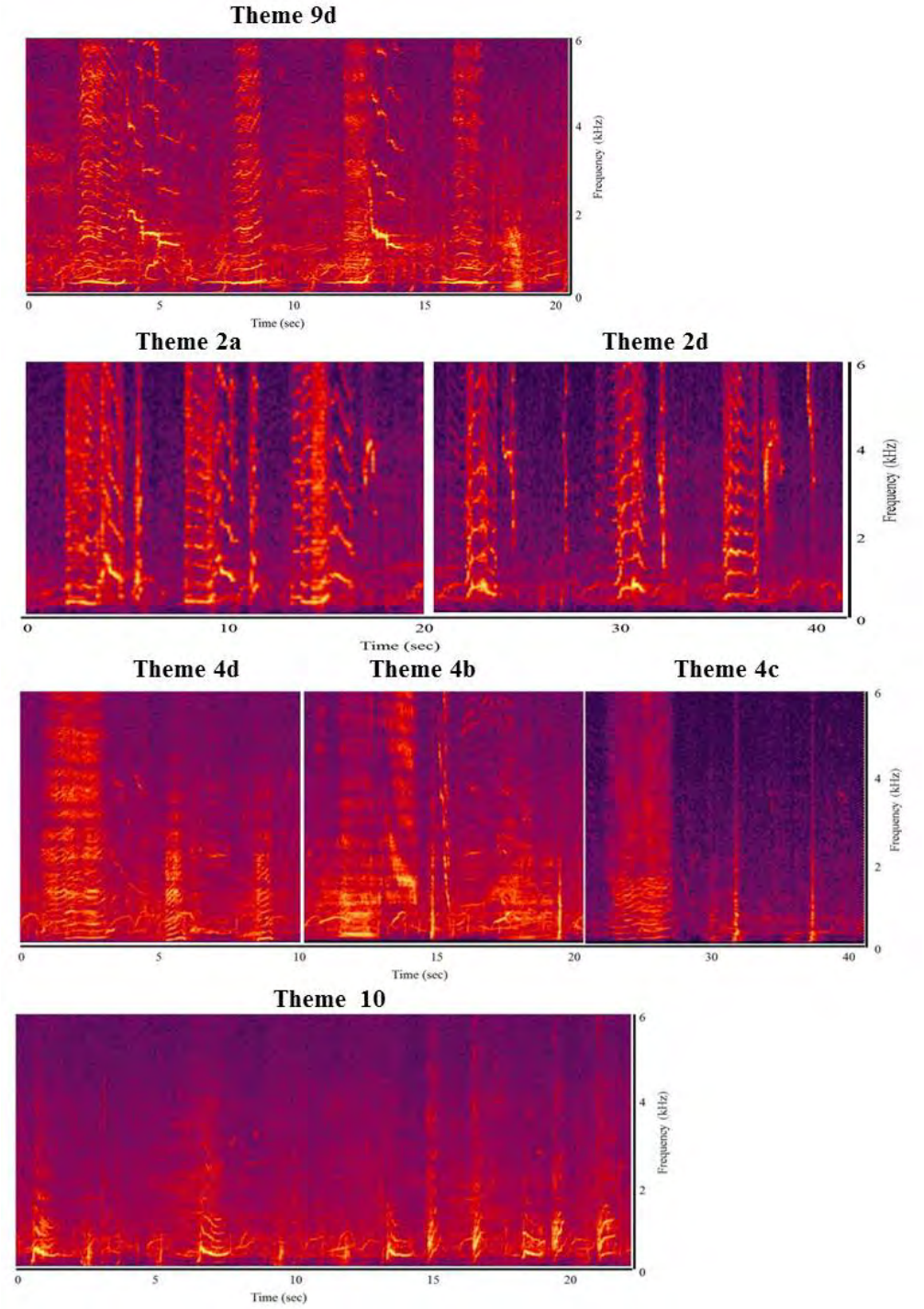


Figure 7.3: Spectrograms showing the sequence of themes in the 2009 song sung by humpback whales in the channel of Ste Marie in Madagascar in August 2009.

The full structure of the songs of following years is presented later in the chapter for comparison purposes but units from these recordings were not in the analysis of the automatic classification because the data of 2010 was of extremely poor quality. This was due to the field season taking place at the end of August when very few whales

were present on site. This was extremely unusual as humpback whales are commonly observed in the Ste Marie Channel until the beginning of October.

7.2 Analysis of song structure

In the previous section, the general structure was described of the humpback whale populations that breed in the area of Ste Marie. This description was conducted on the theme level of the song which is commonly used by biologists to compare populations across ocean basins (Cerchio *et al.*, 2001b; Garland *et al.*, 2011). The basic building blocks are not generally compared across populations or at least they are not analysed into great detail because it is a very time-consuming process to analyse each unit component of a song and it is also challenging because most units tend to change slightly throughout a song and between individuals as the song evolves. Therefore, the context of the themes in which these units are found needs to be taken into account to ensure that one does not overestimate the number of unit classes based on the fact that minor changes of such vocalisations occur as the song progresses. A catalogue of the units found in the songs of 2009 is presented in *Figure 7.4 (more details are given in Appendix 2).*

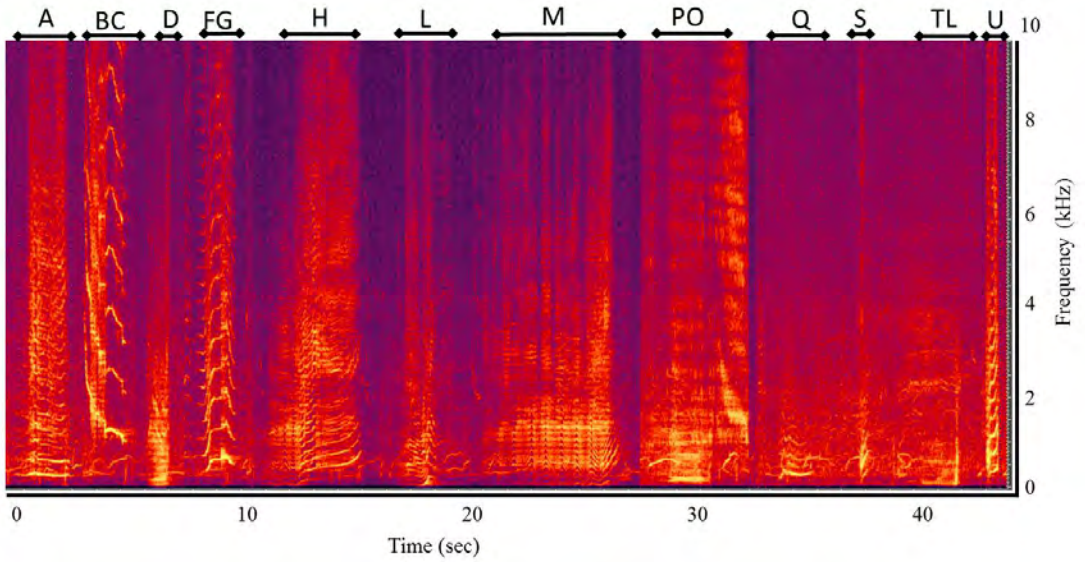


Figure 7.4: Spectrogram of an artificial sequence formed by concatenating one sample of each of the units found in the 2009 song. Multiple samples of each of these classes were used for the training stage of the automatic classification. The classes are named sequentially following the alphabet according to the order in which they were found in the song sequence. The missing letters in the sequence are a consequence of the fact that there were some unit classes that had too few samples to carry out the manual classification and therefore were excluded from the analysis. Sound units with double lettering indicate that this unit can be broken down into two subunits.

The spectrogram above summarises the units that were found in the 2009 song which were used as the starting point for the analysis and automatic recognition performance assessment of the thesis. The labels of the classes used letters and proceeded alphabetically. Four of the units shown above can be split into 2 subunits where there is a marked frequency shift, and the corresponding subunit classes are labelled with a single letter. For instance, unit class 'BC' includes subunit 'B' and subunit 'C'. This catalogue of units and subunits is at the basis of all further analysis and most of these classes are present in the songs of 2007 and 2008, as will be described in the following section.

As one can see from the spectrogram above, vocalisations vary considerably in their frequency characteristics as well as in their duration. Whilst the frequency characteristic, specifically the fundamental frequency and harmonic structure, varies little between units within a song, the duration of the calls can vary considerably as

the song evolves and the singer starts repeating the same sequence of units. A summary of the distribution of the duration of units (*Figure 7.5*) and subunits (*Figure 7.6*) that occurred in the 2009 song is given in below (*Table 7.2*).

Class	Duration (sec)				Frequency (Hz)
	Mean	St Dev	max	min	
A	1.85	0.10	2.71	1.02	290
B	1.03	0.12	1.83	0.34	2500-1000
C	0.68	0.04	0.76	0.47	1000
D	0.79	0.03	0.89	0.69	100
F	0.63	0.13	1.31	0.25	370
G	0.88	0.09	1.31	0.37	990
H	2.61	0.22	3.39	1.77	190
L	0.34	0.02	1.37	0.15	62-400
M	1.92	0.44	4.28	0.69	160
O	1.98	0.16	2.47	1.61	1500
P	2.07	0.13	2.71	0.99	706
Q	1.04	0.07	1.41	0.63	120-360
S	0.53	0.03	0.70	0.39	90-520
T	0.60	0.06	0.75	0.43	120
U	0.45	0.03	0.58	0.27	2500
BC	0.84	0.86	2.59	1.42	2500-1000
NL	0.50	1.14	1.70	1.35	90-280
PO	2.61	2.90	4.61	4.16	252-1643
TL	2.16	2.26	2.97	2.51	120-450
FG	1.53	1.59	2.27	1.91	370-990

*Table 7.2: Table summarising the duration of the vocalisations in each sound class for both units and subunits and the fundamental frequency of the calls. The fundamental frequency of the calls was estimated from the call's spectrograms using the software Raven Lite developed by Cornell Lab of Ornithology. If the fundamental frequency of the calls changed throughout their duration, the average initial and final frequencies are given in the table. For unit/subunit 'M' the * means that it was not possible to estimate a fundamental frequency, rather the band in which most energy of this broadband call is contained is given in the table.*

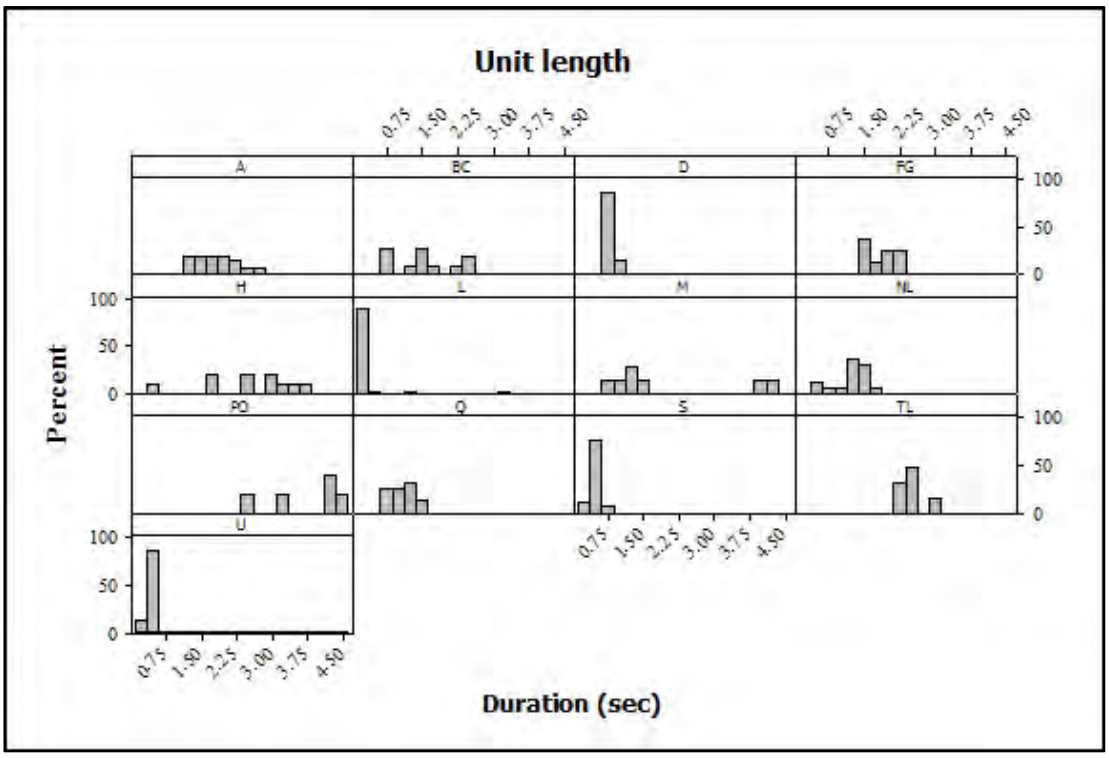


Figure 7.5: Histograms showing the duration of each unit class as a percentage of the total number of units encountered in the recording. Details on the number of calls in each sound class are given in the tables in the next section.

As detailed in the figures above, the duration of units ranged between 00.15 s to 4.6 s for the longest compound units. Some of the units were more variable than others in their duration, for instance, unit 'L' was very short, i.e. less than 1 second, in nearly all the instances, whereas others ranged in length considerably. Indeed, some of the common calls are repeated several times, and as the song progresses they become longer, with more distinct features. The same is true for subunits, whose duration is quite variable for some of the harmonic and broadband calls (*Figure 7.6*).

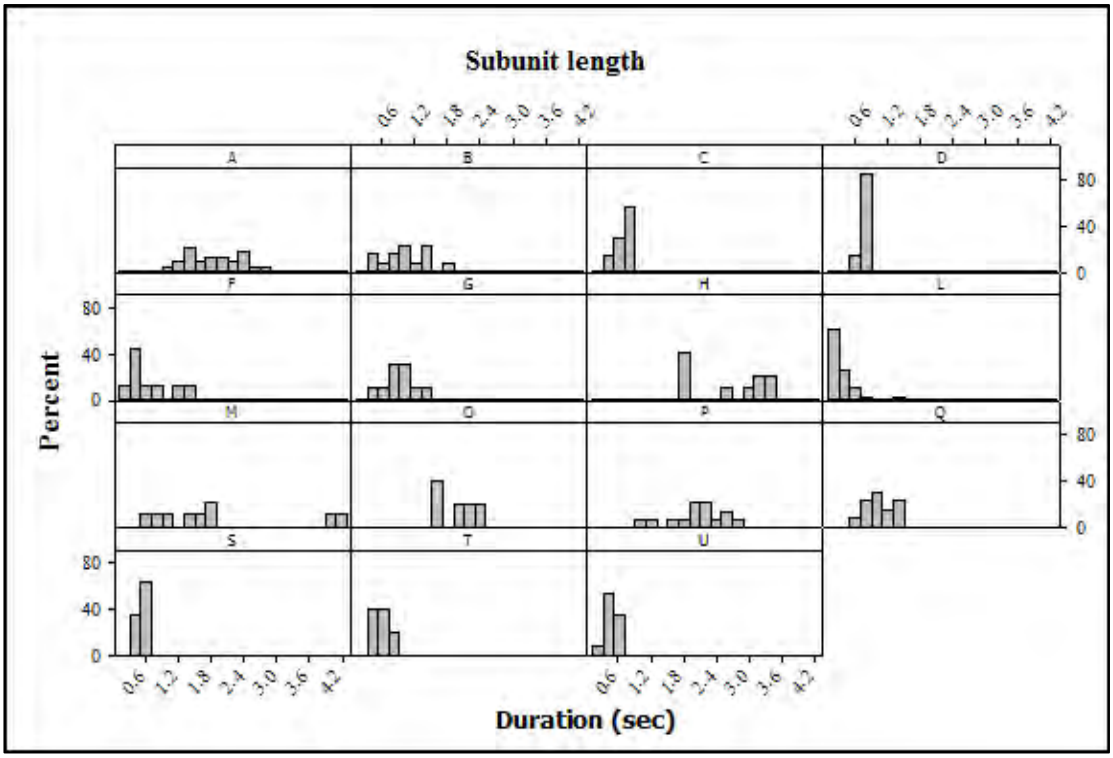


Figure 7.6: Histograms showing the duration of each subunit class as a percentage of the total number of units encountered in the recording. Details on the number of calls in each sound class are given in the tables in the next section.

Overall, as expected, subunits were shorter than units in duration, with more calls being distributed in the frames below and around 1 second.

The results of the classification performance and how this is affected by the features of different calls are presented in the next section.

7.3 Automatic classification performance

The automatic segmentation of the songs was checked against their manual segmentation to determine the efficiency of the detection algorithm, and the efficiency of the HMM classification was determined by comparison with manual classification.

7.3.1 Classification of songs per year

Initially, we looked at the manual classification of a single song recorded at 08:50 am on the 12th August 2009. For each class of units we used half of the calls chosen at random to train the model, and the remaining half for the testing. Six classes were removed from the analysis because they included less than 5 calls, too small a sample to carry out both the training and testing, given that the minimum input number for the training set had to be 3 calls.

The results obtained are presented in the tables and graph below and compared to the subunits model (*Table 7.3, Table 7.4, Table 7.5 and Figure 7.7*). The rows of the confusion matrices presented below show the input calls whereas the columns represent the outputs.

Output Input	A	B	C	D	F	G	H	L	M	O	P	Q	S	T	U
A	24														
B		13													
C			7												
D				6							1				
F					8	1									
G						10									
H							7				3				
L								70							
M							1		7		1				
O									1	4					
P				1							12	1			
Q												14			
S													11		
T														5	
U		1													10

Table 7.3: Confusion matrix of the results of the HMM classification using the subunit model.

Output Input	A	BC	D	FG	H	L	M	NL	PO	Q	S	TL	U
A	32												
BC		11											
D			7										
FG				8									
H					10								
L						32		1					
M					1		6						
NL							2	14					
PO									5				
Q										15			
S				3							20		
TL												6	
U		1											7

Table 7.4: Confusion matrix of the results of the HMM classification using the unit model.

subunits			**units**		
class	**Total tested**	**correct**	**class**	**total tested**	**correct**
A	24	24	**A**	32	31
B	13	13	**BC**	11	11
C	7	7	**D**	7	6
D	7	6	**FG**	8	8
F	9	8	**H**	10	10
G	10	10	**L**	49	46
H	10	7	**M**	7	6
L	70	70	**NL**	16	14
M	9	7	**PO**	5	5
O	5	4	**Q**	15	15
P	14	12	**S**	23	20
Q	14	14	**TL**	6	6
S	11	11	**U**	8	7
T	5	5			
U	11	10			
Overall	219	208	**Overall**	181	171

Table 7.5: Total number of subunits and units tested and performance of the classification method, also given in percentages.

The performance of both models is very high, as expected given that the model was trained using vocalisations of the same recording albeit that the model was trained and tested on different instances of each unit. Despite the fact that some classes occur much more often than others, the classifier has grouped together calls that are similar to each other. This property is very important because on several occasion, some calls can be observed frequently within a song, whilst others are quite rare. The overall performance of the two models (the unit and subunit models), given in *Figure 7.7*, is nearly equal (95% for subunits and 94.5% for the units).

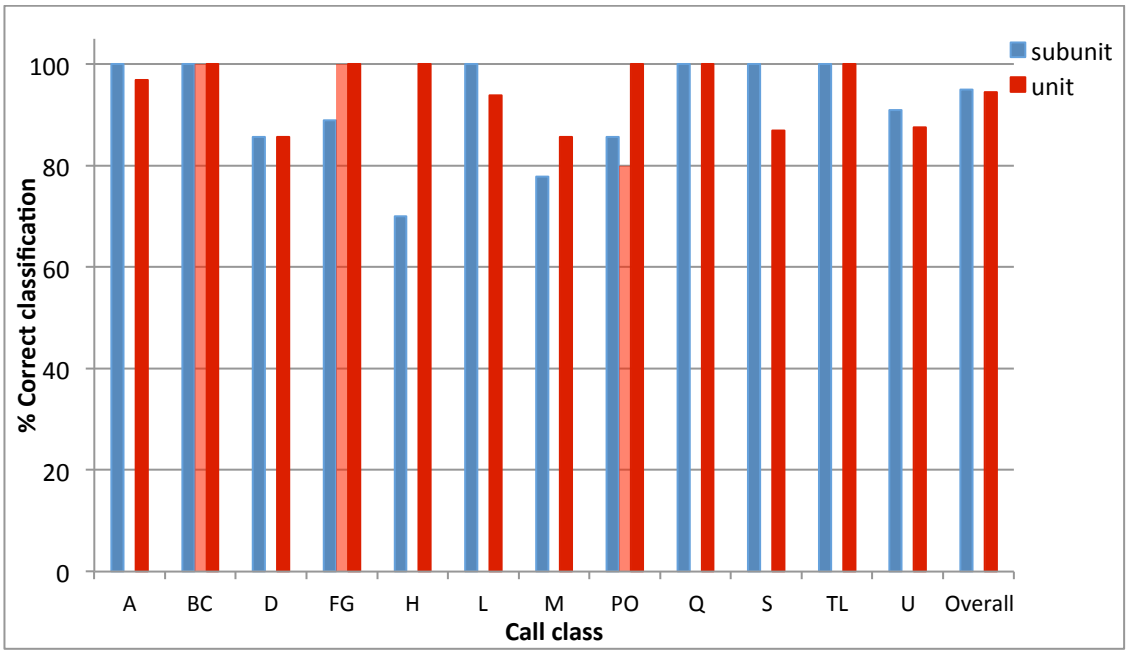


Figure 7.7: Percentage of correctly classified calls using the unit vs subunit model for Mada09a recording. Note that the light red shade corresponds to a second subunit, and by putting together the dark and light red columns we obtain the corresponding unit. In the case of the unit 'TL', the column of the subunit corresponds to subunit 'T' and the blue column corresponds to the unit 'TL'

The unit model performed better for units 'H', 'M' and 'PO', all of which are 'unvoiced' type calls and are not very frequent. The reason of this difference in the performance is not well understood; however, it seems reasonable to suppose that it has to do with the broadband nature of these signals. For instance, the unit 'PO' is composed by two broadband subunits, i.e. subunit 'P' and subunit 'O', the latter of which presents very similar characteristics to subunit 'M'; hence, it is not surprising that the classifier would sometimes exhibit confusion between a 'M' call and a 'O'

call, whilst this doesn't happen when using the unit model since the combination 'PO' is more distinct from 'M'.

It is not surprising that the unit model performs as well as the subunit model in the context of a single song where the training and testing are based on the calls emitted by the same singer over a short period of time. To develop a system which is repeatable across different animals and over different seasons such variability must be included. It is anticipated that the model based on subunits should be better for this task because whilst these small blocks can be associated in different ways by different singers to form a broad vocabulary of units, expectation is that they will be more or less constant in number and not change from year to year.

A cross validation was also performed to measure the variability of the model based on different training and test sets. A ten-fold validation was carried out, which means that during each round of tests, 90 % of the data was used for training the models while the remaining 10% of the data were tested. To conduct this test, all the data was arranged randomly and, each time, a new set containing 10% of the calls (31 or 32 calls) was tested working sequentially through the whole data series. The average classification performance for each unit and subunit obtained from the ten-fold validation test is presented in Figure 7.9*,* whilst detailed results of each test are presented in the tables in Appendix 3.

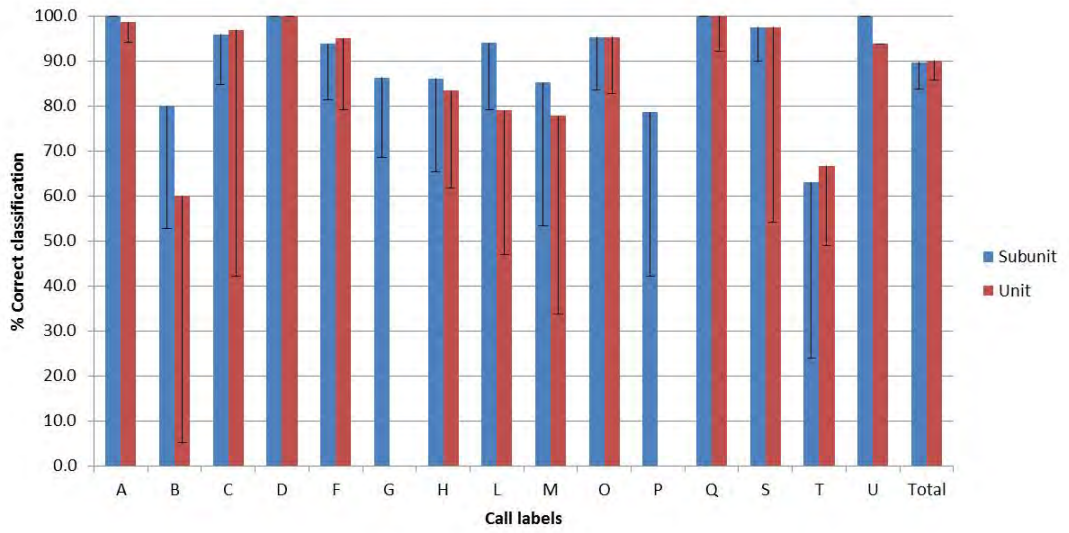


Figure 7.8: Percentage of correctly classified calls using the unit vs subunit model for Mada09a recording during the cross-validation test. The error bars represent theThe standard deviation based on the algorithm performance obtained from 10 repetitions of randomly sampled training and testing datasets. Error bars are not presented in future tests as it was assumed that multiple repetitions of randomly chosen dataset would give similar error levels. The standard deviation (negative value) of the round of testing is shown as a bar on each column. The call labels depicted on the x-axis represent the subunit call label rather than the unit call label. When units are present as a combination of subunits, then the result value is indicated under the label of one of the corresponding unit. For Unit FG the result is presented as bar F, for unit OP in column O, for unit BC in column C and for unit TL in column T.

The overall classification performance of the cross-validation test, where 90% of the data was used to train the models, is very similar to that presented in Figure 7.7, where only 50% of the dataset for each sound class was used to train the model. This applies both to units and subunits. In addition, the standard deviation shows little variation across tests in terms of overall classification (both unit and subunit model standard deviation is less than 5%). However, looking at the classification performance of each call class shows that some units and subunits have different classification performance than in the previous test that was presented in Figure 7.7. In particular, call classes "D", "H" and "M" appear to be more often recognised correctly when using the 90% training test, particularly when classified using the

subunit model. Interestingly, all these calls are broadband vocalisations. This may suggest that some call types may need more training samples than other classes to be able to obtain high levels of correct classification. The effect of different size training sets will be discussed more in detailed in Section 7.3.
The variability in performance between each test round for the cross-validation shows that in all instances but one the subunit model results are more consistent across tests.

The low performance for subunit/unit "B" compared to the previous test, and the large standard deviation, may be a result of the fact that there are very few samples of this call occurring on its own during the recording and, therefore, very few instances were present in the testing set each round, if any. Hence, for the unit model the outcome of the test for this call could only be either 100% or 0% correct classification.

After testing consistency of the classification method for a single recording, i.e. a single song session produced by one singer, we wanted to check how the classification algorithm would perform on other recordings to know whether it could effectively be a practical tool for biologists who may want to compare songs of humpback whales.

A stepped approach was followed to introduce variation in the model gradually. Therefore, the first round of testing on a different recording was conducted on a song of the same year. Choosing a song from the same year and time-period than the one analysed previously means that the same units and subunits will be encountered because research showed that all singers on the same breeding grounds sing the same song during one season. This means that variation is introduced only by the song being sung by another singer and by the recording quality (in terms of its SNR).

Both the unit and subunit models were tested on a different recording from the same year – taken on the 2nd of August 2009. The training set used is the same as the previous test (*Table 7.6* and *Figure 7.9*).

	Subunits		Units	
Class	**Total n.**	**Correct**	**Total n.**	**Correct**
A	11	10	11	11
B(BC)	4	3	4	3
C	4	3	-	-
D	4	3	4	1
G	17	17	17	17
H	6	1	6	1
L	11	11	9	5
Q	10	9	10	0
U	7	7	7	6
Overall	74	64	68	44

Table 7.6: List of subunits and units encountered in the recording and number of correctly classified calls in each case.

The number of calls present in this recording is small compared to the previous dataset because the length of the recording was just 15 minutes. Indeed, after the first song session the whale swam away from the boat and no further recording was possible for this individual. The overall performance for the unit model was above 606% and for the subunit model was 808% (*Figure 7.9*).

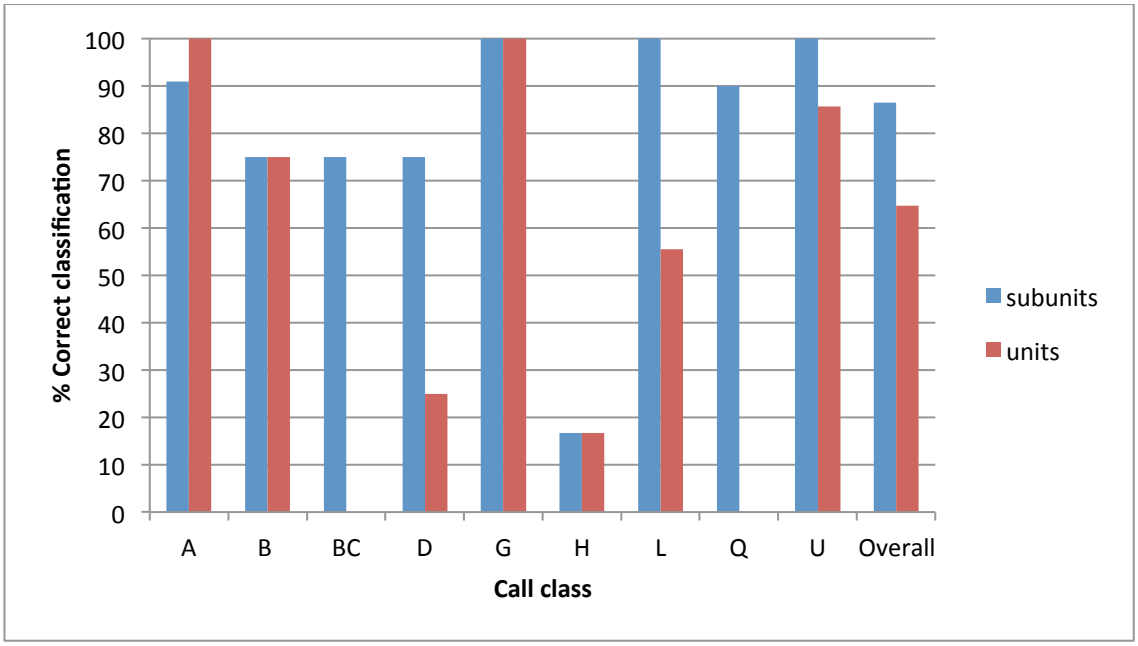


Figure 7.9: Percentage of correctly classified calls according to the subunit vs unit model for each class and overalls. Note that the performance for the recognition of 'Q' is 0 for the unit model because in this recording we always encountered 'Q' in association with other subunits.

The low classification performance for call 'H' with both models is attributed to the fact that there was boat noise when these calls occurred, and the frequency content of the boat noise overlapped that of the whale song. The result was that this call was clustered with other vocalisations.

Although the sample size for this recording was quite limited, the results showed that the classification method can be applied to the songs emitted by other individuals and that, in accordance with our hypothesis, the subunit model performed better than the unit model. The major difference in the results is attributable to the performance in the classification of the class 'Q' calls (90% versus 0%). In this case, the poor performance of the unit model might be due to the fact that 'Q' was not found on its own in the second recording but as part of a new unit formed by subunits 'Q' and 'A' and in all instances this was classified as 'A', perhaps because the last section of 'Q' is similar to 'A' in that its fundamental frequency is the same as the first harmonic of 'A'. This example emphasises the importance of having a more flexible system for the classification task. If we updated the training set of the unit model to include a new unit formed by 'Q' and 'A' together we would expect to improve the

performance of the classification method significantly. Indeed, the subunit model recognised both ‘Q’ and ‘A’ in the unit ‘QA’.

The robustness of the model was further tested on a recording of the previous year 2008, which was trained using only the training dataset of 2009 to start with because we wanted to test whether the subunit model performed differently from the unit model provided that from year to year new units are incorporated in a song. We would expect that subunits change less rapidly within the repertoire because they can be associated in different combinations to form novel units.

The recoding from August 2008 analysed here has a high SNR in the first 10-15 minutes, but it then decreases during the remaining 15 minutes probably as a result of the boat drifting away from the focal singer. In fact, during the second part of the recording there are a lot of overlapping calls due to the presence of multiple singers in the range of the hydrophone. Once again the training set for the model was the same as the one used in the first test but this had to be updated to include unit ‘B’ and unit ‘GF’. The results of this analysis are presented in *Tables 7.7* and *7.8* and *Figure 7.10.*

Output Input	**A**	**B**	**C**	**D**	**F**	**G**	**H**	**L**	**M**	**P**	**Q**	**S**	**T**	**U**
A	11								1		2			
B		3												
D				17				10	7	16	1		1	
F					8							1		
G			6			18					2	4		
H							3							
L	6		2				1	27	4	10	8		2	
M									6	2				
P	1	1							3	27	6			
Q	1						1				1			
S												2		
U														5

Table 7.7: Confusion matrix of the results obtained for the 2008 recording with the subunit model. The rows of the matrix represent the input subunits whereas the columns represent the ouputs.

Output Input	A	B	C	D	FG	GF	H	L	M	NL	PO	Q	S	TL	U
A	4						4				6				
B		1												2	
D				6				6	4	26	10				
FG					6					1			2		
GF						8									
H							3								
L	5						10	16		4		24	1		
M				2			1		4					1	
P											0				
Q												1			
S					2								0		
U								1				1			1

Table 7.8: Confusion matrix of the results obtained for the 2008 recording with the unit model. The input units are presented along the rows of the matrix and the outputs along the columns.

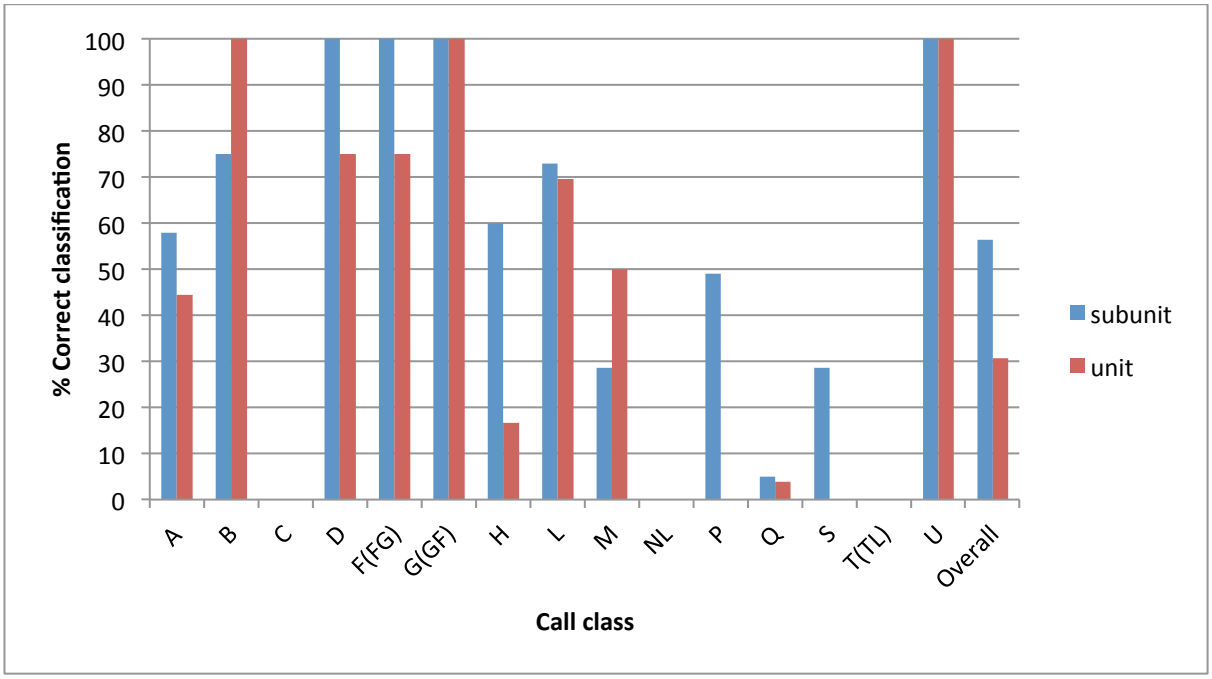


Figure 7.10: Percentage of correctly classified unit vs subunits for the song of 2008. 36 calls were removed from the analysis because they could not be manually classified by the author as they were overlapping with other calls. Note that a new unit class appears, namely 'GF' which is made up of exactly the same subunits ('F' and 'G') found in unit 'FG' but in reverse order. The percentage of calls that were correctly classified as belonging to subunit class 'F' are given in the column labelled 'GF' and the calls belonging to 'G' are given in the column labelled 'GF'.

The results give further demonstration that the model based on segmenting the songs into subunits is more robust than the one based on simply dividing the song into calls and silences. In this recording we had to include two brand new unit classes, which were just a different combination of the subunits observed in the previous recordings. To accommodate this, no adjustment of the model using subunits is required.

The overall drop in classification performance can be attributable to multiple reasons, which are all related to the fact that the training set used to conduct these tests incorporated only calls from a single song of one singer on a different year. This means that few samples were used to train the HMM for each call class accounting for little variability in the vocalisations. Variability needs to be taken into account because calls can vary greatly between individuals and also the ambient conditions from one recording to another may be different. Therefore, for the automatic classifier to be efficient, i.e. achieving a performance of 70% and above, one has to compile a robust training set. Later in the chapter, results will be presented of the classification performance using a more complete training dataset Before doing so, we present data on the same modelling shown above but testing the classifier on a song of 2007, which, theoretically, should be even more different from the song of 2009 than the 2008 song.

The recording of 2007 included 5 brand new units that were not encountered in previous songs. Therefore, training required updating the catalogue to include 50% of the calls found in the 2007 song. However, the remaining classes were tested without updating the training set with sample vocalisations emitted by the 2007 singer.

The results of the automatic classification carried out for the 2007 song are described below (*Figure 7.11*).

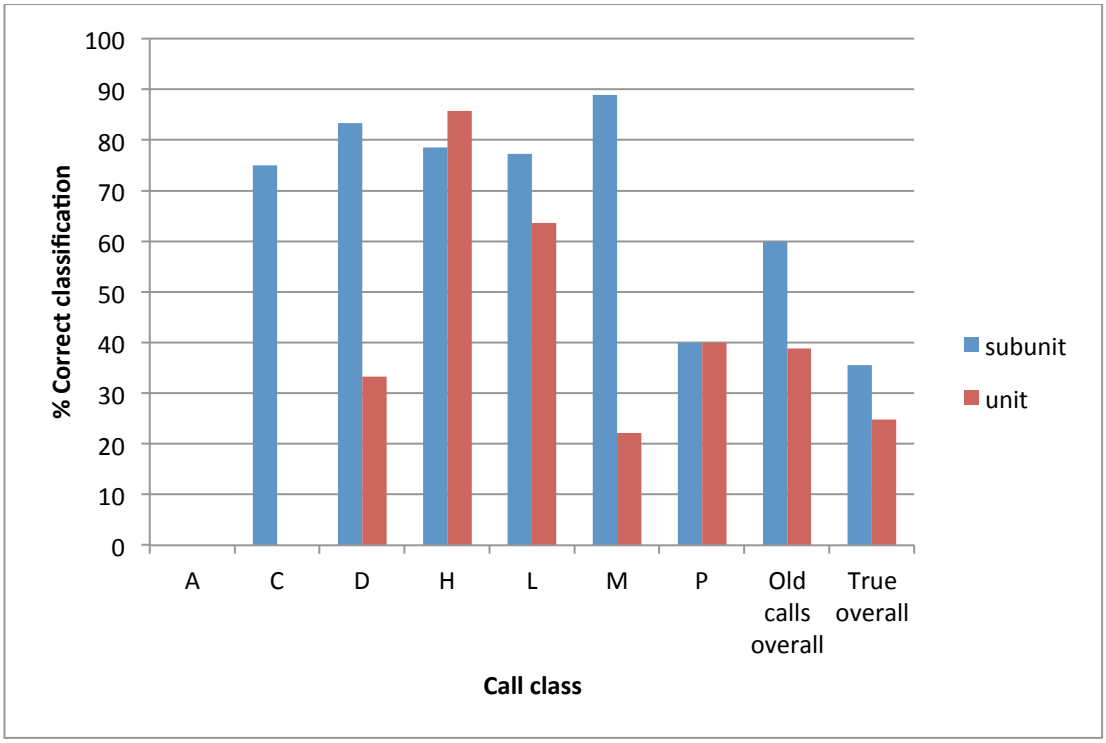


Figure 7.11: Percentage correct classification of the vocalisations present in the 2007 recording. Two overall values are given in this graph because a large proportion of the calls present in this recording is represented by calls that were not present in the previous recordings analysed. The 'old calls overall' represents the percentage of correctly classified calls based on the total number of calls belonging to classes that were present in the training set, whilst the 'true overall' was obtained by diving the correctly classified calls by the total number of calls present in the recording.

Overall, the classification using the subunit model was 10% better than using the unit model; the subunit model outperformed the unit model in all instances, except for one, specifically unit class 'H'H', which was trained using calls from the same singer. The medium performance of the classification overall is affected particularly by the fact that the calls in class 'A' were not classified correctly in any instance and these calls were quite common during the recording. Most of these vocalisations were mis-classified as unit/subunit 'H' instead, which is another sound with many harmonics. The main distinction between the two calls is that the fundamental frequency, which is lower for sound 'A' and that 'H' has a slight upsweep at the end. The fundamental frequency of 'H' is close to the first harmonic of 'A' and this fact might lead to mis-classification of the call considering that the 2007 song was sang

by a different singer and was more noisy than the 2009 recording on which the HMMs were trained.

7.3.2 Comparison of classification across years

In the previous section, results were presented for the classification performance using Hidden Markov Models that were trained on a small dataset of calls which were all produced by the same individual during one song of 2009. In this section, comparisons are drawn on the performance of the classifier across years and we present results of testing conducted after retraining the previous dataset with the addition of calls from all the songs analysed to account for more variability in the vocalisations. A summary of the recordings analysed for the songs analysed is given below, and comparisons are drawn for the overall classification performance using HMMs that were trained with 50% of the dataset (*Table 4.1)*.

The data from recording number 4, of 17 minutes, was manually classified into a total of 300 units and 369 subunits. These calls were randomly sampled so that 50% of the calls in each sound class were chosen as training set while the other half was used for the testing. The total training sample size was 119 units and 150 subunits. A consequence of this is that the minimum sample size for each sound class was set to 6, in which case 3 sample calls are used for training and the remaining 3 were testing. Intuitively, subunits and units that were completely new could not be recognised as they were not included in the 'vocabulary' of calls. Then the training sets were updated with the new calls appearing each year and the data were classified again.

The results of the overall classification for the 4 recordings analysed are presented below (*Figure 7.12*).

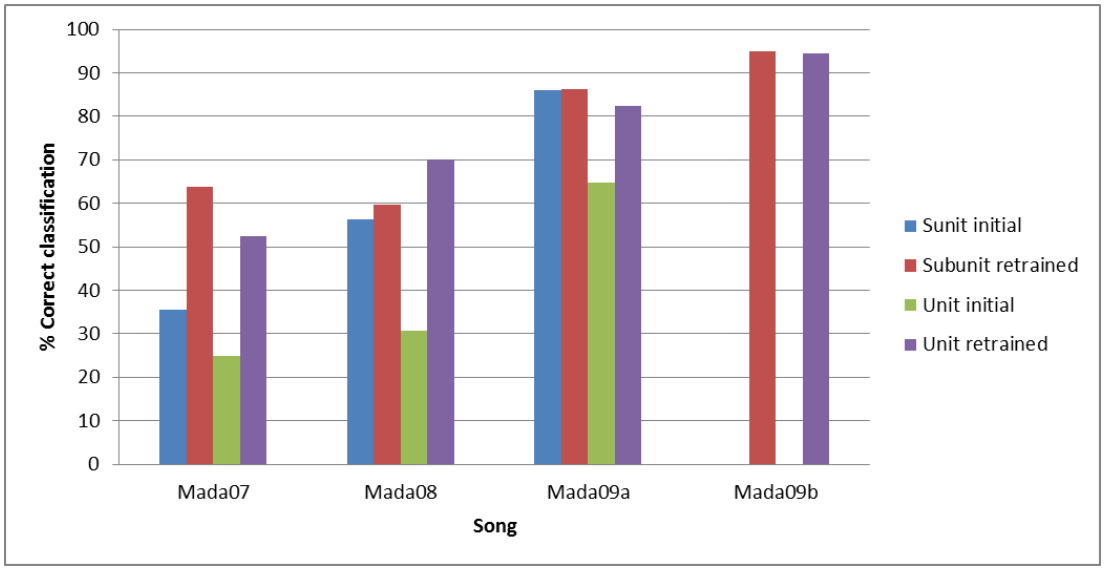


Figure 7.12: Correct classification rate as a percentage of the number of calls in the relevant recording. Results are shown for the Unit model and the Subunit (SU) model, for two conditions, "trained" when the training data includes samples from the specific recording, as well as data from recording 4 and "initial" when training is only performed using data from recording 4.

The recording Mada09b (12th August, 2009) was used for the initial training of the HMMs, therefore there is only one set of results for the subunit model and one for the unit model. The performance of the classification was extremely high and similar for both models. A further recording of 2009 was tested to check the performance of the models with a different singer, but expecting the calls to be similar to the previous recording given that humpback whales tend to copy each others' songs (Noad, *et al.*, 2000).

Intuitively, without updating the training set, the subunit model presents an advantage over the unit model in that it allows recognising new units that are formed by different combinations of previously known subunits. As expected, training using that year's data improved the classification in all instances and especially when analysing the recording of 2008 because a large percentage of the calls were new.

Overall the performance of the classifier decreased in the analysis of the 2 recordings of previous years, and particularly with the classification of calls in the 2007 song. The reduction in performance might be a reflection of the quality of the recordings: recording number 4 was chosen for the initial training because it was the one with

the highest SNR. In recording 1, there are overlapping calls from other individuals and there is masking by the biological noise from the coral reef.

The performance of the classification was next studied on the most popular calls that were found both as unit and subunits in all recordings. The results of the classification before training of the individual units (*Figure 7.13*) and subunits (*Figure 7.14*) detected in the recordings from different years are presented below for the most popular calls encountered.

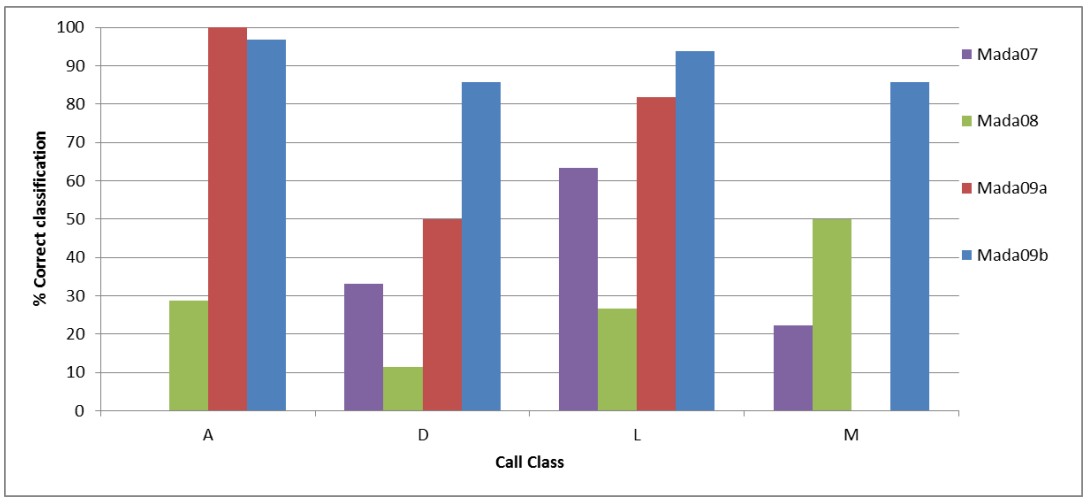


Figure 7.13: Correct classification rate (%) for the most common calls present in the 4 songs from 3 different years using the unit model. Note that unit 'm' is not found on its own in recording 1, leading to a performance of 0.The numbering in the legend refer to the recording numbers outlined in Table 7.5.

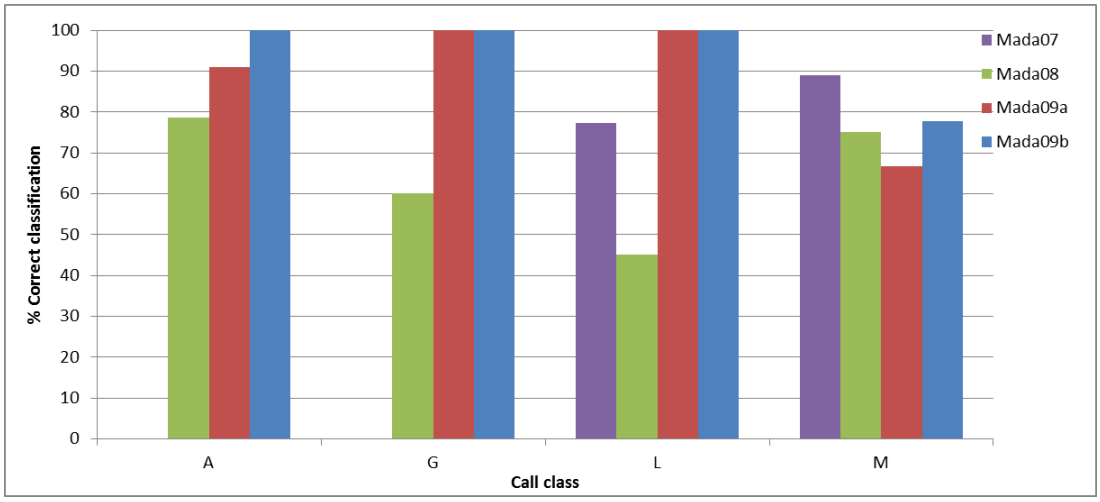


Figure 7.14: Correct classification rate (%) for the 4 most common calls present in 4 songs from 3 different years using the subunit model. Note that subunit 'a' and 'g' are not present in the 2007 recording. Subunit g forms the second part of a unit in the 2009 recordings and the first part of a unit in the 2008 recording, in addition to being found on its own in all three recordings.

Classification with the unit model was much better in the first recording; whereas, with the subunit model the classification without additional training is around or above 60% in all but one case. Subunit 'L' corresponds to unit 'L' in this case and it is a very common call in all recordings (*Figure 7.15*). This sound was observed in songs from populations in other parts of the world (Dunlop, *et al.*, 2007). This is either encountered on its own or preceded by other subunits of various sort. The lower performance of the classifier in recognising this call in the 2008 recording may be due to the fact that it was often overlaid with other calls emitted from animals in the proximity of the singer. The fast upsweeps of class 'L' are indeed very short in duration and their frequency sweeps up very rapidly; therefore, their characteristics may be easily missed if masked by other calls.

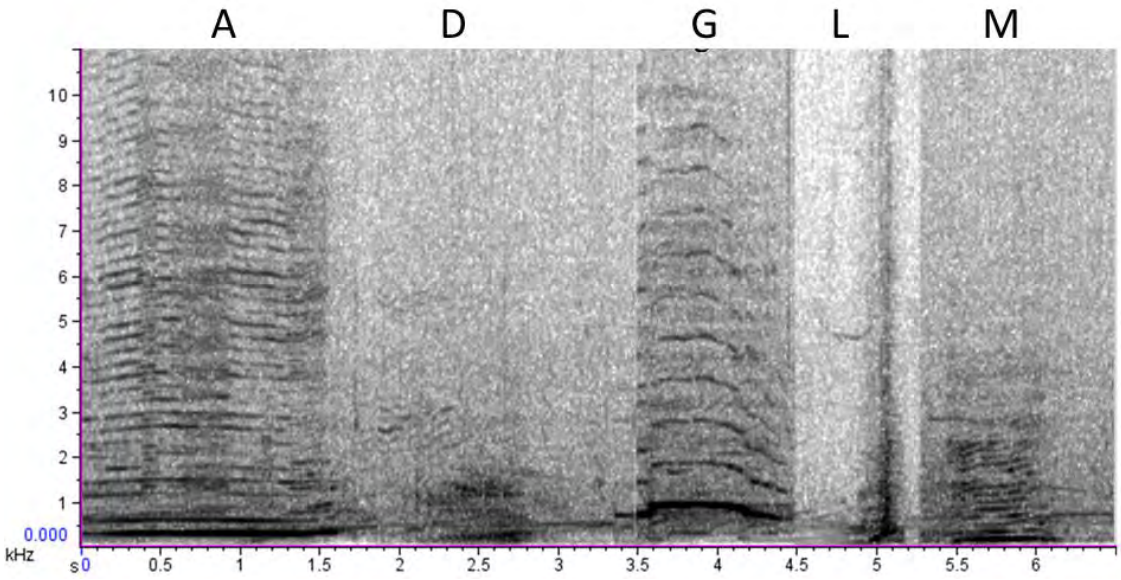


Figure 7.15: Spectrograms (computed with a Hamming window and resolution 22 Hz) of the calls which are the subject of Figures 5 and 6. The above data set is formed by concatenating clips recording 3.

The poor classification of unit 'D' in all the songs except the first one may be attributable to two main factors: in the recording taken on the 2nd August 2009 there was boat noise partially masking the call; in the 2008 and 2007 songs there were several 'amplitude-modulated' calls and the MFCCs may fail to capture accurately the characteristics that distinguish such calls.

Furthermore, unit 'A' in the song of August 2008 in most cases had a different start from the previous years in that it had an initial upsweep which led its misclassification. The subunit model instead captured this modification of the call and classified 'A' correctly in more than 70% of the cases. Indeed, the classification performance improved when a HMM model for a new unit was trained into the system which contained the additional subunit observed before 'A'.

Similarly, in recording 4 the call 'M' was only found in association with call 'L'; therefore the performance of the classifier with the subunit model is higher than when using the unit model. Again in this case training should improve the efficiency of the HMM; however, this could not be tested in this instance because the sample size was too small (N=3) to carry out both training and testing.

7.3 Classification performance with different training sets sizes

As an additional test, we tested the classification performance of the unit model with different training dataset sizes on the recording of 2009 to understand if a smaller percentage of the calls could be used during the training without reducing the accuracy of the classification. This factor is important when considering how time consuming and computationally expensive can be to train the model on an extensive dataset of songs. The number of calls used to train each category for each training scenario is given in Table 7.9.

Call class	**Training**			**Correctly classified (% out of 181)**		
	50 %	**25%**	**10%**	**50 %**	**25%**	**10%**
A	12	11	4	96.88	100.00	84.38
BC	10	5	3	100.00	100.00	100.00
D	6	3	3	85.71	100.00	85.71
FG	7	4	3	100.00	50.00	100.00
H	10	5	3	100.00	60.00	10.00
L	10	10	4	96.97	90.91	84.85
M	8	4	3	85.71	57.14	14.29
NL	10	6	3	87.50	81.25	37.50
PO	5	3	3	100.00	100.00	100.00
Q	10	6	3	100.00	93.33	93.33
S	9	8	3	86.96	95.65	91.30
TL	6	3	3	100.00	100.00	100.00
U	7	4	3	87.50	100.00	87.50
Overall	110	72	41	94.48	89.50	77.90

Table 7.9: Table showing the call types identified in the recording analysed, as well as the number of calls used during the training stage for each of the training scenario. The classification performance for each of the scenarios is presented as a percentage of the total number of calls tested for each call type. The training set number denoted by a '' mean that the actual number of calls used for the training stage should have been less than three if we calculated the appropriate percentage of calls for the training scenario; however, to train the HMM a minimum of 3 calls are required. Also note that the number of calls used for the training has been rounded to the nearest integer.*

The results show that the best performance overall was achieved using 50% of the data for training the HMMs and the other 50% for testing the classifier; however, this was not true of all the call classes as can be seen in *Figure 7.16*.

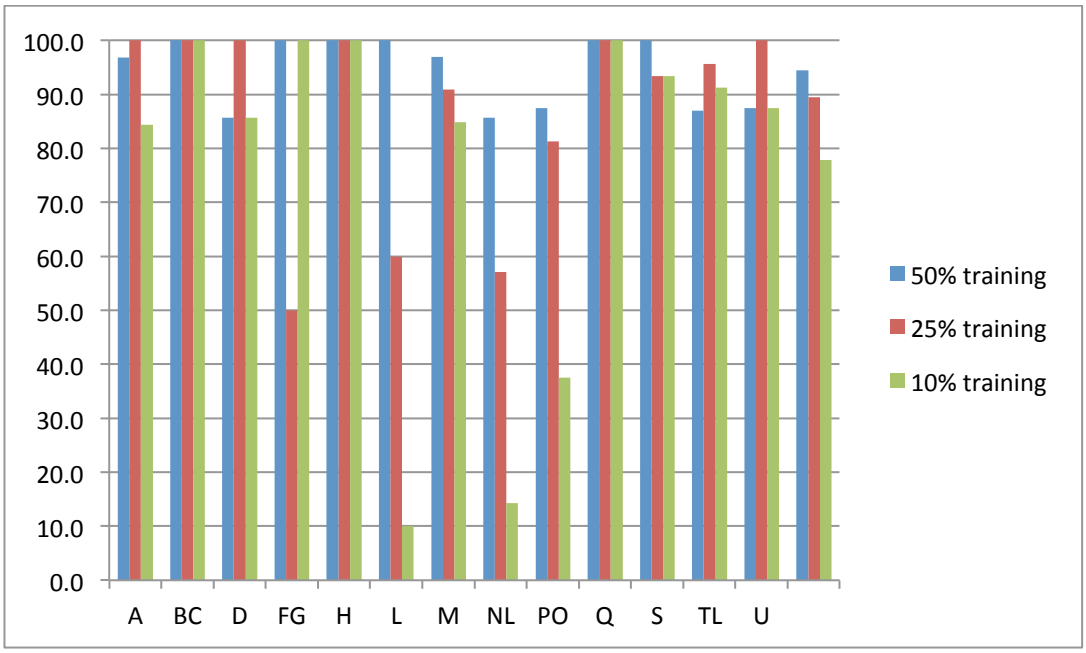


Figure 7.16: Percentage correct classification of the Hidden Markov Modelling classification obtained for three different training scenarios for each call type (or unit type) and overall.

With a 25% percentage reduction in training data, the classification performance decrease only by 4% but the mistakes affected different call types differentially. Indeed, in three instances, namely units 'L', 'NL' and 'PO', the classification accuracy halved (or nearly halved). On the other hand, there are 3 instances in which more units were correctly classified when there were 25% rather than 50% calls used for training.

In the last training scenario, when the HMMs were trained using only 10% of the data (or slightly more) the overall classification performance reduced to 78%. Again here some call types were more affected than others by the change in training set size. Specifically, units 'L', 'NL' and 'PO' were classified very poorly (<40% correct classification), whilst the classification of the other unit types was nearly equal to the one obtained with the other training scenarios.

Comparison with songs from other years and different locations will be presented in the next chapter to validate the applicability of the HMM classification and discuss biological factors relating humpback whale song production.

7.4 Conclusions

The results presented in this chapter showed how the classification based on Hidden Markov Models performed when presented with recordings of variable quality and with different levels of training. For an automatic classifier to be successfully implemented for the analysis of large datasets, the classification performance needs to be very high (at least 70% correct classification), especially when the task involves classifying a sequence of sounds to be able to recognise song components between years because mis-classification of a single may result in portions of songs being incorrectly classified. Ultimately this could lead to the same phrase or theme being ascribed to different categories (or vice versa).

The classification based on the unit model achieved quite poor performance when songs from different years and singers were introduced; indeed, in many instances, less than 70% of the calls were incorrectly classified using unit recognition (Figure 7.8-7.12), which suggests that it would not be advantageous to use an automatic classifier in these cases. However, by looking at the performance before and after training using units from the same recordings as the test set, it is clear that classification performance increases when samples of calls from the same singer/song are included in the training set. The associated performance improvement can be attributed to two factors: the training set containing more samples, representing the population of calls more accurately, and the fact that the new training sample contains calls that are more similar in characteristics to the ones present in the test set, facilitating correct recognition. This observation also applies to the subunit model but to a lesser extent, as the classification performance of the HMMs for this model is higher before and after re-training.

Overall, in order maximise the performance of the automatic classifier several samples should be included for each call category and of recordings of different quality so that variability amongst calls is taken into account. The results showed that certain calls are incorrectly classified more often than others and this should be taken into account when evaluating the overall performance of the classifier and to understand if the problem is related to the Hidden Markov Model associated with that call category or if it lies with the fact that the feature sets used for that particular call is unsuitable to describe its characteristics. In general, the classifier performance improves as the number of samples included in the training set increases.. This

means that when one chooses the suitable training set size, some prior information regarding the calls should be taken into account so that one might choose to use a higher percentage of training samples for a particular call category, whilst using a small training set for other calls that present more distinctive features. Another factor to take into consideration is knowing if certain calls are more stereotyped than others; stereotyped calls will change little across singers and different years, making it easier for the classification algorithm to identify new calls as belonging to a pre-existing category. Indeed, acquiring prior information about the vocalisations requires more human input; therefore, one will always need to trade-off the effort required with the advantages that will be obtained in terms of performance and reduction in computational load.

8. Song comparisons

In the previous chapter we analysed the structure of humpback whale songs recorded in Madagascar in the Channel of Ste Marie between 2007 and 2009, and presented the results of the automatic classification of these songs using Hidden Markov Models (HMMs) based on the recognition of units and subunits. The results showed that subunits are preserved more than units across years and that the automatic classification performance can be relatively high even with small training datasets. The first section of this chapter compares the overall song structure of the songs of Madagascar analysed in the previous chapter to understand the relationship in the changes at the level of the building blocks and those at higher levels of the song hierarchy. Specifically, we will look at the themes that are shared across years and check if these were present in other recordings of songs of Madagascar and then look in more detail at the finer structure of the songs and at which units and subunits are preserved through the years. The rest of this chapter compares the overall structure of previously recorded Madagascar songs that have been recorded in the waters off the East coast of the Island and the structure of songs from populations of the opposite hemisphere, specifically humpback whale songs recorded in Hawaii and Mexico during various years. Whilst the biological significance of the comparison and the overall performance of HMMs will be discussed in the next chapter, here we will present the results of the automatic classification conducted on the recordings of Hawaii and Mexico to test the reliability of the method and to examine if the subunit concept can be extended to songs from these populations too.

8.1 Comparison with known Madagascar songs

The songs of Madagascar have been less intensively studied compared to songs of populations of humpback whales that breed in other parts of the world. As a consequence, there is limited information about their structure and particularly about their evolution through the years. Scattered data are available that have been published in peer-reviewed literature and can be compared to the songs recorded in the Ste Marie channel between 2007 and 2011. Previous studies described songs at the theme or phrase level which is easily comparable across populations. Here, we will compare such sequences but will look in more detail at the building blocks that

make up phrase, i.e. the units, to build upon the vocabulary of vocalisations of humpback whales presented in the previous section.

As discussed in the previous chapter, the first description of Madagascar song was produced by Razafindrakoto (2001) who presented the themes composing songs that were recorded in Antongil Bay, a bay just North of Ste Marie Island, where the recordings for this thesis were collected. Although this song description was based upon two recordings taken during the Madagascar WinterW, we can assume that this account is an appropriate representation of the song sang by humpback whales in that breeding area in 1996 (*Figure 8.1*).

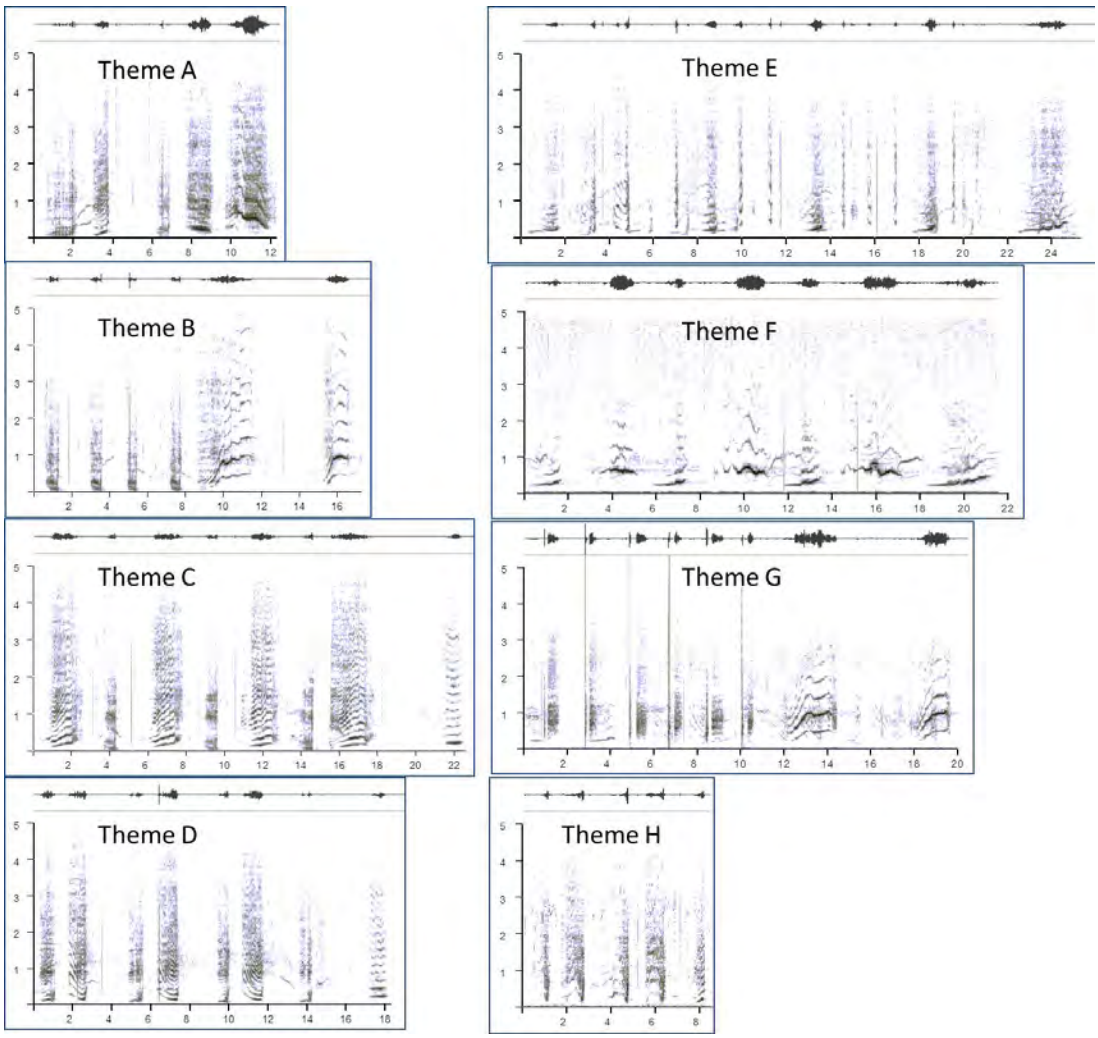


Figure 8.1: Composition of humpback whale song recorded in Antongil Bay in 1996 (adapted from Razafindrakoto (2001)). An example of one phrase is given for each theme.

The song described was composed of 8 themes, each of which is made up by a sequence of phrases. The song was compared to those recorded during the same year

in various locations in the Southern Ocean, and was found to be very different to any of the Australian songs and also from the song recorded of the coast off Columbia in South America (Helweg *et al.*, 1998). Given the proximity of Antongil Bay to Ste Marie Island, and the genetic evidence showing relatedness in the DNA sequences of these populations, it is feasible that animals singing in these areas belong to the same population (Rosenbaum *et al.*, 1997); indeed, singers might spend part of the season in one area and then move to the other one (*Figure 8.2*). Therefore, we can postulate that the song recorded in Antongil Bay is a precursor of the songs recorded in the Ste Marie channel between 2007 and 2011.

Figure 8.2: Close up view of the North of Madagascar showing the relative position of Ste Marie Island (red rectangle) and Antongil Bay (yellow rectangle)(created using Google Earth).

Comparison of the song of 1996 to the ones recorded in Ste Marie shows no themes or phases are shared amongst them. This is consistent with the idea that humpback whale songs evolve through the years and as a consequence, after a few years, the song has completely been replaced by a different one, as was demonstrated by studies conducted in Australia (Noad *et al.*, 2000; Garland *et al.*, 2011). However, close inspection of the spectrograms shows that there are shared building blocks between these song sequences (although detailed analysis is not possible without the original audio data), which suggests that these vary less frequently than phrases and

other components that are higher in the song hierarchy do. Indeed, this evidence supports that there may a limited number of building blocks that are arranged in different ways to form all the possible song sequences observed. Such observation motivates development of an algorithm for automatic classification of units or subunits, as explained in the introductory chapters of this thesis.

More recently, the song of Antongil Bay of 2006 was compared to Western Australian recordings (*Figure 8.3*).

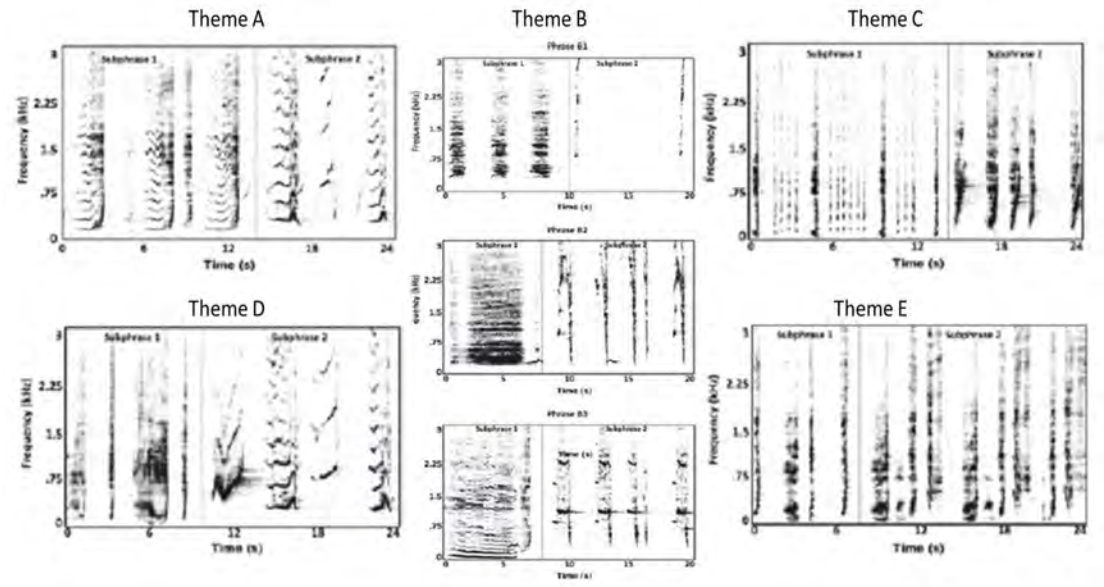


Figure 8.3: Theme composition of the 2006 song of Madagascar. All the themes presented above were unique to the Madagascar song except for theme B which was shared with the Western Australia song of the same year. The spectrograms above show only the basic themes that constitute the song and do not include themes that were formed by a combination of phrases taken from two different themes, which are known as transitional themes (adapted from Murray et al. (2012)).

Three of the themes presented above are shared with the 2007 song that we recorded in the channel of Ste Marie Island; specifically, these were themes B, C and D. This is consistent with the fact that songs of the same population of humpback whales evolve from year to year so that a portion of the song includes elements that were sang the previous year whilst it incorporates new themes that bring variety to the repertoire. A colour-coded summary of the Madagascar song composition across different years is given below to aid the comparison (*Figure 8.4*).

2006	1	2	3	4	5	-
2007	6	3	7	2	6	4
2008	8	2	9	10	3	2
2009	10	3	2	11	-	-
2010	12	2	3	-	-	-
2011	13	12	2	3	-	-

Figure 8.4: Colour coded Madagascar song description where each box represents a theme type present in the song sequence. Different colours represent different themes and the white boxes with a dash mean that no additional theme is present within the song. Transitional themes are not included in this comparative analysis. Note that the 2006 description is based solely on the investigation of the spectrograms presented in Murray et al. (2012).

The comparison of songs across years shows that between 2006 and 2011 two themes were preserved in the song structure and were mostly performed consecutively one after the other. Slight changes at the call level were observed in these themes within songs of the same year and also across songs of different years which may be due to individual variation as a consequence of the fact that different singers were performing the song and also to the natural progression in the evolution of the song. The observation that two themes are conserved whilst the others disappear from year to year or within a two year period could have various biological implications which will be discussed in the next chapter.

Whilst this section provided an overview of the general song structure at the theme level for songs of the C3 stock population (Ste Marie) of Madagascar (Rosenbaum *et al.*, 1997), the next section will describe the themes of the songs of humpback whales recorded in a different ocean basin in the Northern Hemisphere, namely the Pacific Ocean, which are known to be different from the themes of the songs produced by whales in the Southern Hemisphere to understand whether the automatic classification algorithm can be applied to songs produced by humpback whales globally and the same pattern is observed in terms of variability at the unit and subunit levels.

8.2 Songs of Hawaii and Mexico

Songs of Hawaii and Mexico were obtained from two affirmed researchers who have conducted studies on humpback whale songs for many years, specifically Dr Salvatore Cerchio and Dr Danielle Cholewiak (Cerchio *et al.*, 2001b; Cerchio *et al.*, 2005; Cerchio *et al.*, 2008). Testing the automatic classifier on songs different from the Madagascar ones had two purposes; first we wanted to test the reproducibility of the method and ensure that the classifier would work with songs emitted by different singers in different locations. Secondly, songs of Mexico and Hawaii were chosen specifically because these areas are in the opposite hemisphere of Madagascar which means that the populations of humpback whales in these regions should never mix because they are geographically isolated given that their feeding and breeding seasons occur at different times of the year. As a consequence, singers belonging to populations of the Northern Hemisphere should not share themes within songs because they would not be able to copy each other as they do not come into contact at any time during their lifetime. Hence, we are able to test whether the HMM classifier would work on songs emitted by completely different populations and check if the subunit concept is suitable to analyse their songs too and to identify which, if any, of the building blocks are shared across these non-mixing populations. Songs of Hawaii and Mexico (Socorro Island just off the West coast) have been shown to be exactly the same during the same time of the year; in other words, humpback whales of Mexico and Hawaii share their song repertoire despite being nearly 5,000 km apart (Cerchio *et al.*, 2001b)(*Figure 8.5*) .

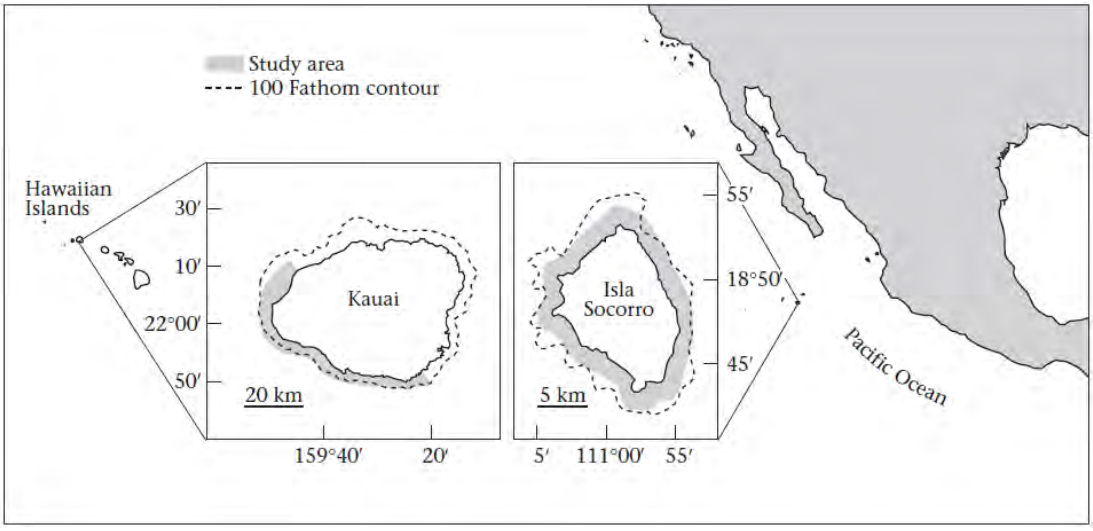


Figure 8.5: Map showing the location and relative size of Kauai and Socorro Island where the songs of humpback whales were recorded. The grey shaded area around the Islands represents the study area where recordings took place and they were all within the 100 meters water depth contour Cerchio et al. (2001b). The different environmental conditions between these recordings and the ones taken off the island of Ste Marie are evident because in the latter ambient recordings are much more reverberant because the channel is shallow compared to the waters around the Pacific Islands and there are numerous echoes of the sounds produced by the singer, as well as those of non-focal animals.

Given the fact that songs, and therefore their building blocks, are the same for Mexico and Hawaii within a year, we analysed songs of one or the other locations during different years. However, since songs will be broken down into their smallest components it is important to mention that small differences are observed at the unit level between songs of different geographical areas, which have been referred to as micro-geographic variations (Norris *et al.*, 2000), that can affect the classification performance. Such differences are comparable to the way in which words or syllables are pronounced differently in different dialects when describing human speech, and they are opposed to macro-variations in the song structure which refer to the differences at the phrase and theme levels; in other words differences in the sequences that compose the song structure. The latter macro-variations have been referred to as different dialects in the songs of humpback whale in previous literature, which may lead to confusion (Winn *et al.*, 1981; Parsons *et al.*, 2008).

Hence, in this chapter for each case we will specify at what level differences are observed.

A classification of the themes that compose the Hawaiian song of 1989 and 1991 was provided which was compared to the independent classification of the same recording conducted by the author and another trained observer to check the accuracy of the manual classification. The classification obtained through these three independent observers had a high level of agreement >98%. The few calls that were classified differently belonged to phrases that were transitions between a theme and the next. Transitional phrases are observed in all humpback whale songs and include vocalisations that are peculiar because they are usually stumped versions of the calls that will form the core of the following theme or are formed by a mixture of calls from the previous and following phrase. Hence, these transitional phrases occur only one or twice within a song and they look different on each occurrence. The manual classification obtained by the two trained observers was then compared to the classification at the theme level obtained from Dr Salvatore Cerchio as ultimate test of the reliability its accuracy. The themes observed in the 1989 song of Hawaii are presented below (*Figure 8.6*) and as expected they are different from the ones observed in the Madagascar recordings.

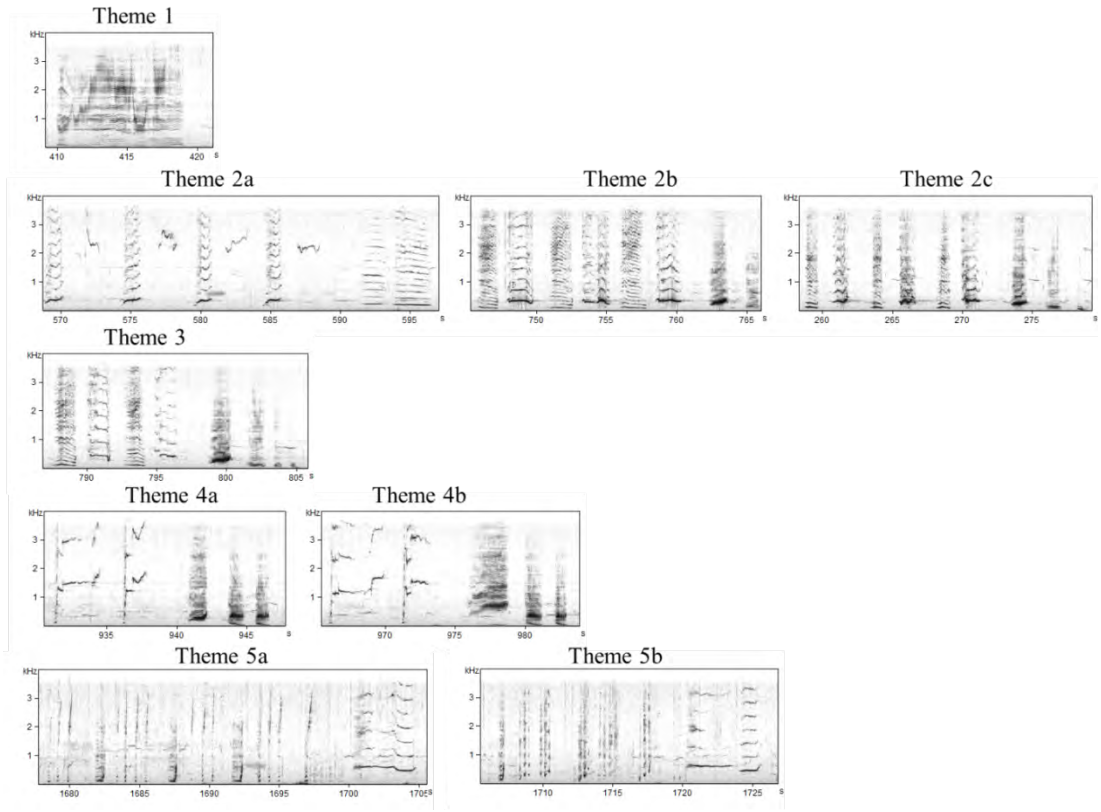


Figure 8.6: spectrograms representing the themes that compose the 1989 song of Kauaii (adapted from document given by Dr Cerchio of Whale Conservation Society). The numbers used to label the themes presented here are chronological and have no bearings to the numbering used to label the themes of the songs recorded in Madagascar described previously.

The spectrograms shown above represent a sample phrase for each of the themes observed; the actual theme is formed by repeating the phrase illustrated *n* times. The number of repetitions is variable and can change from one repetition to the next of the same song. As discussed in chapter 2, the song structure is extremely repetitive and it is unknown if this redundancy is a means of memorising the song component or if it actually plays a role in the meaning of the song. Research is being conducted to better understand this aspect of the singing behaviour of humpback whales (Suzuki *et al.*, 2006; Miksis-Olds *et al.*, 2008). In some cases, a theme is subdivided into versions a) and b) to describe themes that are essentially the same but present minor differences in the number of repetitions of a particular unit or if one unit in the sequence is substituted by a different one (e.g. Themes 4a and 4b, Figure 8.6).

So far, this section has introduced the generic structure of songs of humpback whales recorded in different years and different geographical areas at the theme level and

highlighted how themes vary between years and how they differ completely between populations of opposite hemispheres in terms of their macro-structure, i.e. the sequence of calls that are juxtaposed to obtain the full song. By contrast, the rest of this chapter will look at the lower level structure of the song, examining the characteristics of individual units and subunits and consider any changes that may occur in them through years and across populations. Finally, the results of applying the automatic classifier to these songs are presented.

8. 3 Call evolution within and between songs

Recall that themes and songs evolve over the course of a breeding season and also between years with new components being incorporated in the themes and the order of phrases within the themes being rearranged, leading to the generation of original songs. Despite the fact that numerous studies have looked at the way songs evolve, no one has yet described the way in which individual calls change through the years as the songs change. This section aims to provide a description of such changes which concern the building blocks of humpback whale songs. Specifically, a description will be given of the vocalisations which were identified as shared across the songs of different ocean basins. The purpose here is to understand how whales modify single vocalisations through the years rather than looking at the minute details in the signal characteristic, whose differences can be ascribed to individual variations of the vocal apparatus of the singers. In other words, this section will summarise the calls shared amongst songs and look at their substructure, i.e. if they are always encountered as individual units or if they can be decomposed into subunits and in the latter case, specify the different associations of subunits found in the recordings. All the figures presented below were produced using Raven Lite.

The first call that was shared amongst recordings was unit ‘A’; this vocalisation was always found on its own separated from other calls by silences at the start and at the end of the call, which means that the unit and subunit for this call class correspond (*Figure 8.7*).

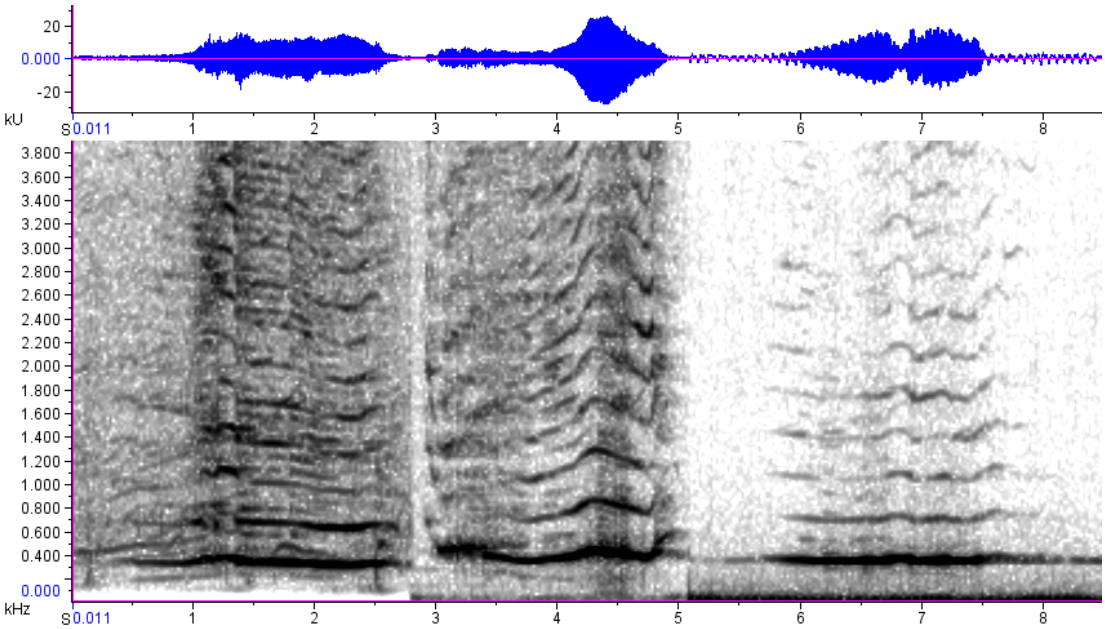


Figure 8.7: Amplitude (top) and spectrogram (bottom) of a sample vocalisation 'A' taken from a recording of Madagsacar (left), Hawaii (middle) and Mexico (right). The y-axis of the spectrogram shows frequency in kHz and the x-axis indicates time in seconds.

Vocalisation 'A' was found in all recordings of Madagascar, Hawaii and Mexico with most energy in the fundamental frequency and numerous harmonics. The overall shape of the call varied within recordings from completely straight to slightly arched. Overall, the call was frequent in all recordings and highly stereotyped in the sense that changes in duration, overall shape and fundamental frequency were minimal. In some instances, the formation of subharmonics was observed at multiple integers of half the value of the fundamental frequency.

Another shared call amongst songs of all three geographic areas was vocalisation 'F' which unlike the previous unit was mostly found in combination with some other call before or after it without any silence to separate them. In other words, call 'F' was mostly found as a subunit and occasionally observed on its own as a unit (*Figure 8.8*).

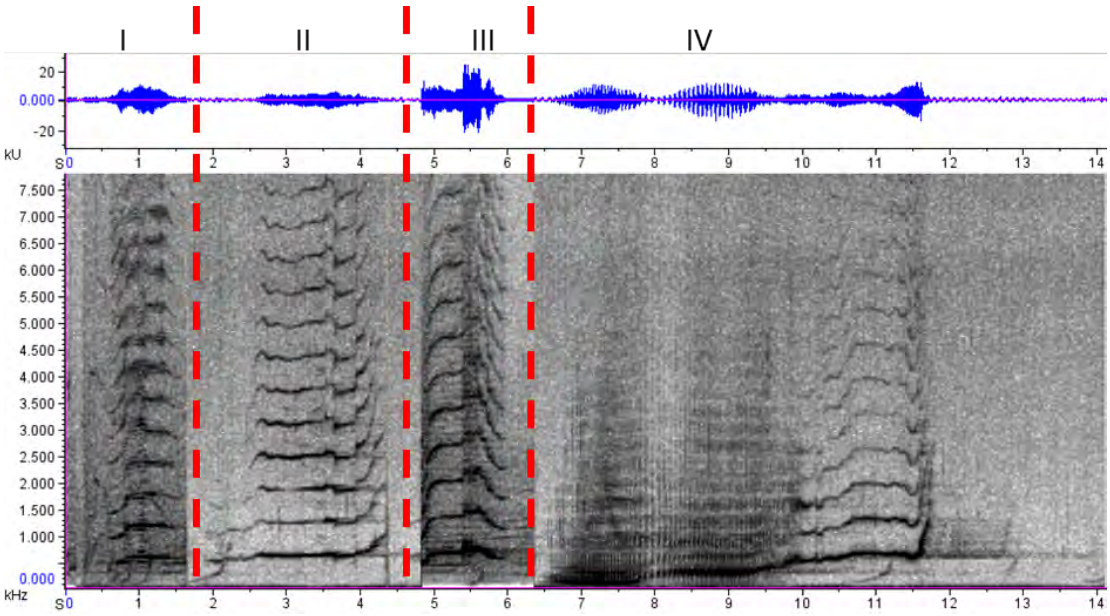


Figure 8.8: Amplitude (top) and spectrogram (bottom) of a sample vocalisation 'F' taken from a recording of Madagascar (left), Hawaii (II). The y-axis of the spectrogram shows frequency in kHz and the x-axis indicates time in seconds. In the songs of Madagascar this call was found in association with another call, namely call 'G' (III)) which was both found before or after call 'F' to form units 'FG' and 'GF' respectively. In the songs of Mexico 'F' was found on its own or as the last component of a long vocalisation made up of three subunits (call IV).

Subunit 'L' was the most frequent call in the Madagascar recording and was present in themes from all the years analysed. Hence, it is unsurprising that this vocalisation was present in the songs of the other two locations across years (*Figure 8.9*).

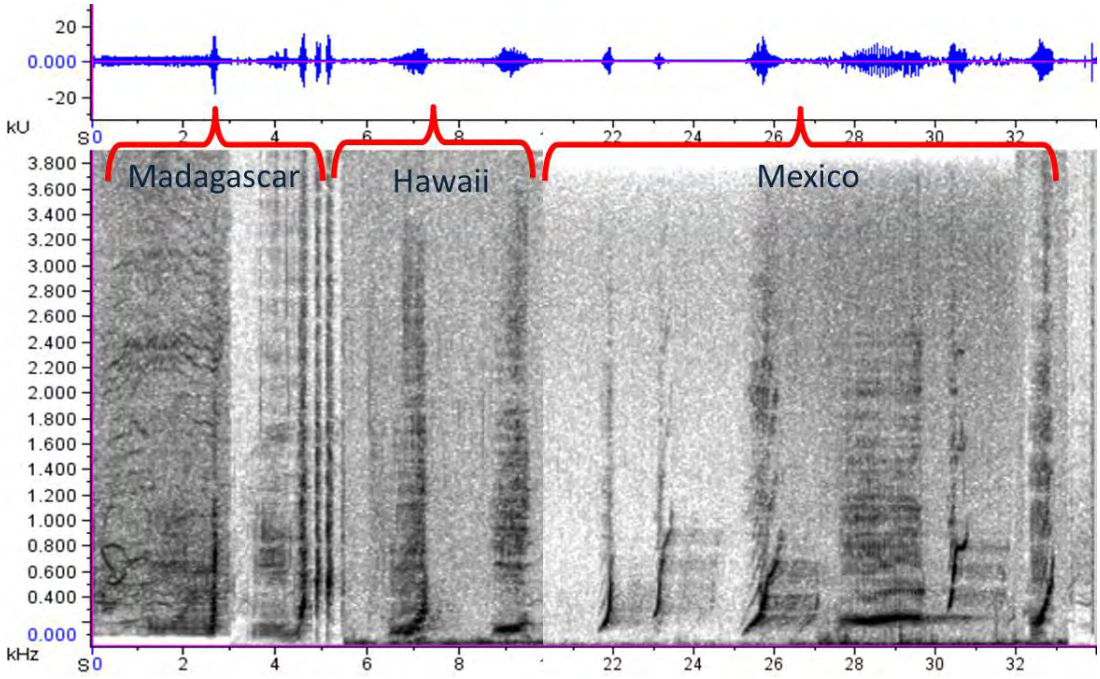


Figure 8.9: Amplitude (top) and spectrogram (bottom) of various samples of the vocalisation 'L' on its own and with associated subunits in recordings from Madagascar (left), Hawaii (middle) and Mexico (right). The y-axis of the spectrogram shows frequency in kHz and the x-axis indicates time in seconds. In the songs of Madagascar this call was found in association with another call, namely call 'T' (first vocalisation from the left) which was only observed before call 'L. In the songs of Mexico 'L' was found on its own or in a slightly 'stumped' version as the starting subunit in unit 'LC', of which there were many versions (two of them are shown in the spectrogram as the last two calls of the Hawaii sequence). In the recordings of Mexico, 'L' was found always on its own but sometimes just after the flat frequency call represented in the spectrogram above (3rd call from the right).

Whilst subunit 'L' was extremely common in the Madagascar recordings and also in the songs of Socorro Island, it was quite rare in the songs recorded off the coast of Kauai between 1989 and 1993. A 'stumped' version of 'L' was observed in the Hawaii recording though, which was common and found in association with other subunits to form many variations of the unit 'LC'. The difference between the Hawaiian call and the more recent instances of Madagascar and Mexico lies in the fact that in the latter two locations call 'L' had a fast frequency upsweep so that the majority of the energy in the final section of the call was distributed equally in a band spanning few Hertz to about 800 Hz, as depicted in Figure 8.9. Although the

sweep rate can vary slightly between calls, their duration and frequency is fairly consistent. These sounds have indeed been shown to be extremely stereotyped in a recent study conducted in Hawaii, suggesting they can be used to detect humpback whales using passive acoustic monitoring techniques (Stimpert *et al.*, 2011).

Broadband calls are less frequent than the other types of calls within songs, and although a clear fundamental frequency cannot be established just by observing a spectrogram, the band where most energy is contained can be used as indicator to compare the call between songs and determine if it is indeed the same vocalisation that we observe in the different songs. Some broadband calls were also shared amongst song repertoires, as shown in the figure below (*Figure 8.10*).

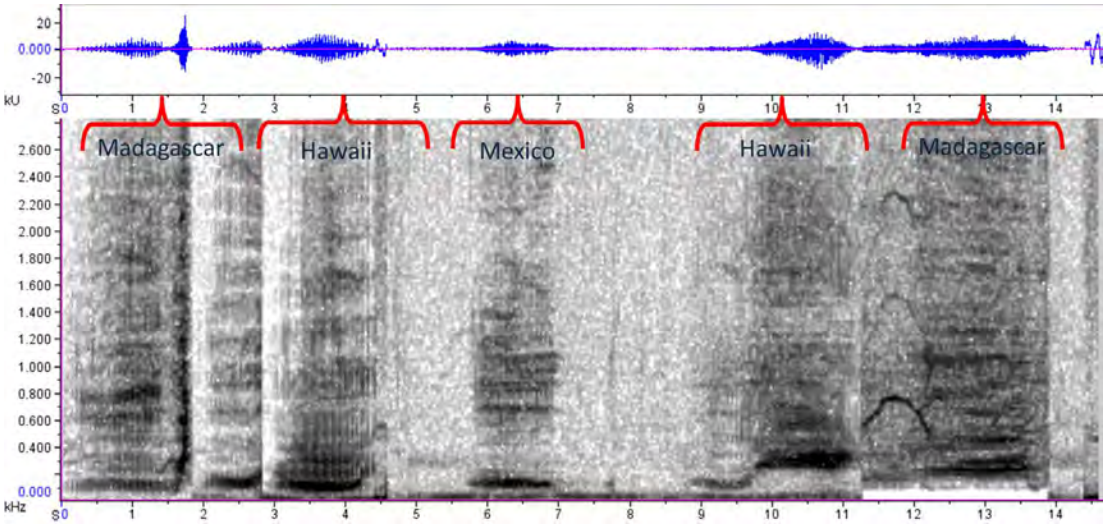


Figure 8.10: Amplitude (top) and spectrogram (bottom) of various samples of the shared broadband vocalisations encountered on their own and with associated subunits in recordings from Madagascar (left), Hawaii (middle) and Mexico (right). The y-axis of the spectrogram shows frequency in kHz and the x-axis indicates time in seconds. In the songs of Madagascar the broadband call labelled 'T' was found in association with another call, namely call 'L' (first vocalisation from the left). In the songs of Mexico and Hawaii 'T' was found both on its own and in association with other broadband calls. The spectrogram above shown subunit 'T' associated with 'P' (last two calls on the right), and in Figure 8.8 it was shown in an instance were 'T' was the first subunit of a unit composed by three elements, specifically subunits 'T', 'P' and 'F'.

The duration of the calls described above was approximately the same in all the songs and so was the frequency band where most energy was contained. Even

though many combinations of units were observed which contained subunit 'T', this call was always the first component of the unit, the end of which was a subunit with a higher fundamental frequency was attached.

Some calls were extremely variable even within songs in that their fundamental frequency shifted from one repetition of a phrase or theme to the next. This means that it was not possible to determine confidently if these calls were shared between songs of different populations. The problem arises because the fundamental frequency is the main characteristic used to discern calls, whilst the overall shape of the call is a secondary factor. There are instances of calls that have the same overall shape and duration but the attack and end frequency of these sweeps is different and therefore the calls were ascribed to different call classes.

Table 8.1 provides a summary of the subunit combinations and defines which ones are shared.

SHARED CALLS	Hawaii 1989	Hawaii 1991	Mexico 2006	Madagascar 2009
Hawaii 1989	-	4	5	2
Hawaii 1991	3	-	2	1
Mexico 2006	6	2	-	1
Madagascar 2009	4	2	2	-

Table 8.1: Overview of the number of subunit types (pink shading) and unit types (blue shading) shared amongst recordings.

The number of call types shared across recordings demonstrates that more subunits than units are present in songs of different years and different locations, suggesting that the smaller building blocks are less variable. Although the number of shared subunits (pink shading) may seem small, one needs to consider that the song sampled for these comparisons is very limited because accessing data from other locations that could be compared to the Madagascar songs was difficult, and that there is a 20 year gap between the songs of Hawaii and Madagascar. This comparison between shared units and subunits took into account the fundamental frequency of the calls and their harmonic structure and call types were assigned to the same category only if there was absolute certainty that the calls observed were shared amongst songs. In certain cases, very similar calls were observed that had minor differences in structure or whose fundamental frequencies differ by about 100 Hz lower or higher than another call. In such cases, the calls observed were assigned to different call categories even though one might argue that the differences might be

due to individual differences or environmental factors. Indeed, no information was available about either the singer or the depth at which the animal sang, whilst both of these factors could play an important role in the features of the observed calls. One problem with analysing variations in calls at the unit and subunit levels extracted from recordings of different geographic regions is that the sound production mechanism in baleen whales is still largely unknown. Therefore one cannot determine whether minor differences in call structure are related to the physical differences of individual singers and the sound propagation mechanisms through the environment, or if some of these differences truly represent a novel call type.

8.4 Automatic classification across songs

The automatic classification algorithm was tested to classify songs of humpback whales across years and populations. In order to do this, the recordings were all converted to a consistent sample. The chosen sample rate was 16 kHz because this allowed the inclusion of most of the recordings that were available to the author whilst including in the analysis all of the calls and most of the harmonics of the vocalisations present in the songs.

The initial tests were conducted on the recordings from Hawaii 1989 because they were the first chronologically. One recording from 1989 was used for training the models initially and running all the subsequent tests. Initially, the HMMs were trained on the first recording of 1989(A) and tested using the unit grammar to recognise calls of the same recording and of another recording of 1989(B) to test if the performance was consistent across recordings that contained the same units performed by different individuals (*Figure 8.11*).

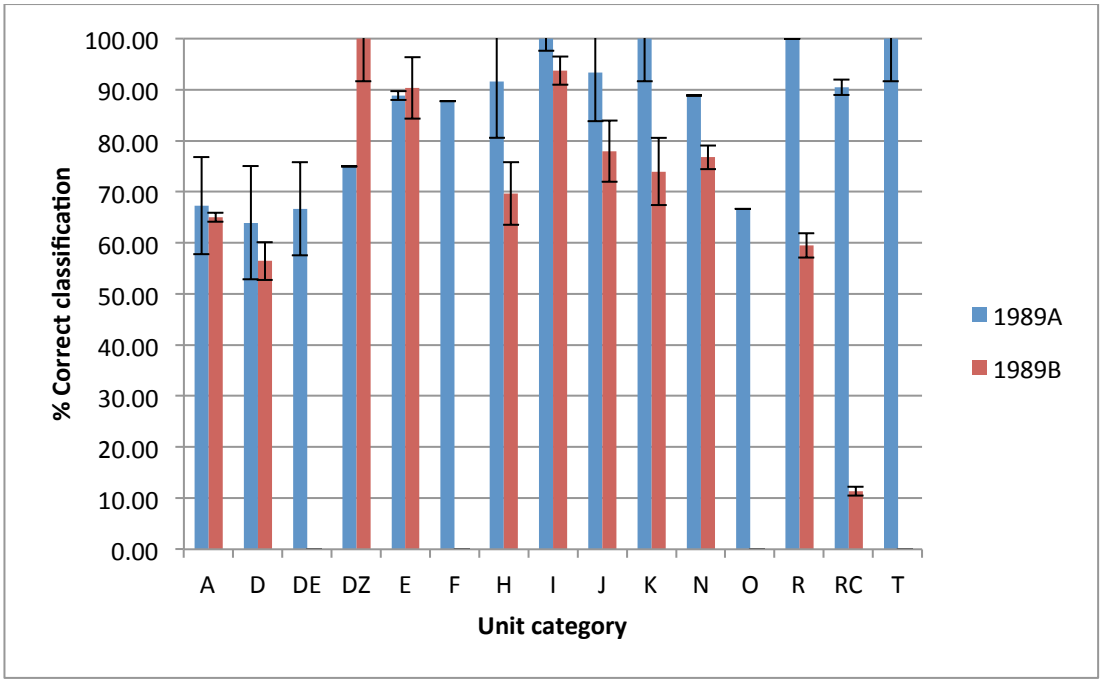


Figure 8.11: Classification performance of the automatic classifier based on HMMs trained with the unit grammar. The results show the automatic classification performance of two recordings, both taken in 1989 in Hawaii but on different days which means that we assume a different singer was performing the same song. The recordings capture different portions of the songs and for this reason, three of the units encountered in the first recording were not found in the second one (which is why there is no pink bar for three units). The performance bars in this figure are the average performance of the automatic classifier obtained from running the test 3 times, each of which was carried out using 50% of the data for each call type for training. For each round of testing 50% of the data of each call type manually classified was randomly selected as the training set and the remainder was used as the testing set. Error bars indicate the standard error for the results obtained for the three rounds of testing.

The variability associated with conducting multiple tests using different random selections of training samples show that the results are consistent independently of which calls are chosen for training the HMMs. This suggests that the results obtained from trials were only one round of training is carried out are reliable and represent the performance of the automatic classification algorithm. The variance was higher for those call types whose sample size was very small where a difference in classification of just one call resulted in a relatively high error rate.

In terms of classification performance, the results show that all the units present in the first recording were accurately classified also in the second recording apart from call type 'RC'. For the latter the performance dropped drastically to approximately 10%, this is because this unit is extremely variable both in duration and frequency (*Figure 8.12*).

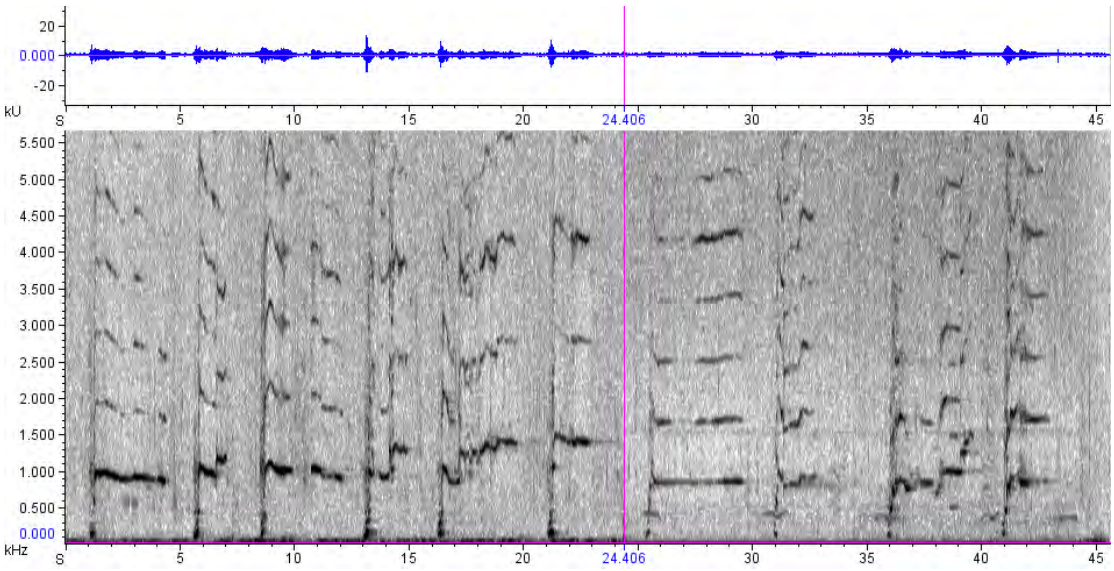


Figure 8.12: Spectrogram (NFFT 1024, Hanning window (1024)) and amplitude of the unit RC from recording 1989A (left of the pink line) and 1989B (right of the pink line).

The unit presented in Figure 8.12 was classified as unit 'N' in nearly all the cases possibly because this was the most similar unit to the second part of the structure of call type 'RC', and the classifier did not group this call with unit 'R' which corresponds to the first short low frequency pulse at the beginning of each 'RC' call. The fact that call 'RC' was misclassified nearly in all cases in the tests conducted on recording 1989(B) trained on the dataset from recording 1989(A) is attributable to great variability of the second part of this call. Even though there was much variability associated with this call type, all these vocalisations were grouped in the same unit category because when investigating the song structure at the phrase level, they were found to be in corresponding phrases.

So far we looked at the robustness of the classification performance when HMMs are trained with different random samples of call types but the amount of data fed in the system was unvaried; specifically training was always done using 50% of the calls identified for each call category. The next set of results investigates the performance

of the classifier when the amount of training data is altered. In a similar fashion to what we didin the previous chapter for the vocalisations detected in the songs of Madagascar, we tested the classification on the Hawaiian calls when 50%, 25% and 10% of the dataset for each call type used for training the HMMs. The number of calls that were tested for each unit class is summarised in *Table 8.2*.

	Number of calls tested			**Number of calls correctly identified**		
Class	**50%**	**25%**	**10%**	**50%**	**25%**	**10%**
A	51	76	91	34	91	78
D	11	14	16	8	16	8
DE	4	5	5	3	5	5
DZ	4	4	4	3	4	4
E	5	10	11	5	11	2
F	15	13	15	10	15	15
H	4	4	4	4	4	4
I	4	7	7	5	7	7
J	34	53	63	33	63	56
K	10	14	14	9	14	12
N	3	3	3	3	3	3
O	6	8	8	4	8	4
R	9	17	17	10	17	16
RC	6	9	9	6	9	8
T	13	22	27	12	27	23
Total	179	259	294	148	294	245

Table 8.2: Summary table of the unit classes of vocalisations encountered in the Hawaiian 1989 song. The table shows the number of calls tested for each call class according to the amount of training (expressed as a percentage of the total number of calls for each category) carried out on the data and the relative number of calls correctly ascribed to each class. When the number of calls to be trained did not correspond to an integer, the value was approximated to the nearest integer number.

The dataset used for this test was song 1989(A) and as before we used vocalisations taken from the same recording to train the automatic classifier, which means that the first column of the following figure will match the results presented before for the same recording. We would expect these new tests to have higher classification performance because calls from the same singer are used for training and testing (*Figure 8.13*).

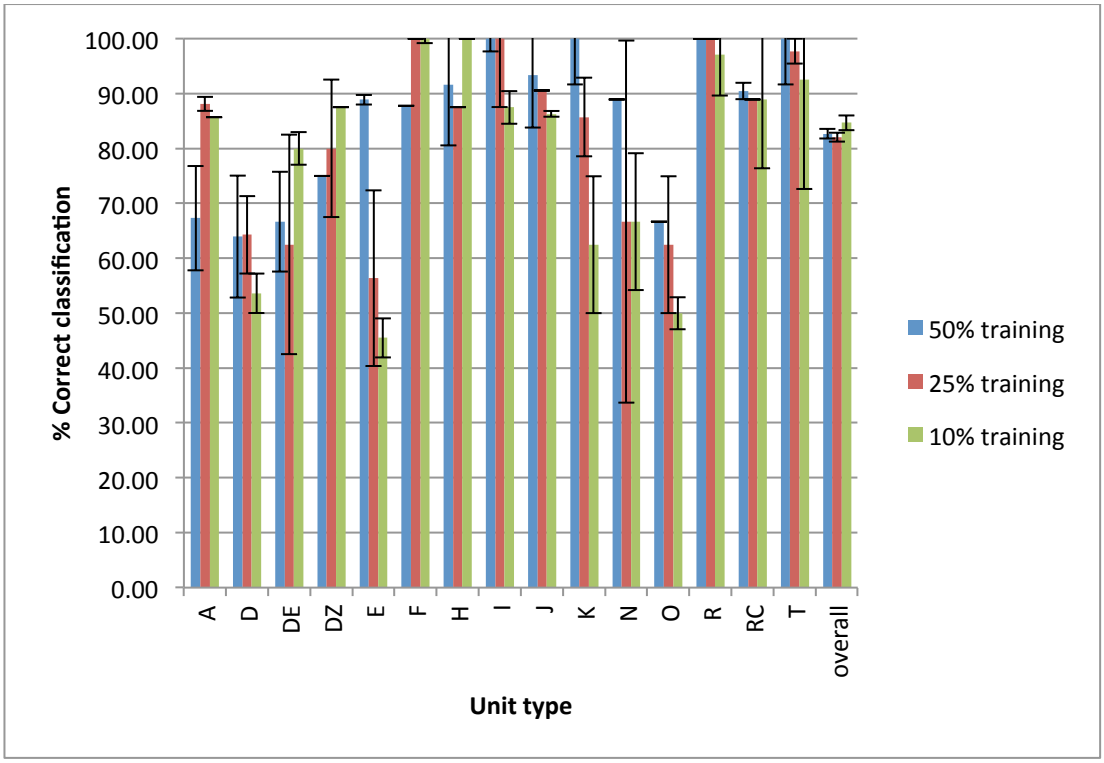


Figure 8.13: Performance of automatic classification algorithm tested on the Hawaiian recording from 1989 where the HMMs were trained with different amounts of calls. Each trial was performed three times using randomly chosen training sets, and the variability the results obtained using different training sets is expressed by the error bars.

The results show that overall the automatic classification performance obtained using different amounts of training data was very similar and above 80% with little variability amongst trials. However, the amount of training performed seemed to affect the recognition performance of unit types differentially. Unlike for the other unit types, in the case of units 'DE' and 'DZ' the number of calls used for training was not modified for the different trials because there were too few samples of these calls present in the recording; however the performance of the automatic classifier in correctly identifying these calls changed between trials. This observation just highlights the fact that the classification of all of the vocalisations is linked to the training set presented and therefore, to achieve the best possible results, the recognition algorithm must be fine-tuned for each class. Indeed, one needs to choose the most suitable scenario to obtain a performance that is suited to the task at hand and decide on the trade-off between training set and overall recognition performance, as we will discuss in more detail in the next chapter.

Now that we have looked at the specific performance of the automatic classification using HMMs based on the unit grammar to classify songs of the same year recorded in Hawaii, we will investigate the performance of the algorithm across years and songs of different populations where the fine characteristics of calls might be different, as mentioned in the previous section. The system was trained using 50% of the data for each unit call type of the song of Hawaii of 1989(A) and the tests were conducted on one recording for each of the other locations and one additional song of Hawaii recorded in 1991 to test if the performance of the classification algorithm was robust enough to recognise shared calls that present micro-geographical differences in the same way as a human listener can.

The results obtained using the unit dictionary described in previous chapters is compared to the performance of the automatic classifier based on subunit classification of the calls. Both of these scenarios are presented for two cases: firstly the training set of the first song of Hawaii (1989) was used as the only training sample for the recognition task of all the other songs, which means that some of the calls that are new each year will not be correctly classified (*Figure 8.14*). Secondly, the training set was updated to include some call samples for each call category for each song.

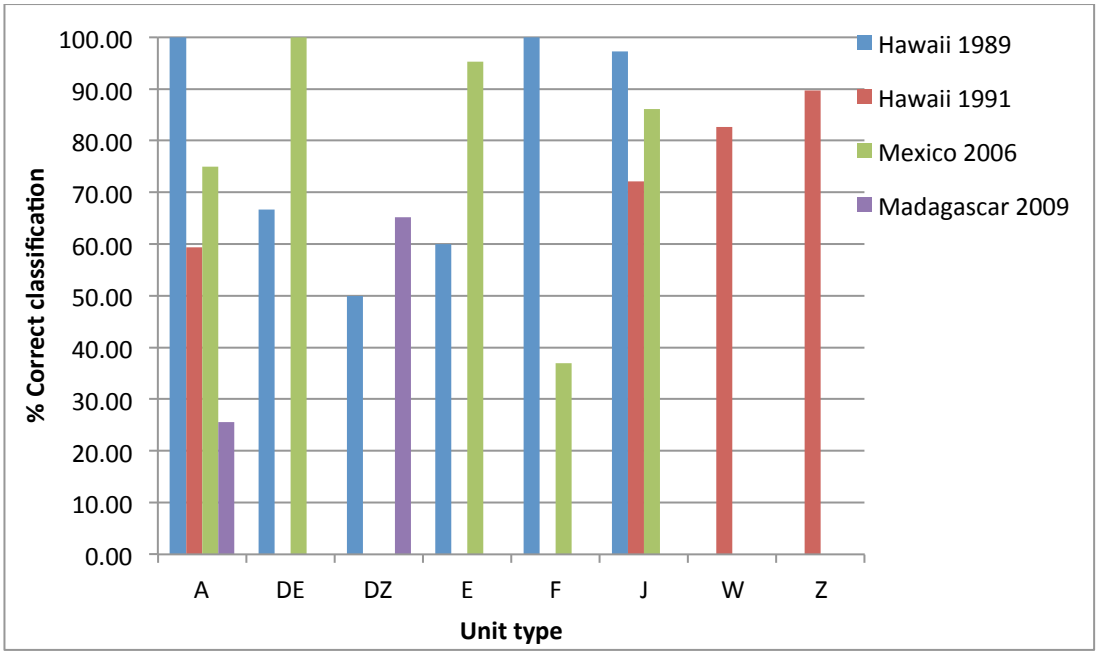


Figure 8.14: Classification performance of the algorithm using HMMs trained only on the call types found in the recording of Hawaii 1989. Note that when there is no bar for a data series it means that that particular unit was not found in the song of that year not that the performance of the classifier scored 0%. In addition, although call types W and Z were found in the recording of 1989 used for training, their sample size was too small to allow for testing the classification performance for that call type on the unit in that year. However, the samples presented in 1989 for those two units were trained in the HMMs and successfully classified calls in the song of 1991.

The results showed that in all but 2 cases the classification performance of song units using HMMs trained on a different recording from a different year and location was 60% or above. In the case of unit 'DZ' the classifier did better when presented with units of another recording than its own which was unexpected given that calls emitted from the same singer are generally more alike and should be easier for the classifier to cluster together. We can attribute this unusual result to the small sample size for this call, which included only 3 samples for training and the same number for the test set.

Call type 'A' was the only one that was encountered in all 4 songs analysed and it was one of the most common vocalisations in all of them. However, the performance of the classifier for this call type dropped dramatically when presented with calls of the Madagascar song recorded in 2009. This is surprising because earlier results of

the units classification carried out using the Madagascar songs for training (Chapter 7) showed that unit 'A' was correctly classified most of the time. This large difference may be due to two factors: 1) the training set built using the Hawaiian song contains one or more units that are very similar in characteristics to call type 'A', or 2) the quality of the Madagascar recording is very different from that of Hawaii leading the feature sets (MFCCs) to describe the call inaccurately in some cases.

Whilst 8 out of the 17 sound unit types found in the song of Hawaii recorded in 1989, only a maximum of 5 units in total were shared with a song from a different year, specifically the Mexico 2006. These findings show that song evolution is not only at the theme and phrase levels but that it also involves the building blocks of songs. Indeed, observing the changes that occur at the theme or phrase level one can see that some themes are completely preserved from one year to the next whilst others are completely replaced by new themes that are made up of different phrases from the songs of the previous year. This process leads to the song changing thematic structure entirely over the course of a few years. On the other hand, if we observe the changes in units composition from year to year, there does not seem to be complete replacement of unit types after a few years, instead some units are more or less common in different years and some are formed by rearranging small chunks of their components to form novel combinations, in a similar way to how birds and humans rearrange syllables during their utterances. For this reason the number of shared subunits across recordings is greater than the number of shared units despite the fact that the overall number of subunits forming the 'vocabulary' of songs is smaller. This is an attractive characteristic from the classification point of view because using fewer building blocks means that the computational load and the amount of training needed will be reduced. Results of the automatic classification performance for shared subunits are presented below (*Figure 8.15*).

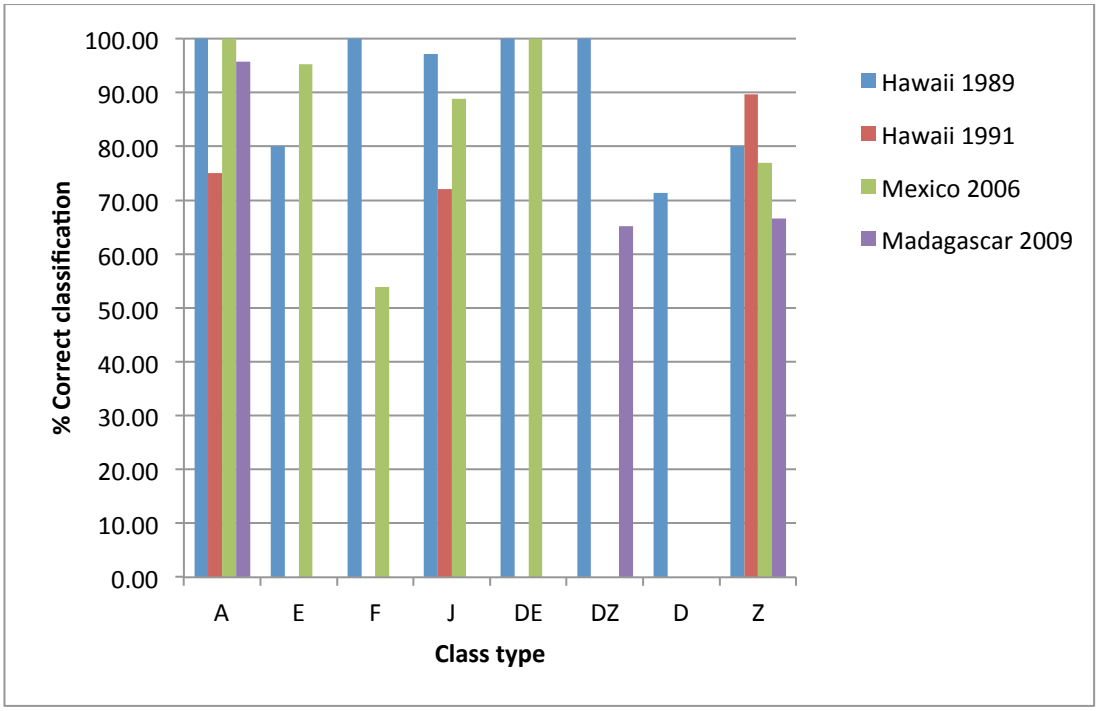


Figure 8.15: Automatic classification performance of all the songs analysed from 3 different locations using the HMM modelling based on subunit recognition.

As expected by definition, the number of shared subunit types across songs was greater than the number of unit types which is advantageous because it allows comparison of more recordings with less and less training required since one needs to add new vocalisations for fewer instances. In addition, the recognition performance of the classifier based on the subunit grammar was better than the one obtained using the unit modelling. Subunit 'Z' was not found as a unit in any of the recordings apart from the one used to train the algorithm but was found in association with other subunits in all the recordings. For this reason we have the recognition results for 'Z' in *Figure 8.15* but not for the unit recognition (*Figure 8.14*).

As it will be discussed in more detail in the next chapter, the identification of a shared 'vocabulary' amongst humpback whale populations that are spatially isolated could have important implications in biological terms because it may give cues as to how they produce sounds if physical constraints to sound production can be identified and it may provide an insight into the usage of the calls and the song purpose. In addition, identification of common calls amongst the repertoire of humpback whales and description of how their characteristics change over time are

necessary information for scientists who wish to develop automatic classifier and detection algorithms, particularly for localising humpback whales in real time.

Lastly, a cumulative test was carried out to test the performance of the algorithm trained on the original training set of Hawaii 1989 but with the addition of all the new units and subunits encountered in the other recordings to test if the recogniser would cope well with a cumulative training set that included sample of calls from all the recordings analysed (*Figure 8.16*).

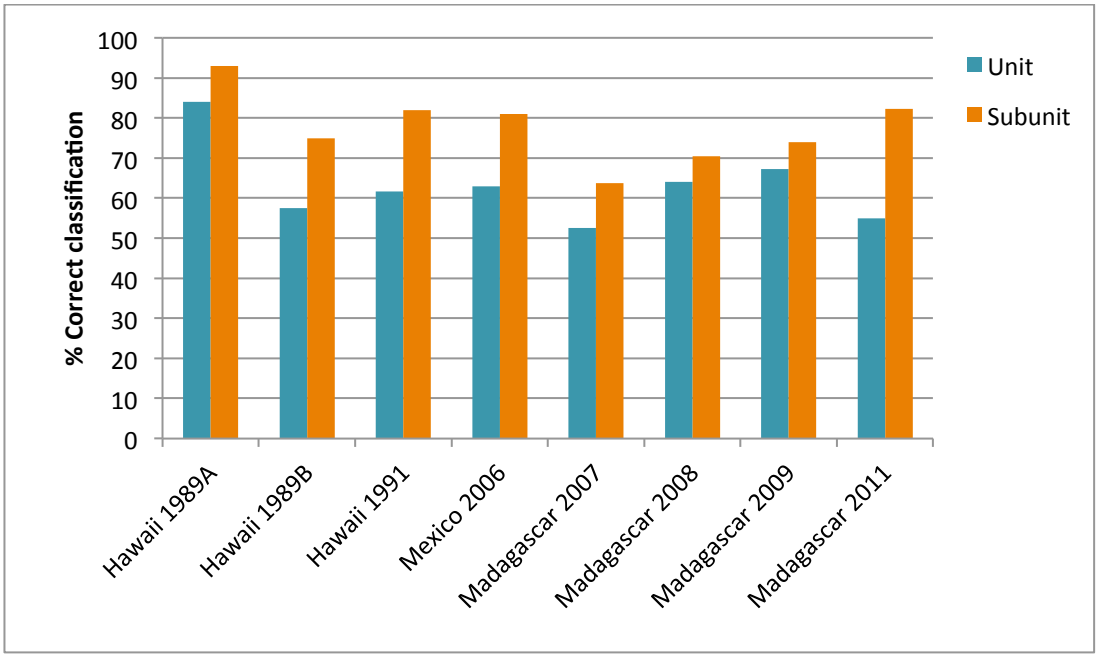


Figure 8.16: Percentage of correct classification overall of the units and subunits contained in a variety of humpback whale songs from different years and geographical areas.

The results of the recognition tests carried out on all the songs using a cumulative training set which included 50% of the calls of the Hawaii 1989(A) recording and 50% of the total number of calls for each new call type that was introduced from the other recordings showed that high levels of correct classification can be achieved, particularly if one implements the subunit grammar. The overall performance of the classifier was also related to the quality of the recording; indeed, the worst results were observed for the recordings of Madagascar 2007 and 2008 which included several overlapping calls because more than one singer was recorded on those occasions. Also during these tests, as observed in previous recognition tasks, some calls were classified correctly nearly every time whereas others were more prone to misclassification, which affected the overall performance percentage presented in

Figure 8.16. Broadband calls were especially problematic probably because the MFCCs used for characterising the signals are not particularly well-suited to describe such signals. In addition, some incorrect classification occurred for very short signals, which could be a result of the fact that the MFCCs were calculated using the same window length for all calls. However, the structure of the HMM toolkit does not allow to change frame sizes depending on the type of call presented and most biologist who might need to use an automatic classifier would not want to change the acoustic parameters. An overview of the ways in which it would be possible to improve the automatic classifier and to extend its applications will be given in the Chapter 9.

8.5 Conclusions

This chapter presented an overview on the whole structure of the songs recorded in Madagascar, in the Ste Marie channel, by the author and compared them to other songs produced by the same humpback whale population in previous years. Because Madagascar is a very remote area, little research has been put into studying the structure of humpback whale songs in this area compared to other regions of the World where songs have been recorded since the 1ate 1980's. It was shown how some of the subunit components defined in this thesis were shared across populations that reproduce in separate hemispheres, suggesting that the use of these building blocks could aid classification of songs and the identification of shared features between populations.

The results presented in this chapter showed how Hidden Markov Models can be used to successfully classify songs of humpback whales produced by multiple singers on different breeding grounds. This can be achieved simply by training the existing model with new sound subunits. This means that once the HMMs are setup then only limited manual input is required to adapt the model for recognition of new songs and to quickly run comparisons of the repertoire of different populations, as discussed more in detail in the final chapter of this thesis.

9. Discussion and perspectives

9.1 Summary of findings and original contributions

As it will be discussed in more detail in the following sections, the following objectives have been achieved:

- The development of an automatic classification algorithm for classifying humpback whale songs. The results presented in chapters 7-8 showed that Hidden Markov Models can be used to classify the elements that compose humpback whale songs accurately. As for all supervised algorithms, a robust training set containing several examples for each call class are necessary to achieve high levels of performance.
- For the first time, subunits were formally defined as the smallest building blocks of humpback whale songs. Although the subunit nomenclature existed in the first description of humpback whale songs, given by Payne and McVay, they were never defined nor used for classifying songs and comparing them across populations.
- A catalogue of subunits was created, highlighting the elements that remain constant across years and/or populations versus the ones that are more variable. In addition, throughout the thesis, we presented examples of subunits that are arranged in a different order to compose different and novel units.
- A record of songs was created for the East Madagascar population from 2007-2011 that were compared to songs of other populations across the world, allowing an insight on similarities and differences. Further comparisons of such songs may improve our understanding of how songs are learnt and modified through the years, and on the movements of the Madagascar population in relation to other populations in the Southern Hemisphere.
- We showed that classification based on sound units is not the best approach for categorising humpback whale songs when performing comparisons of songs across years and for different populations that may arrange building blocks in different ways. However, if one wanted to mimic the manual approach of comparing song elements at a higher

hierarchical level, for instance to compare phrases, the Hidden Markov models could be adapted to fit this purpose.

9.2 Automatic detection

The automatic detection algorithm was developed with the only goal to segment the songs recorded into their individual unit components to enable further processing of the data and to input individual calls in the automatic classifier. To achieve this, a basic energy detector was developed based on a double energy threshold. The double threshold was chosen because it has been widely observed that in most cases the loudness of humpback whale vocalisation is greater at the start of a call than at the end. Under this assumption, it is reasonable to choose a rather high energy threshold to mark the start of a vocalisation to avoid false positives that might be triggered by noise, and a lower threshold value to mark the end of the call to ensure that most of the vocalisation was captured between the two limits. Indeed, the detector developed was crude and extremely susceptible to varying the quality of recordings. As pointed out in previous chapters, the calls of humpback whales are extremely variable and some of them are characterised by frequency jumps, which by definition have small segments where the local energy is zero because the flow of air through the vocal folds is constrained temporarily. Intuitively, this causes a problem in an energy based detector because one needs to incorporate a window so that the local drop in energy will not be detected as the end threshold separating the call into smaller segments each time a frequency jump occurs. Although it is easy to apply a windowing function, which was done in the detector used in this project, this introduces a problem, which has two important implications. Applying a window means that the user has to set the window length introducing subjectivity to the automatic detector, and the choice of this parameter will also impact the two energy thresholds that are set to determine the start and end of each signal. Therefore, one is faced with a trade-off between reducing errors due to frequency jumps and choosing a window that is short enough to place the start and end markers of the call at the right locations, without missing large chunks of each unit. Moreover, as previously mentioned, the duration of humpback whale vocalisations is extremely variable which is again a factor to consider when setting an appropriate window size for the call; indeed,

choosing a window that is too long will impede the correct segmentation of the shortest calls that are 0.1 s long.

The problem of subjectivity due to human input is also linked to setting the initial threshold values for the segmentation because these two parameters need to be adjusted by the user according to the quality of the recordings to maximise the detector performance. Indeed, if the noise floor is very low one can set relatively low energy thresholds to be able to identify the exact point at which the signal starts; however, if the quality of the recording is poor, one needs to set a higher threshold which will not be triggered by distant callers or other sources of noise. The fact that humpback whales sing in a chorus in Madagascar presents a huge challenge for building a robust algorithm for the segmentation of their songs because unless one is recording an animal that is very close to the hydrophone, the recording is bound to contain signals emitted from more than one singer. Hence, during the signal processing stage, one needs to trade-off between maximising the number of calls detected whilst reducing the false positives caused by vocalisations of conspecifics that are singing in the vicinity of the focal animal. As in most field studies, the researcher needs to tailor the recording tools to ensure that it will be possible to answer the study questions with the collected data. Practically, this means that one should be aware of the limitations of the equipment available to record and of the tools available during the signal processing stage to ensure that it will be possible to obtain meaningful results. For instance, if one aims to compare the songs produced by humpback whales in a given location between years, then it is important to maximise the quality of the recordings obtained which means reducing the total amount of data obtained in each season, to be able to record a singer in isolation so that all the vocalisations present in the song sequence will be detected and included in the analysis. If the song quality deteriorates, then the detector may not be accurate and the song structure estimated using the automatic segmentation algorithm may be affected. If this was the case, then one might end up with a completely mis-read song structure, an effect similar to frame-shift errors that occur in DNA sequences during replication.

A couple of ways to improve the detector accuracy in noisy conditions are to implement a pre-whitening filter on the data to de-correlate the noise, which could also reduce the human input in terms of selecting different start and end thresholds

for each recording, and remove the effect of transducer response, which is dependent on the hydrophone and recording system used to collect the data. Indeed, such adaptation might not be effective in very high levels of noise because the algorithm would not know where to mark a sound start. Whilst noise from boats and other biological sources such as snapping shrimp and fish is an issue that can be mitigated by applying a variety of filters that are widely developed, the biggest problem faced during the detection stage is that multiple singers might be present in a recording because humpback whales sing in chorus. Whilst one can take much care to get as close as possible to the singer before starting to record, the boat will usually drift away from the singer whilst recording, or the singer might move away from the boat between song cycles. This means that even if one starts recording very close to the animal and sets the minimum gain settings to reduce the ambient noise, the data are likely to deteriorate very quickly. Hence, in most recordings, especially in an area as restricted as the channel of Ste Marie which is less than 20 km wide and very shallow, there are numerous overlapping calls being recorded and reflections of the signal due to the reverberant environment. A system for the automatic segmentation of humpback whale calls which is robust to noise was recently presented by researchers at Scripps Institute, California, which showed high level of accuracy (Probability of detection = 95% and probability of false alarms < 6%) in detecting calls buried in shipping noise (Helble *et al.*, 2012). The detector is based on the calculation of the signal energy with a generalised power low introducing some modifications to cope with non-stationary coloured noise and removing the manual input that is commonly required to set the energy thresholds. However, unsurprisingly the algorithm is not well-suited to separating calls of a focal animal buried in the chorus of calls from conspecifics because the calls from background animals are exactly the same as the ones that are produced by the focal singer. A trial of this algorithm was carried out to test if it would detect the song components accurately but the results were extremely poor because there was too much overlap between singers (*Figure 9.1*).

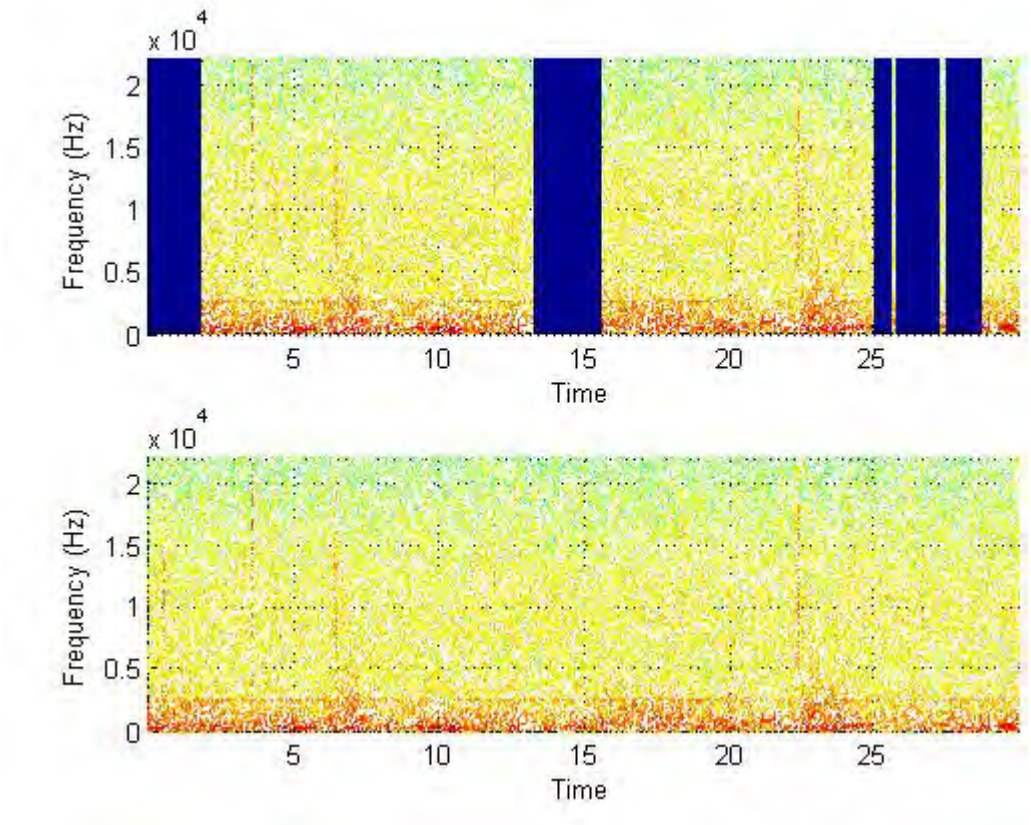


Figure 9.1: Spectrograms showing the spectrogram of a 30 seconds song segment with poor signal to noise ratio (bottom) and the detection results where the blue shading represents silence because the detector identified that portion as noise, and the spectrograms segment represent sections that are identified as one single call of humpback whale (top).

As shown in Figure 9.1, most of the energy of the units that constitute this song segment is below 5 kHz, and many calls overlap in frequency and time, making it impossible even for a human trained human observer to segment this song into its components. Thus, it is not surprising that the automatic did not cope well with such extensive overlapping of songs, which were emitted by more than 3 singers simultaneously.

To conclude, the automatic detection of humpback whale song components is feasible only if a human observer would be able to tell the calls apart. If two singers are present in a recording, it is usually possible to separate their calls apart and work out the song sequence because the calls will not be completely overlapping but some will fit in the silence of the other singer and vice versa. However, if three or more

singers are present then it becomes nearly impossible to discriminate the calls of different singers recorded with a single hydrophone because the vocalisations will overlap extensively. Therefore, in such instances, systems of multiple hydrophones will be required to separate the calls emitted by many singers present in a small area.

9.2 Automatic classification performance

Previous research on humpback whale song classifiers (Mercado III and Kuh, 1998; Deecke and Janik, 2006a; Pace *et al.*, 2009) proved that the task of automatic classification for this species is extremely challenging because, unlike other baleen whales (Mellinger and Clark, 2000; Mellinger *et al.*, 2011), their calls are extremely variable in both the time and frequency domains which means that matched filter algorithms are ill-suited for this task For this reason, this project developed a classifier based on Hidden Markov Models (HMMs) which are robust to variations in the signal duration and that have been widely developed for speech recognition making them easily accessible to users.

The results demonstrated that HMMs can be successfully employed for the classification of humpback whale songs. This method has several advantages: the training is not computationally expensive, therefore it can be carried out quickly, the process is easily repeatable and less subjective than a full manual classification, and the training set can be updated using new calls without having to run the whole algorithm from the start. In addition, a relatively small sample size, in this case approximately 150 subunits, is required to train the system which means that fewer calls have to be discarded from the analysis. This is particularly important when one wants to preserve the integrity of the song sequence to study the theme structure and to compare it to that of songs recorded from distinct populations.

In this context, it is important to note that the subunit model needs less training and, as shown for certain vocalisations in the previous Chapters 7 and 8, it allows one to identify a call which appears only 3 times in the song segment. Again, this has important implications in the inter-population comparisons because one might be able to identify the appearance of a new vocalisation within the repertoire of a specific population and investigate if this is copied from another population, giving insights into the possible migration routes followed by the whales. Indeed, analysis

of a larger dataset whose songs span over several years and different locations is needed to corroborate these ideas.

It is widely recognised that songs evolve from year to year and that some form of cultural transmission takes place since adult humpback whales are capable of learning new songs (Noad *et al.*, 2000). Changes in songs based on phrase comparisons across populations have also been recorded within a season (Cerchio *et al.*, 2001a), leading to interesting discussions on the evolution and cultural transmission of humpback whale songs.

Although all males within a breeding area sing very similar songs at any given time (Payne and McVay, 1971; Winn *et al.*, 1981), these songs change gradually within the same season with the introduction of new units, which will cause changes at all levels of the song. Such modifications result in songs being very different when one compares songs recorded in years separated by extended periods (4 or 5 years), as first noted by (Payne and Payne, 1985) and later confirmed by the Australian song revolution reported by Noad and colleagues (2000).

Subunit analysis over a wider year range will enable one to establish if the evolution of songs occurs at this lower level too and determining the rate of change of subunits over different seasons. We expect subunits to change less than units over time and, if they really can be considered the basic building blocks of the songs, they should converge to a fixed number as the dataset increases. Ultimately, the modifications at the unit level should be amenable to different associations of subunits. A similar approach based on subunit classification was used for killer whales' calls (Shapiro *et al.*, 2011) and the results obtained confirmed that analysis based on subunits helps reducing the computational load and that stereotyped calls often contained different combinations of subunits (in 75% of the cases). Although, killer whales do not produce call sequences to form songs, the variety of calls they produce present some common characteristics with the sound units of humpback whales, including frequency jumps and fast sweeps. The results comparing the performance of the automatic classification showed that the model based on subunit recognition was more successful than the unit model at correctly classifying the individual song components overall and also required less training to be carried out when comparing songs of different singers as they sometimes form different association of subunits to constitute new units. These findings were consistent across the analyses carried out:

overall the subunit classification was better in the results obtained on the analysis conducted on Madagascar songs alone and also when the HMMs were trained using Hawaiian data to recognise vocalisations of songs from different years and populations. In a few instances, looking at the classification performance for each call type the unit model out-performed the subunit one, particularly if some instances of the calls emitted from one singer were included in the training set to test a song produced by the same singer.

The reliability of the Hidden Markov Modelling results was tested by replicating tests on the same songs using different training and test datasets chosen at random to assess the error after each trial, as described in the Chapters 7 and 8. The recognition appeared to be consistent across trials; the performance diverged by 1, or rarely 2, calls per class from one trial to another. Nevertheless, the error in some cases was large because the total number of calls tested for a particular sound type was very small, meaning that the proportion of correct to incorrect calls could vary significantly if there was a difference in classification performance of one vocalisation. Indeed, with larger datasets for each call type we could have obtained a better estimation of the reliability of the automatic classification algorithm and if this remains constant as training and testing datasets are randomly chosen. This is quite an important factor to be established especially given that humpback whale vocalisations are extremely variable and do evolve in structure throughout a song because one needs to be sure that including peculiar vocalisations in either the training or testing set will not impact the results negatively.

The impact of changes in the training set on the automatic classification results was also tested with three scenarios in which the HMMs were trained using different numbers of calls for the training stage. Knowing how the classifier performs using different proportion of training and testing data is crucial for the practical application of the algorithm. Indeed, it is universally accepted that the more the data used for training, the better a classifier should perform because by increasing the amount of training one will obtain a more accurate representation of the statistical properties of that particular population, given the characteristics of the observations fed into the system. However, it is impractical for an end user to train the model using hundreds or thousands of observations because it is a very time consuming process, and often the data available to the analysis will not be enough to do so, as in the case of this

project. Indeed, it should be possible to identify a level of training past which the increase in accuracy starts levelling off so that an increase in amount of training will result only in a very slight increase in accuracy. Therefore, we compared the performance of the automatic classifier using different training percentages to be able to inform users regarding the trade-off between accuracy and manual labour. The results presented both for Madagascar and Hawaiian songs showed that the performance of the classifier was higher for most sound categories when the HMMs were trained using 50% of the dataset of each call type. Nonetheless, high levels of classification performance were also achieved by using only 25% of the data for the training, both overall and looking at the call categories individually with 80% or more of the calls being correctly classified. On the other hand, when we reduced the training set to 10% of dataset for each call type, the individual performance of the classifier per call type reduced significantly in several of the categories, despite the fact that the overall classification results did not diverge hugely between trials. This is because the overall classification performance of the classifier was calculated as the percentage of total calls correctly classified over the total number of calls tested; specifically, this means that if a few calls were classified incorrectly in categories which contained only a few calls, the individual classification performance for that particular call class would be very much lower, but overall the percentage of misclassified calls would increase by a small amount. In other words, when assessing the performance of any automatic classification algorithm that deals with vocalisations as varied as those of humpback whales it is necessary to take into account not only the values obtained for the overall number of calls correctly classified, but to look at the distribution of incorrect calls and if some sound types are always or never ascribed to the wrong category. The results obtained in this project showed that some vocalisation types were classified correctly more often than others. Whilst some variation in the performance can be expected partly because the amount of calls included in the training set is different for each call category, however, instead of depending on the number of calls used in the training which is a positive indicator of the reliability of the system, the performance of the classifier also depends on the nature of the calls. In particular, broadband calls are mis-classfied more often than any other type of calls. This is probably due to the fact that Mel-Frequency Cepstrum Coefficients (MFCCs) are more suited to

characterising harmonic signals, being based on the Fourier transform on the signal. In addition, in noisy conditions, the frequency bands of broadband calls are often corrupted by noise because they occupy a broader frequency range than harmonic calls. Furthermore, broadband calls include both the longest and the shortest vocalisations identified within songs, and the MFCCs are calculated using a fixed window length. Indeed, it would be useful to be able to modify the frame length accordingly to the type of call to be classified. However, this might be difficult to realise practically without introducing the need for some prior knowledge about the signals. One way in which this could be achieved would be to run a preliminary analysis during which all the vocalisations inputted in the algorithm are split into broadband and non-broadband, similarly to how speech is divided into voiced or unvoiced in some tasks. Once the calls are ascribed to these pre-defined categories, then one could proceed with the calculation of the MFCCs with different window length for each of the two groups. One could even decide to use different types of features for the two groups. This would significantly complicate the analysis following this step given the way in which HTK is structured.

For the purpose of this thesis we analysed the automatic classifier performance based on the classification of individual calls and investigate its effectiveness using two 'vocabularies' based on the traditional unit segmentation and the subunit approach. However, given that HTK allows introducing additional modules for sentence recognition systems for instance, one could extend the algorithm to classify the songs based on sequences of units or subunits. Since the earliest stages, researchers noted that the songs of humpback whales appeared have a hierarchical organisation, hence the definition of song proposed by Payne and McVay (1971), which was recently confirmed quantitatively using an information theoretic analysis (Suzuki *et al.*, 2006). In a later study, entropy was used to quantify redundant information of the songs and to validate the observations on the hierarchical structure of humpback whale songs (Suzuki *et al.*, 2006; Miksis-Olds *et al.*, 2008). The fact that songs are formed by a hierarchy of sound components suggests that the automatic classification effort could be extended to longer dependencies of songs than the units or subunits, for instance phrases, which may contain biologically meaningful information. Indeed, the hierarchy of sounds could be learned by the HMMs during the training stage. Song classification using HMMs could be extended to the analysis

of more complex song components, such as phrases. Phrases are units that are juxtaposed to form sequences that are repeated several times to form song themes (Payne and McVay, 1971). The length of phrases is variable both in terms of duration and in the number of units that are associated to form it. Typically they are between 2 and 6 units long and last between 5-30 seconds. Research has focused on the study of phrases in an attempt to understand the pattern of song evolution and transmission. Using HTK, it is possible to train the models to recognise sequences of HMMs by simply tuning the language structure. This function was created to recognise words or entire sentences for speech recognition purposes. Intuitively, the system could be adapted to reflect the hierarchy of humpback whale songs to recognise, for instance, phrases. Although the increased level of complexity will require a larger training set and a greater computational load, it would help processing large amounts of data objectively but safeguarding the biological significance of the analysis. A pilot study for the recognition of phrases of humpback whales songs of Madagascar using HMMs was carried out by William Scott-Hartley in the past few months as part of an undergraduate project at the Institute of Sound and Vibration Research, showing promising results. The classifier developed treated the sound components as a string of units, from which the silent intervals were removed and was successful and identifying most phrases within a song given a training set of a few samples of the labelled components of individual phrases. This was achieved applying only few modifications of the algorithm at the grammar level; in other words, the new structure to be recognised was defined in the HTK to include the new hierarchical structure of the sounds that were found in the dataset presented to the algorithm. Although problems such as the fact that some phrases every so often contain a variable number of repetitions of a particular vocalisation were not addressed, the methodology shows the huge potential of applying HMMs to the recognition of humpback whale songs.

9.3 Comparison of songs of Madagascar across years

In Chapter 7, the songs recorded in Madagascar between 2006 and 2011 were compared to understand how song in this region evolve. The spectrograms of the song of 1996 presented in Razafindrakoto (2001) could not be inspected in detail and included in the analysis because the quality of the figures were of poor and only one

representation of phrase was presented with no detailed description of how they varied within and between songs. The analysis conducted showed that songs were composed of a varying number of themes, some of which were short and repeated more several times within a song, whereas others were constituted by more themes which were repeated only once per song. The evidence presented shows that some themes are preserved more than others across years. Whilst we cannot confirm if this fact applied to the songs of other populations of humpback whales or if this is true long term in any of the breeding grounds, it opens the debate as to whether some themes are less affected by the evolution process. If whales conserve some themes of their songs in preference to others, there are several factors to consider both in terms of the cognition capability of these whales and the general song usage and purpose.

The fact that some themes are conserved for a longer period than other themes across years may be due to chance or to a decision of the whale to produce those themes specifically. The theory that themes with a high level of similarity are repeated for 5 consecutive years as a result of chance seems extremely unlikely; however, more robust evidence is needed to implicate a decision of the whale to repeat certain themes over others for several years. If whales chose to preserve specific themes for a long period of time compared to the general evolution of the song structure with time, it means that there must be some advantage to doing so. For instance, certain themes might just have a simpler structure than others or be easier to remember and therefore have a lower cost of production to the singer. Alternatively, the common themes could serve some purpose in the communication amongst whales that belong to the same population.

9.4 Comparison of individual vocalisations that constitute humpback whale songs

Subunits were defined here for the first time as the shortest continuous sound that can be encountered on its own or in association with other subunits within a song. The frequency characteristics of a subunit are less variable than those of a unit; therefore they should be more easily classified using automatic algorithms. There are similarities between a subunit a phoneme in speech analysis; phonemes being the building blocks of human language. By drawing this comparison with speech we are not implying that humpback whales convey their mental representation via a sound;

nor are we suggesting that we are able to assign the meaning of the units that constitute a song by distinguishing their subunit components. We merely aim at describing humpback songs through less complex blocks which eases the automatic classification task by reducing the number of components necessary to describe the wide variety of calls produced by these marine mammals.

The analysis based on subunits rather than units appears to improve the classification of humpback whales vocalisations. Indeed, according to our definition subunits are less variable than units and they are usually of shorter duration. This fact allows one to more accurately model them with stationary models. Subunits should be able to describe the whole repertoire of calls. This means that subunits should be repeated from year to year, whereas the units may change. Comparison of songs collected over different years showed that a larger proportion of subunits were preserved during the years compared to the number of units; however, given the limited data available for the analysis it was not possible to converge towards an invariant number of components that are able to represent the entire vocabulary of humpback whales. The dataset was integrated with songs obtained from other regions of the World, namely Hawaii and Mexico, which showed that some of the units and subunits are shared amongst populations even though they are geographically isolated from each other and songs have never been shown to contain shared phrase or themes across different ocean basins. Again, the number of shared subunits was greater than the number of shared units suggesting that using the subunit model for recognition can lead to advantages in reducing the manual input and therefore the time effort at the training stage. Although at this stage it is not possible to apply the classification algorithm developed based on a fixed predefined training set that would match the vocabulary of all humpback whale populations, we showed that it is possible to achieve high level of classification performance integrating the training set with samples of the new calls that appear in each different song. More detailed information about the dynamics in which the song evolves is required to be able to achieve the goal of a universally applicable automatic recogniser that needs no updating of the training set. At present, research on song evolution has focused on the phrase and theme levels, possibly because it is much easier and less time consuming to compare dataset over these higher levels of the hierarchy than it is to look at the details of their building blocks. However, it would be extremely valuable

to obtain a detailed picture the variability of units and subunits between individuals that perform the same song, how these change between years, and if the same amount of variability is encountered across populations. Some studies showed the geographical variability of songs meaning that humpback whales in different regions may sing the same song but the themes or phrases within these songs might contain slightly different units or different numbers of the same units, forming what they term 'dialects'. Nonetheless, the details regarding the variability of units, which could be compared to pronunciation difference, has been overlooked. Certainly the fact that little is known about the mechanism of sound production in this species, and that we cannot associate a meaning with the sounds emitted makes it difficult to conduct comparative studies on the building blocks of humpback whale songs. However, research focussed on the behaviour associated with social sounds will give an insight on the intentions of whales when they emit particular calls (Dunlop *et al.*, 2007c; Dunlop *et al.*, 2008). If calls with similar structure are associated with the same behaviour repeatedly, then one can assume that such sounds are the same and they might be 'dialects' of the same sounds.

9.5 Future work

The work presented in this thesis shows that Hidden Markov Models can be employed for recognising calls of humpback whale songs with high levels of precision across a variety of sources, despite the limitations highlighted in the previous sections.

Future research could be focused on the development of the classification algorithm to recognise elements of humpback whales songs at a different level of the hierarchy. Because biologists tend to classify the phrases within songs, it would be useful to build this level of song hierarchy within the recognition grammar of the HTK so that one could immediately visualise the sequence of phrases that make up the theme, and ultimately the song for that year and location. This added level of complexity in the algorithm could be implemented within the HTK toolkit with minor modifications to the algorithm from the methods presented in this thesis. However, putting together the training set of phrases for such models would be very labour intensive.

Another area of development for the model would be to include a class of unidentified calls within the classification algorithm so that, if unsure, the model

could class new calls within this category rather than trying to fit them within existing call classes.

Furthermore, future work could be dedicated to developing a way to interface the results of the HMM classification with some sound analysis software (e.g. Adobe Audition) that is commonly employed by researchers for inspecting their recordings so that the results of the classification would be embedded, for instance, within their Spectrogram or waveform viewer as they scroll through the recording. This could be achieved by turning the results of the HMM recognition into marker cues that are appended to the wav files. Such application would make any recognition algorithm really useful and readily implemented by groups who are working on humpback whale songs across the globe. This could be taken even further by running the algorithm in real-time on a vessel as one listens to the songs, giving near real-time song classification.

Appendix I

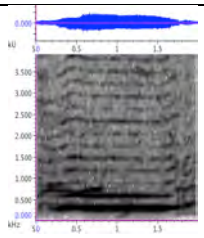
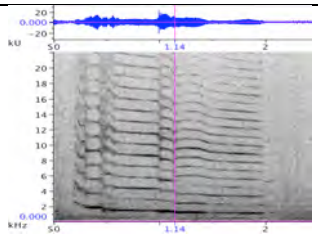
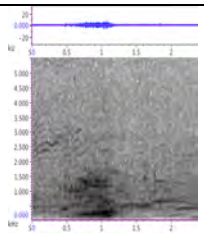
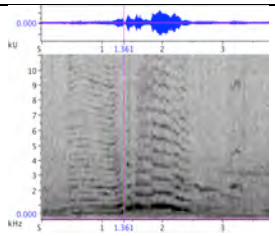
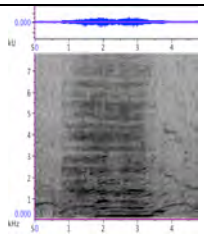
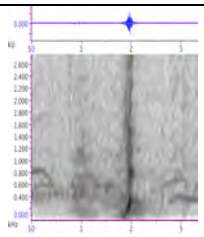
Matlab script for the segmentation of songs into their component units.

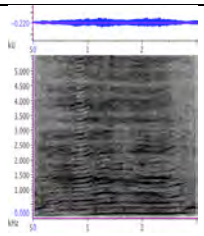
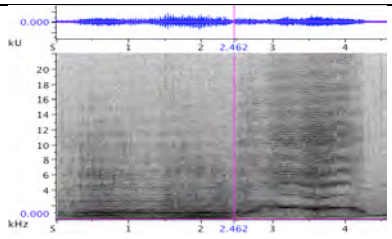
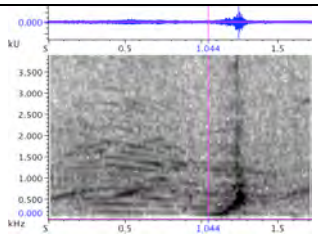
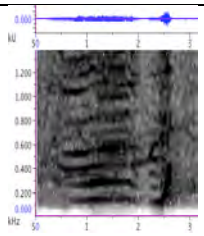
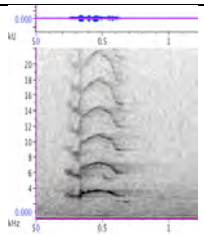
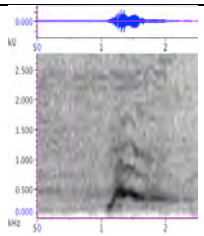
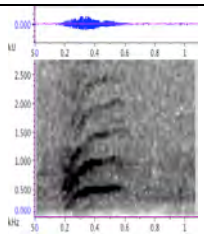
```
fs=44100;

tseg=30;
rec_length=901*fs;
kstart=1;
kstop=kstart+(tseg*fs);

j=1;
while kstop<=rec_length

    xf=wavread('08h50',[kstart kstop]); % only get a segment of the
data

    t=[0:(length(xf)-1)]/fs;
    L=512; % window length
    [C,Ekm]=kleiwer_mertins(xf,L); % estimation of the signal energy
    C=Ekm;


    %% calculation of the energy with double-thresholding system %%
    tb=0.005; % threshold of signal beginning
    te=0.0005;
    E=zeros(size(C));
    D=E;
    %figure(1),hold on,plot([0 30],[1 1]*tb) % to plot threshold
line on the
    %energy plot
    det_flag=0;
    for n=1:length(C)
        if C(n)>tb
            det_flag=1;
        else
            if (det_flag==1)
                if (C(n)<te)
                    det_flag=0;
                end
            end
        end

        if (det_flag==1)
            E(n)=C(n);
        end
        D(n)=det_flag;
    end

    %% find position of start/end units %%%

    dD(:,j)=diff([0;D(:);0]);
    positions(j).start=find(dD(:,j)==1); % vector of start position
of each vocalisation
    positions(j).stop=find(dD(:,j)==-1)-1;

    kstart=kstart+tseg*fs;
    kstop=kstop+tseg*fs;
```

```
    j=j+1;

end

%%
n=length(vocals);
for k=1:n
    voc=xf(vocals(k,1):vocals(k,2));
    LPCs(k,:)=lpc(voc,20); % matrix with lpc coefficients. each row
is a different vocalisation
end
n=length(vocals);
banks=20;
noverlap=50;
window=hanning(512);
for k=1:n
    voc=xf(vocals(k,1):vocals(k,2));
    mfccs(k,:)=mymfcc_prw(voc,window,noverlap,banks,fs); % structure
with mfcc coefficients. each column is a different vocalisation
end
```

Appendix 2

Catalogue of calls present in Mada 2009a song.

Unit label	**Subunit labels**	**Spectrogram (frequency resolution 43 Hz/bin) Pink line represents split between subunits**	**Fundamental frequency (Hz) Start (end)**	**N. of HMM states Unit (subunit)**
A	A		290	1 (1)
BC	B AND C		2500 (1000)	3 (B=2 , C=1)
D	D		100 (Energy up to 1290)	2 (2)
FG	F AND G		370 (990)	3 (F=1, G=2)
H	H		190	1 (1)
L	L		62 (400)	1 (1)

M	M		160	2 (2)
PO	P AND O		706 (1500)	2 (P=2 , O=2)
NL	N AND L		130 (800)	2 (N=1, L=1)
TL	T AND L		120 (450)	2 (T=1, L=1)
U	U		2500 same start and end; max frequency =3700	3 (3)
Q	Q		120 (360)	2 (2)
S	S		90 (520)	2 (2)

Appendix 3

Results of each round of tests for the ten-fold validation conducted on recording Mada2009a.

Unit	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9	Round 10	Mean	Sd
A	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	85.7	**98.6**	4.5
FG	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	50.0	100.0	**95.0**	15.8
B		100.0	0.0	100.0					100.0	0.0	**60.0**	54.8
BC		100.0	100.0	100.0	100.0	100.0		100.0	100.0	75.0	**96.9**	8.8
D	100.0	100.0	100.0	100.0		100.0		100.0	100.0		**100.0**	0.0
H	100.0	100.0	50.0	100.0	75.0		100.0	75.0	50.0	100.0	**83.3**	21.7
L	85.7	100.0		75.0	66.7	100.0	100.0	83.3	0.0	100.0	**79.0**	32.1
M	100.0	100.0	100.0	0.0	0.0		100.0	100.0	100.0	100.0	**77.8**	44.1
PO	100.0		100.0	100.0	100.0	66.7	100.0		100.0		**95.2**	12.6
Q	100.0	100.0	100.0		100.0	100.0	100.0	100.0	100.0	100.0	**100.0**	0.0
S	100.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**97.5**	7.9
TL	100.0		0.0	100.0	50.0	50.0	100.0	100.0	100.0	0.0	**66.7**	43.3
U	100.0	100.0	100.0	100.0		100.0	100.0	50.0	100.0		**93.8**	17.7
TOTAL	96.8	96.8	83.9	87.1	87.1	90.3	93.5	90.3	87.5	87.5	**90.1**	4.4

Subunit	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9	Round 10	Mean	Sd
A	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
B		100.0	66.7	100.0	100.0	100.0		100.0	33.3	40.0	80.0	27.3
C		100.0	100.0	100.0	100.0	66.7		100.0	100.0	100.0	95.8	11.0
D	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
F	100.0	100.0	100.0	71.4	100.0	66.7	100.0	100.0	100.0	100.0	93.8	12.4
G	75.0	100.0	100.0	71.4	100.0	66.7	100.0	100.0	50.0	100.0	86.3	17.8
H	100.0	100.0	100.0	100.0	50.0		100.0	75.0	50.0	100.0	86.1	20.8
L	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.0	50.0	100.0	94.0	15.0
M	100.0	100.0	100.0	66.7	0.0		100.0	100.0	100.0	100.0	85.2	31.9
O	100.0		100.0	100.0	100.0	66.7	100.0		100.0		95.2	11.7
P	100.0		100.0	50.0	100.0	100.0	100.0		0.0		78.6	36.4
Q	100.0	100.0	100.0		100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0
S	100.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.5	7.5
T	100.0		0.0	66.7	50.0	50.0	100.0	100.0	100.0	0.0	63.0	39.1
U	100.0	100.0	100.0	100.0		100.0	100.0	100.0	100.0		100.0	0.0
TOTAL	97.4	97.1	94.7	84.4	95.2	88.1	100.0	94.1	80.5	90.0	89.7	6.0

Reference list

Abbot, T. A., Premus, V. E., and Abbot, P. A. (**2010**). "A real-time method for autonomous passive acoustic detection-classification of humpback whales," The Journal of the Acoustical Society of America **127**, 2894-2903.

Arraut, E. M., and Vielliard, J. M. E. (**2004**). "The song of the Brazilian population of Humpback Whale *Megaptera novaeangliae*, in the year 2000: individual song variations and possible implications," An Acad Bras Cienc **76**, 373-380.

Au, W. W. L., Pack, A. A., Lammers, M. O., Herman, L. M., Deakos, M. H., and Andrews, K. (**2006**). "Acoustic properties of humpback whale songs," The Journal of the Acoustical Society of America **120**, 1103-1110.

Baum, J. F., Petrie, T., Souler, G., and Weiss, N. (**1970**). "A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains," Am. Math. Statist. **41**, 164-171.

Beskow, J., and Sjolander, K. (**2000**). "School of Computer Science."

Brown, J. C., Hodgins-Davis, A., and Miller, P. J. O. (**2006**). "Classification of vocalizations of killer whales using dynamic time warping," The Journal of the Acoustical Society of America **119**, EL34-EL40.

Brown, J. C., and Miller, P. J. O. (**2007**). "Automatic classification of killer whale vocalizations using dynamic time warping," The Journal of the Acoustical Society of America **122**, 1201-1207.

Brown, J. C., and Smaragdis, P. (**2009**). "Hidden Markov and Gaussian mixture models for automatic call classification," The Journal of the Acoustical Society of America EL221-EL224.

Cerchio, S., Ersts, P., Pomilla, C., Loo, J., Razafindrakoto, Y., Leslie, M., Andrianirivelo, N., Minton, G., Dushane, J., Murray, A., Collins, T., and Rosenbaum, H. C. (**2008**). "Revised estimation of abundance for breedingstock C3 of humpback whales, assessed through photographic and genotypic mark-recapture datafrom Antongil Bay, Madagascar," SC/60/SH32 presented to the to the IWC.

Cerchio, S., Jacobsen, J. K., Cholewiak, D. M., Falcone, E. A., and Merriwether, D. A. (**2005**). "Paternity in humpback whales, Megaptera novaeangliae: assessing polygyny and skew in male reproductive success," Animal Behaviour **70**, 267-277.

Cerchio, S., Jacobsen, J. K., and Norris, T. F. (**2001a**). "Temporal and geographical variation in songs of humpback whales, Megaptera novaeangliae: synchronous change in Hawaiian and Mexican breeding assemblages," Animal behaviour **62**, 313-329.

Cerchio, S., Jacobsen, J. K., and Norris, T. F. (**2001b**). "Temporal and geographical variation in songs of humpback whales, Megaptera novaeangliae: synchronous change in Hawaiian and Mexican breeding assemblages," Animal behaviour **62**, 313-329.

Constantine, R., Russell, K., Gibbs, N., Childerhouse, S., and Baker, C. S. (**2007**). "Photo-identification of Humpback Whales (Megaptera novaeangliae) in New Zeland and their migratory connections to breeding grounds of Oceania," Marine Mammal Science **23**, 715-720.

Cranford, T. W., Amundin, M., and Norris, K. S. (**1996**). "Functional morphology and homology in the odontocete nasal complex: Implications for sound generation," Journal of Morphology **228**, 223-285.
Darling, J. D., and Berube, M. (**2001**). "Interactions of singing humpback whales with other males," Marine Mammal Science **17**, 570-584.
Deecke, V. B., Ford, J. K. B., and Spong, P. (**1999**). "Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (Orcinus orca) dialects," The Journal of the Acoustical Society of America **105**, 2499-2507.
Deecke, V. B., and Janik, V. M. (**2006a**). "Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls," The Journal of the Acoustical Society of America **119**, 645-653.
Deecke, V. B., and Janik, V. M. (**2006b**). "Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls," J The Journal of the Acoustical Society of America, 645-653.
Deller, J. R., Proakis, J. G., and Hansen, J. H. L. (**1993**). *Discrete-time processing of speech signals* (Macmillian Publishing Company, New York).
Dunlop, R., Noad, M. J., Cato, D., and D, S. (**2007a**). "The social vocalization repertoire of east Australian migrating," The Journal of the Acoustical Society of America, 2893-2905.
Dunlop, R. A., Cato, D. H., and Noad, M. J. (**2008**). "Non-song acoustic communication in migrating humpback whales *Megaptera novaeangliae*," Marine Mammal Science **24**, 613-629.
Dunlop, R. A., Noad, M. J., Cato, D. H., and Stokes, D. (**2007b**). "The social vocalization repertoire of east Australian migrating humpback whales (Megaptera novaeangliae)," The Journal of the Acoustical Society of America, 2893-2905.
Dunlop, R. A., Noad, M. J., Cato, D. H., and Stokes, D. (**2007c**). "The social vocalization repertoire of east Australian migrating humpback whales (Megaptera novaeangliae)," The Journal of the Acoustical Society of America **122**, 2893-2905.
Erbe, C. (**2002**). "UNDERWATER NOISE OF WHALE-WATCHING BOATS AND POTENTIAL EFFECTS ON KILLER WHALES (ORCINUS ORCA), BASED ON AN ACOUSTIC IMPACT MODEL," Marine Mammal Science **18**, 394-418.
Fagot, R. F. (**1961**). "A model for equisection scaling," Behavioral Science **6**, 127-133.
Fitch, W. T., Neubauer, J., and Herzel, H. (**2002**). "Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production," Animal Behaviour **63**, 407-418.
Forney, G. D. (**1978**). "The Viterbi Algorithm," Proceedings of the IEEE **61**.
Frazer, L. N., Frazer, L. N., and Mercado, E., III (**2000**). "A sonar model for humpback whale song," Oceanic Engineering, IEEE Journal of **25**, 160-182.
Garland, Ellen C., Goldizen, Anne W., Rekdahl, Melinda L., Constantine, R., Garrigue, C., Hauser, Nan D., Poole, M. M., Robbins, J., and Noad, Michael J. (**2011**). "Dynamic Horizontal Cultural Transmission of

Humpback Whale Song at the Ocean Basin Scale," Current biology : CB **21**, 687-691.
Gedamke, J., Costa, D. P., and Dunstan, A. (**2001**). "Localization and visual verification of a complex minke whale vocalization," The Journal of the Acoustical Society of America **109**, 3038-3047.
Gillespie, D. (**2004**). "Detection and classification of right whale calls using an “edge” detector operating on a smoothed spectrogram," Canadian Journal of Acoustics **32**, 39-47.
Gold, B., and Morgan, N. (**2000**). *Speech and audio signal processing. Processing and perception of speech and music.* (John Wiley & Sons Inc., USA).
Helble, T. A., Ierley, G. R., D'Spain, G. L., Roch, M. A., and Hildebrand, J. A. (**2012**). "A generalized power-law detection algorithm for humpback whale vocalizations," The Journal of the Acoustical Society of America **131**, 2682-2699.
Helweg, D. A. (**1996**). "Geographic and temporal variation in songs of humpback whales," The Journal of the Acoustical Society of America **100**, 2609-2609.
Helweg, D. A., Cato, D. H., jenkins, P. F., Garrigue, C., and McCauley, R. D. (**1998**). "Geographic Variation in South Pacific Humpback Whale Songs," Behaviour **135**, 1-27.
Holmes, J. N. (**1988**). *Speech synthesis and recognition* (Van Nostrand Reinhold, Wokingham).
IWC, S. C. (**1996**). "General Principles for Whalewatching," (IWC website).
Janik, V. M. (**2000**). "Source levels and the estimated active space of bottlenose dolphin (Tursiops truncatus) whistles in the Moray Firth, Scotland," Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology **186**, 673-680.
Janik, V. M., and Slater, P. J. (**1998**). "Context-specific use suggests that bottlenose dolphin signature whistles are cohesion calls," Animal behaviour **56**, 829-838.
Johnson, M. P., and Tyack, P. L. (**2003**). "A Digital Acoutic Recording Tag for Measuring the Response of Wild Marine Mammals to Sound," IEEE Journal of Oceanic Engineering **28**, 3-12.
Jurafsky, D., and Martin, J. H. (**2009**). *Speech and language processing*.
Ketten, D. R. (**1994**). "Functional Analyses of Whale Ears: Adaptations for Underwater Hearing," in *OCEANS 1994*, pp. I264-I270.
Kogan, J. A., and Morgan, D. (**1998**). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," The Journal of the Acoustical Society of America **103**, 2185-2196.
Leighton, T. G. (**1998**). "Fundamentals of underwater acoustics," in *Fundamentals of noise and vibration*, edited by F. Fahy, and J. Walker (Spon Press, London), pp. 373-444.
Leigthon, T. G., Richards, S. D., and White, P. R. (**2004**). "Trapped within a ‘wall of sound’. A possible mechanism for the bubble nets of humpback whales," Acoustics bulettin **29**, 23-29.
Mazhar, S., Ura, T., and Bahl, R. (**2007**). "Vocalization based Individual Classification of Humpback Whales using Support Vector Machine," in *Oceans 2007*, pp. 1-9.

Mazhar, S., Ura, T., and Bahl, R. (**2008a**). "An analysis of Humpback whale songs for individual classification," The Journal of the Acoustical Society of America **123**, 3774-3774.
Mazhar, S., Ura, T., and Bahl, R. (**2008b**). "Effect of Temporal Evolution of Songs on Cepstrum-based Voice Signature in Humpback Whales," in *OCEANS 2008 - MTS/IEEE Kobe Techno-Ocean*, pp. 1-8.
Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," The Journal of the Acoustical Society of America **107**, 3518-3529.
Mellinger, D. K., Martin, S. W., Morrissey, R. P., Thomas, L., and Yosco, J. J. (**2011**). "A method for detecting whistles, moans, and other frequency contour sounds," The Journal of the Acoustical Society of America **129**, 4055-4061.
Mercado III, E., Green, S. R., and Schneider, J. N. (**2008**). "Understanding auditory distance estimation by humpback whales: A computational approach," Behavioural Processes **77**, 231-242.
Mercado III, E., Herman, L. M., and Pack, A. A. (**2003**). "Stereotypical sound patterns in humpback whale songs: Usage and function," Aquatic Mammals **29**, 37-52.
Mercado III, E., and Kuh, A. (**1998**). "Classification of humpback whale vocalizations using a self-organizing neural network," in *IEEE World Congress on Computational Intelligence*, edited by N. N. Proceedings.
Mercado III, E., Schneider, J. N., Pack, A. A., and Herman, L. M. (**2010**). "Sound production by singing humpback whales," The Journal of the Acoustical Society of America **127**, 2678-2691.
Miksis-Olds, J. L., Buck, J. R., Noad, M. J., Cato, D. H., and Stokes, M. D. (**2008**). "Information theory analysis of Australian humpback whale song," The Journal of the Acoustical Society of America **124**, 2385-2393.
Miller, P. (**2006**). "Diversity in sound pressure levels and estimated active space of resident killer whale vocalizations," Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology **192**, 449-459.
Miller, P. J. O., and Bain, D. E. (**2000**). "Within-pod variation in the sound production of a pod of killer whales, Orcinus orca," Animal behaviour **60**, 617-628.
Murray, A., Cerchio, S., McCauley, R., Jenner, C. S., Razafindrakoto, Y., Coughran, D., McKay, S., and Rosenbaum, H. (**2012**). "Minimal similarity in songs suggests limited exchange between humpback whales (Megaptera novaeangliae) in the southern Indian Ocean," Marine Mammal Science **28**, E41-E57.
NOAA, N. O. a. A. A. (**1991**). "Final Recovery Plan for the Humpback Whale (*Megaptera novaeangliae*)," edited by H. W. R. T. f. t. N. M. F. Service (Silver Spring, Maryland), pp. 1-105.
Noad, M. J., Cato, D. H., Bryden, M. M., Jenner, M. N., and Jenner, K. C. S. (**2000**). "Cultural revolution in whale songs," Nature **408**, 537-537.
Norris, T. F., Jacobsen, J. K., and Cerchio, S. (**2000**). "A comparative analysis of humpback while songs recorded in pelagic waters of the eastern north pacific: preliminary findings and implications for

discerning migratory routes and assessing breeding stock identity," in *NOAA Technical Memorandum*.
Nowak, M. A., Plotkin, J. B., and Jansen, V. A. A. (**2000**). "The evolution of syntactic communication," Nature **404**, 495-498.
Pace, F., Benard, F., Glotin, H., Adam, O., and White, P. (**2010**). "Subunit definition and analysis for humpback whale call classification," Applied Acoustics **71**, 1107-1112.
Pace, F., White, P. R., and Adam, O. (**2009**). "Comparison of feature sets for humpback whale song analysis," in *Bio-acoustics 2009* (Institute of Acoustics, Loughborough), pp. 136-144.
Pace, F., White, P. R., and Adam, O. (**2011**). "Hidden Markov Models for classification of humpback whale songs collected in Ste Marie, Madagascar, over three years.," The Journal of the Acoustical Society of America **In Review**.
Parsons, E. C. M., Wright, A. J., and Gore, M. A. (**2008**). "The Nature of Humpback WHlae (Megaptera novaeangliae) Song," Journal of Marine Animals and Their Ecology **1**, 22-31.
Payne, K., and Payne, R. (**1985**). "Large Scale Changes over 19 Years in Songs of Humpback Whales in Bermuda," Zeitschrift für Tierpsychologie **68**, 89-114.
Payne, R., and Mc Vay, S. (**1971**). "Songs of Humpback Whales," Science, 585-597.
Picone, J. (**1990**). "Continuous speech recognition using Hidden Markov Models.," IEEE ASSP Magazine **7**, 26-41.
Potter, J. R., Mellinger, D. K., and Clark, C. W. (**1994**). "Marine mammal call discrimination using artificial neural networks," The Journal of the Acoustical Society of America **96**, 1255-1262.
Rabiner, L. R. (**1989**). "A tutorial on Hidden Markov Models and selected applications in speech recognition," in *IEEE*, pp. 257-286.
Rabiner, L. R., and Juang, B.-H. (**1993**). *Fundamentals of speech recognition* (Prentice Hall Inc., London).
Rabiner, L. R., and Shafer, R. W. (**1978**). *Digital processing of speech signals* (Prentice-Hall International Inc., London).
Rabiner, L. R., and Shafner, R. W. (**2011**). *Theory and Applications of Digital Speech Processing* (Prentice-Hall Inc).
Razafindrakoto, Y. (**2001**). "First description of humpback whale song from Antongil Bay, Madagascar," Marine Mammal Science **17**, 180-186.
Reilly, S. B., J.L., B., Best, P. B., Brown, M., Brownell Jr., R. L., Butterworth, D. S., Clapham, P. J., Cooke, J., Donovan, G. P., Urbán, J., and Zerbini, A. N. (**2008**). "*Megaptera Novaeangliae*. In: IUCN 2009," (IUCN).
Reindenberg, J. S., and Laitman, J. T. (**2007**). "Discovery of a low frequency sound source in Mysticeti (baleen whales): anatomical establishment of a vocal fold homolog," The Anatomical Record **290**, 745-759.
Ren, Y., Johnson, P. C., Darre, M., Suart Glaeser, S., Osiejuk, T. S., and Out-Nyarko, E. (**2009**). "A Framework for bioacoustic Vocalization Analysis Using Hidden Markov Models," Algorithms, 1410-1428.
Reynolds, J. E., and Rommel, S. A. (**1999**). *Biology of Marine Mammals* (Smith-sonians Institution Press, Washington, DC).

Rickwood, P., and Taylor, A. (**2008**). "Methods for automatically analyzing humpback song units," The Journal of the Acoustical Society of America **123**, 1763-1772.
Rosenbaum, H. C., Walsh, P. D., Razafindrakoto, Y., Vely, M., and DeSalle, R. (**1997**). "First description of a humpback whale breeding ground in Baie d'Antongil, Madagascar," Conservation Biology **11**, 312-314.
Seekings, P., and Potter, J. (**2008**). "Classification of marine acoustic signals using Wavelets & Neural Networks," in *Proceeding of 8th Western Pacific Acoustics conference (Wespac8)* (Australia).
Seekings, P., and Potter, J. R. (**2003**). "Classification of marine acoustic signals using Wavelets & Neural Networks," (Proceeding of 8th Western Pacific Acoustics conference (Wespac8), Australia).
Shapiro, A. D., Tyack, P. L., and Seneff, S. (**2011**). "Comparing call-based versus subunit-based methods for categorizing Norwegian killer whale, Orcinus orca, vocalizations," Animal behaviour **81**, 377-386.
Slater, P. J. (**2001**). *Essentials of Animal Behaviour* (Cambridge University Press).
Smith, J. N., Goldizen, A. W., Dunlop, R. A., and Noad, M. J. (**2008**). "Songs of male humpback whales, Megaptera novaeangliae, are involved in intersexual interactions," Animal Behaviour **76**, 467-477.
Sousa-Lima, R., and Clark, C. W. (**2009**). "Whale sound recording technology as a tool for assessing the effects of boat noise in Brazilian marine park," Park Science **26**, 59-63.
Southall, B., Ann E. Bowles, William T. Ellison, James J. Finneran, Roger L. Gentry, Charles R. Greene Jr., Kastak, D., Darlene R. Ketten, James H. Miller, Paul E. Nachtigall, W. John Richardson, Jeanette A. Thomas, and Tyack, P. L. (**2007**). "Marine mammal noise exposure criteria: Initial scientific recommendations," Aquatic Mammals **33**, 411-521.
Stevick, P. T., Neves, M. C., Johansen, F., Engel, M. H., Allen, J., Marcondes, M. C. C., and Carlson, C. (**2011**). "A quarter of a world away: female humpback whale moves 10 000 km between breeding areas," Biology Letters **7**, 299-302.
Stimpert, A. K., Au, W. W. L., Parks, S. E., Hurst, T., and Wiley, D. N. (**2011**). "Common humpback whale (Megaptera novaeangliae) sound types for passive acoustic monitoring," The Journal of the Acoustical Society of America **129**, 476-482.
Strager, H. (**1995**). "Pod-specific call repertoires and compound calls of killer whales, Orcinus orca Linnaeus, 1758, in the waters of northern Norway," Canadian Journal of Zoology **73**, 1037-1047.
Suthers, R. A. (**1990**). "Contributions to birdsong from the left and right sides of the intact syrinx," Nature **347**, 473-477.
Suzuki, P., Buck, J. R., and Tyack, P. L. (**2006**). "Information entropy of humpback whale songs," The Journal of the Acoustical Society of America **119**, 1849-1866.
Tervo, O. M., Christoffersen, M. F., Parks, S. E., Kristensen, R. M., and Madsen, P. T. (**2011**). "Evidence for simultaneous sound production in the bowhead whale (Balaena mysticetus)," The Journal of the Acoustical Society of America **130**, 2257-2262.

Thompson, P. O., Cummings, W. C., and Ha, S. J. (**1986**). "Sounds, source levels, and associated behavior of humpback whales, Southeast Alaska," The Journal of the Acoustical Society of America **80**, 735-740.
Thompson, P. O., Cummings, W. C., and Kennison, S. J. (**1977**). "Sound production of humpback whales, *Megaptera novaeangliae*, in Alaskan waters," The Journal of the Acoustical Society of America **62**, S89.
Titze, I. R. (**2008**). "Nonlinear source--filter coupling in phonation: Theory," The Journal of the Acoustical Society of America **123**, 2733-2749.
Tyack, P. (**1981**). "Interactions between singing Hawaiian humpback whales and conspecifics nearby," Behavioral Ecology and Sociobiology **8**, 105-116.
Tyack, P. L., and Miller, E. H. (**2002**). "Vocal anatomy, acoustics communcation and echolocation," in *Marine Mammal Biology: an evolutionary approach* (Blackwell Publishing, Oxford UK), pp. 142-175.
Tyson, R. B., Nowacek, D. P., and Miller, P. J. O. (**2007**). "Nonlinear phenomena in the vocalizations of North Atlantic right whales (Eubalena glacialis) and killer whales (Orcinus orca)," The Journal of the Acoustical Society of America **122**, 1365-1373.
Viterbi, A. J. (**1967**). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Trans on Inf Th **13**, 260-269.
Wada, H. (**2010**). "The Development of Birdsong," Nature Education Knowledge **2**, 16.
Wartzok, D., and Ketten, D. R. (**1999**). "Marine Mammal Sensory Systems," in *Biology of Marine Mammals* (Smith-sonians Institution Press, Washington DC), pp. 117-175.
Weinrich, M., and Corbelli, C. (**2009**). "Does whale watching in Southern New England impact humpback whale (Megaptera novaeangliae) calf production or calf survival? Biological Conservation " Biological Conservation **142**, 2931-2940
Weisburn, B. A., Mitchell, S. G., Clark, C. W., and Parks, T. W. (**1993**). "Isolating biological acoustic transient signals," in *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing* (Minneapolis, MN).
Wilden, I., Herzel, H., Peters, G., and Tembrock, G. (**1998**). "Subharmonics, biphonation, and deterministic chaos in mammal vocalization," Bioacoustics, Vol 9, 171-196.
Williams, H., and Staples, K. (**1992**). "Syllable chunking in zebra finch (Taeniopygia guttata) song," Journal of Comparative Psychology **106**, 278-286.
Winn, H. E., Thompson, T. J., Cummings, W. C., Hain, J., Hudnall, J., Hays, H., and Steiner, W. W. (**1981**). "Song of the humpback whale — Population comparisons," Behavioral Ecology and Sociobiology **8**, 41-46.
Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (**2000**). "The HTK book," (Microsoft Corporation).

Yurk, H., Barrett-Lennard, L., Ford, J. K. B., and Matkin, C. O. (**2002**). "Cultural transmission within maternal lineages: vocal clans in resident killer whales in southern Alaska," Animal behaviour **63**, 1103-1119.
Zimmer, W. M. X. (**2011**). *Passive acoustic monitoring of ceataceans* (Cambridge University Press).