Department of Physics and Astronomy

Centre for Marine Science and Technology

Automatic Detectors for Underwater Soundscape Measurements

Shyam Kumar Madhusudhana

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

December 2015

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: ……

Date: ……

Abstract

Environmental impact regulations require that marine industrial operators quantify their contribution to the underwater noise scene. Automation of such assessments becomes feasible with the successful categorisation of underwater sounds into broader classes based on the type of source – biological (e.g. echolocation clicks, whistles and moans), anthropogenic (e.g. vessel noise, offshore construction and airgun surveys) and natural (e.g. underwater quakes and eruptions, rain and ice-cracking). Such an automatic characterisation system would enable regulators and operators to readily determine "noise budgets" (the contribution to underwater acoustic energy by source type) on a large spatiotemporal scale. Though industry and national governments alike are being increasingly concerned with mitigating human impact on marine life, little progress has been made in developing a unified approach for the study of underwater soundscapes.

Previous approaches to detection and classification in underwater passive acoustic monitoring (PAM) have mostly been limited to one or a few specific sources of interest such as dolphin whistles or signature noises of vessels. A fundamental problem in employing such techniques for soundscape characterisation lies in the varied notion about noise. A majority of these approaches are designed to discard or overcome interfering sounds that are considered "noise" for the particular application. Successful soundscape characterisation cannot afford to overlook any class of sounds. Also, some of the existing techniques exhibit claimed levels of performance only in specific recording scenarios. A recording scenario involves a combination of factors including recording environment, time of year, recording equipment configurations, etc. On the other hand, assembling a soundscape characterisation system as a combination of carefully handpicked independent PAM techniques is prohibitive given the variety of sounds prevalent in aquatic environments and the vast repertoire of available PAM techniques.

I propose to tackle the automatic categorisation problem by embracing a different perspective of underwater sounds – as classes of spectro-temporally distinguishable

units, without regard to the source producing the sounds. Such a perspective enables viewing of an automatic characterisation system as a broader 2-phase detection-and-classification system where the first phase involves only the detection of individual sound units and the second phase would perform classification upon examining the characteristics of the detected units. In this thesis, robust independent signal detectors are proposed for the automation of the detection phase. The proposed detectors are developed as modular units with emphasis on achieving a high degree of spectro-temporal context-insensitivity while maintaining faster than real-time performance. A framework is suggested for realising an automated underwater soundscape characterisation system utilising the proposed detectors.

Acknowledgements

Firstly, I would like to express my sincerest gratitude to my supervisors A/Prof. Alexander Gavrilov and Dr. Christine Erbe for allowing me the space and freedom I needed to work and for the unwavering support and guidance offered from day one. For their meticulous supervision and the countless hours they have poured into making this study a wonderful learning experience, I am truly grateful. Additional thanks to Dr. Gavrilov for his patience and understanding and to Dr. Erbe for her inspirational encouragement towards keeping me motivated throughout the course of this study. The learning experience you two have provided has considerably broadened my outlook.

I would like to thank Chevron, who sponsored this work through their funding of research into Marine Noise under their participation in the Western Australian Energy Research Alliance (WA:ERA). I would like to thank Curtin University for offering me the International Postgraduate Research Scholarship. Thanks to the sponsors and organisers of the 6th and 7th international workshops on Detection, Classification, Localization and Density Estimation workshops for the registration and travel allowances provided towards attending and presenting my work at the events.

In the process of completing this thesis, I have received help from a few other people. Thanks to my former (Masters) supervisor Prof. Marie Roch for her valuable guidance through my candidacy and her assistance towards completing a journal article. Thanks to Dr. Alec Duncan for verifying the accuracy of certain aspects of the manuscript. Thanks to the anonymous reviewers of my article published in the Journal of the Acoustical Society of America for their critical reviews which helped improve the quality of my work.

I have always enjoyed the camaraderie of my fellow PhD candidates – Matthew Koessler, Marta Galindo-Romero, Jamie McWilliam and Sarah Marley; and some of the fellow researchers that I have shared many lighter moments with – Miles

Parsons, Klaus Lucke, Iain Parnum. Thanks for all the memories. I would like to extend my gratitude to all students and staff at the Centre for Marine Science and Technology, for their support throughout the PhD programme. In particular, thanks for all the help and comfort offered that enabled me cope better with my numerous injuries.

Thanks to my family for their encouragement and for being ever supportive in more ways than I can describe. Finally, I like to thank the friends I have acquired during my stay in Australia – Robert Main, Vishal Chaturvedi, Meagan Burke, Aparna Warrier, Mark Thomson, Rahul Singh, Elliot Cleaver-Wilkinson and Laura Vermeulen – for all the cheerful moments that helped me maintain a healthy work-life balance through the course of my study.

Table of Contents

List of Figures

List of Tables

Chapter 1.
Introduction

The soundscape of an environment is the superposition of acoustic signals from all the sources within it (Krause, 1987; Schafer, 1977). These acoustic sources can be of geophysical, anthropogenic or biological origin. Geophysical sources include underwater earthquakes, volcanic eruptions and polar ice events, as well as weather-related sources, such as surface waves, wind and rainfall. Anthropogenic sources include, for example, ship noise, seismic airgun surveys and pile driving. Biological sources include marine fauna vocalisations and fish choruses. The acoustic characteristics of sounds change as they propagate through the environment due to spreading losses of acoustic energy, reflections from boundaries, surface and volume scatters. The superposition of sounds produced by various sources at different locations results in soundscapes that are both spatially and temporally heterogeneous. Soundscapes can be characterised by classifying and quantifying the contributions from the various sources.

Wenz (1962) summarised data on ambient sound sources and their spectra, in particular those of geophysical and weather-related origins. The sounds arising out of sea surface agitation caused by rainfall, wind and ice mechanics were further reviewed by Kerman (1988). Examples of geophysical contributions include underwater seismic and volcanic activities and polar ice events (calving, cracking, etc.).

Anthropogenic contributions to marine soundscapes have significantly increased since the onset of industrialisation, and in some regions seem to be steadily increasing (Andrew *et al.*, 2002; Andrew *et al.*, 2011; Chapman and Price, 2011). Some of the most common sources of anthropogenic noise in the ocean include commercial and recreational vessel traffic, seismic exploration of the sea floor, marine construction, as well as military, commercial and fisheries sonar (Richardson and Thomson, 1995; Wyatt, 2008; Hildebrand, 2009).

Cetaceans (whales, dolphins and porpoises), pinnipeds (seals, sea lions and walruses) and sirenia (dugongs and manatees) are all known to produce sounds underwater (Richardson and Thomson, 1995) not only as a means of communication, but also to aid in foraging and navigation as in the case of cetacean biosonar. Other sounds of biological origin in underwater soundscapes include those of fish (Tavolga, 1971) and invertebrates (Popper *et al*., 2001; Hazlett and Winn, 1962; Schmitz, 2002). Some of the other known reasons that marine organisms produce sounds include establishing territories, locating conspecifics, attracting mates and warding off predators (Tyack and Clark, 2000; Madsen *et al*., 2005; Valinski and Rigley, 1981).

Apart from actively producing sounds, marine animals also passively utilise sounds for various purposes, e.g. for habitat selection as in the case of pinnipeds (Miksis-Olds and Madden, 2014), fish (Simpson *et al*., 2008; Radford *et al*., 2011) and coral larvae (Vermeij *et al*., 2010). As soundscapes vary spatially across marine environments resulting from their local geophysical conditions and inhabiting biota, some species utilise sounds as environmental cues (Clark and Ellison, 2004) (e.g. avoidance of predatory regions) and navigational cues (Tolimieri *et al*., 2000; Leis *et al*., 2003; Radford *et al*., 2011).

Sound propagates faster and farther in marine environments than in the atmosphere and over longer ranges with less attenuation than does light (Urick, 1983). While marine organisms have evolved to use this to their advantage (Tyack and Clark, 2000), the very same beneficial property of water as an excellent medium for sound transmission turns detrimental as man-made acoustic disturbances can reach far distances with low attenuations. As a result, added anthropogenic sounds may have several impacts on marine ecosystems (Erbe, 2012), including behavioural responses (Southall *et al*., 2007), masking of communication and echolocation signals (Clark *et al*., 2009), an increase in organisms' stress levels (Wright *et al*., 2007) and temporary or permanent damage to sensory organs (McCauley *et al*., 2003; Solé *et al*., 2013; Kastelein *et al*., 2013). Over the past few decades, growing awareness of the impacts (Laiolo, 2010) have prompted various organisations and national governments to consider acoustic impact mitigation programmes. In consideration of the tight coupling between an environment's ecology and soundscape (Farina and Piretti,

2012), environmental impact assessment programmes have also recently started to broaden the scope of their studies to incorporate soundscape assessments.

Soundscape assessments as well as noise impact mitigation and monitoring programmes often involve Passive Acoustic Monitoring (PAM). It is commonly used as a complement to visual monitoring, and in some situations replaces visual methods (e.g. in conditions of poor visibility or at night time; Erbe *et al*., 2013). Cabled or autonomous underwater acoustic recording equipment can be deployed for extended periods for the collection of long-term data. Recordings are also commonly performed onsite from research vessels using on-board equipment. The collected acoustic data are analysed, either in-situ or after retrieving the recording equipment, to extract 'signals of interest' (e.g., dolphin whistles, odontocete echolocation clicks, seismic survey pulses). Some of the aims of past analyses have been to –

- detect the presence/absence of species of interest (e.g. Klimley *et al*., 1998),
- map species' migratory patterns (e.g. Clark *et al*., 1996),
- estimate species abundance (e.g. Lewis *et al*., 2007; Marques *et al*., 2009),
- monitor anthropogenic noise from industrial operations (e.g. Bailey *et al*., 2010), and
- determine animal responses to noise (e.g. Tyack *et al*., 2011).

The use of software programs for the automatic identification of signals of interest in audio recordings has made PAM more time- and cost-effective. Some of the early efforts in automatic analysis of underwater recordings include the works of Stafford *et al*. (1994), Gillespie and Leaper (1996), Mellinger and Clark (1997), Mann and Lobel (1995a, 1995b) and Chen *et al*. (2000). Since then, automatic PAM approaches have become an increasingly common tool in the monitoring of underwater fauna and have been widely and successfully employed over the last two decades.

The major goal of this study is to develop software tools to help automate the process of underwater soundscape characterisation. The significance of improved tools for soundscape studies is described in Section 1.1. The challenges in their development are discussed in Section 1.2. A brief review of existing automatic detection and recognition techniques is given in Section 1.3. An overview of the proposed approach and its motivations are described in Section 1.4.

1.1. Significance

Environmental impact assessments of marine industrial operations often require that the marine soundscape be measured and quantified prior to industrial operations, in order to understand the baseline conditions before anthropogenic noise is added. Such "before" conditions are assumed to represent the "natural" soundscape. Data analysis typically involves the identification of marine fauna present in the local environment, as well as a statistical analysis of "natural" noise levels and spectra. A baseline soundscape assessment is performed for a certain period either directly preceding operations, or a year earlier during the same season, or over much longer periods to capture any seasonal changes in ambient noise levels and fauna presence. The soundscape is also commonly recorded during industrial operations in order to monitor fauna around the operation sites and to monitor anthropogenic noise emission. Finally, soundscapes may be studied after industrial operations have ceased to determine whether there were any long-term changes as a result of the industrial operations.

Given the efficiency of PAM (unmanned data collection, low cost) and the ongoing technological advances (lower power consumption, higher storage capacity, higher sampling rates, improved recording quality), vast amounts of underwater acoustic data are collected by PAM around the globe every year. An analysis of such high volumes of data is often laborious and the scientific community at large has been looking beyond manual methods. As such, there is an ongoing need to improve automatic analysis tools. Automated soundscape analyses enable quick and efficient processing of real-time data as well as archived data. In addition, automatic tools allow consistent, repeatable and objective analyses.

Given the potential of an automatic system to process large amounts of data quickly and efficiently, archived data can be reused in the analysis of historical soundscapes. Consequently, automatic soundscape analysis may provide a new perspective and methodology for assessing marine ecosystems by enabling researchers to study the long-term dynamics of ecosystems as they respond to climatic and human-induced interferences.

1.2. Challenges and expectations

Our understanding of the origins of the myriad of sounds observed in underwater environments is still very limited. This is because simultaneous visual observation or other means of sampling are necessary to associate sounds with their sources. Given that acoustic signals are capable of propagating over long distances in aquatic media, the sources might be far away and inaccessible for sampling non-acoustically. As a result, the set of underwater sounds currently known to us is only a small fraction of the sounds heard underwater. Furthermore, even in cases where the calls of certain marine species have been described, the automatic detection of species-specific sounds is made difficult by various factors such as variations in the calling behaviour of vocalizing species (Au, 1993, p. 121) and variations in the spectrotemporal characteristics of received sound due to effects of hydro- and geo-acoustic environmental conditions on sound propagation (Dashen *et al*., 2010).

The evolution of automatic PAM solutions has made available numerous independent *ad hoc* approaches to the detection and/or classification of underwater sounds, each targeted at a subset of acoustic signals. These approaches have designs that are rooted in the nature of not only the targeted sounds, but also of the ambient noise at the recording site. A fundamental problem in employing such approaches for soundscape characterisation lies in the varied notion about noise. A majority of these approaches are designed to discard or overcome interfering sounds that are considered "noise" for a particular application. Comprehensive soundscape characterisation cannot afford to overlook any class of sounds. Integrating all of the available disparate approaches to recognise the variety of sounds in the ocean into a single software solution would be prohibitive, from both implementation and operational perspectives.

Another major hurdle in realising a universal solution from existing approaches is that most of the existing techniques were optimised to excel in specific situations that are a combination of the target signals of interest and the ambient noise at the specific time and location. For example, some tools are well suited to work with recordings from a particular sonic environment that is a factor of the underlying

bathymetry and the different habitats in the area; some are designed to work only with audio that was recorded at a particular sampling frequency; some techniques produce reasonable results only when applied to non-noisy recordings. Applying software tools that were developed for a specific scenario (i.e., target signals, background noise, sound propagation conditions and recording equipment) to a new environment is difficult. Certain techniques are based on the use of a previously trained neural-network or a statistical model, and training a new model for use in a different scenario may not be feasible and sometimes may not produce results of high quality. The challenge is to develop algorithms that deliver desirable characteristics such as:

- Robustness – must exhibit little or no degradation in performance in dynamic noise environments;
- Flexibility – must be capable of working with data obtained via different collection methods, with little or no performance variation;
- Adaptability – must exhibit an acceptable level of consistency in performance across recordings from different data collection sites.

While achieving a system with the aforementioned characteristics, it is vital for the research and development phases to maintain focus on achieving high operational throughput, i.e. high efficiency.

1.3. A brief review of available automatic recognition methods

A variety of *ad hoc* methods have been developed for the identification of one or more specific types of underwater sounds. As the literature often uses the terms "detection", "classification", "identification" and "recognition" interchangeably, these terms will be redefined here for convenience in distinguishing different facets of processing in this study. Generally, the goal of an automatic approach is to "recognise" or "identify" sounds that are of interest to a particular application. Automatic recognition of sounds of interest involves separating them from other interfering sounds and background ambient noise. Separation of sound units, which have well-defined temporal and/or spectral bounds, from background noise will be referred to as "detection". The detection process does not identify the sources that

produce the sounds. Often, multiple sounds are of interest to certain PAM applications and the goals of such applications may include "classification" which separates sounds into classes or types of sources. The distinction between detection and classification algorithms in existing literature is rather fuzzy. In some approaches, simple operations achieve recognition without explicitly distinguishing detection and classification steps, e.g. those described in Mellinger and Clark, 2000 and Woodman *et al*., 2004. Some approaches achieve recognition with explicit detection and classification steps, e.g. Soldevilla *et al*., 2008; Madhusudhana *et al*., 2009; Thode *et al*., 2012.

Some of the earlier attempts in the PAM of underwater sound sources involved the use of simple matched filtering methods for the detection of blue whale (*Balaenoptera musculus*) calls (Stafford *et al*., 1998; Stafford *et al*., 1994) and sperm whale (*Physeter macrocephalus*) clicks (Gillespie and Leaper, 1996), a semi-automatic approach using an automatic pulse detector for identifying fish sounds (Mann and Lobel, 1995a; Mann and Lobel, 1995b) and an amplitude peak-counting algorithm for isolating ice-cracking sounds in Arctic underwater ambient noise (Zakarauskas, 1993). Spectrogram correlation has been effectively and widely used for many mysticete calls (Mellinger and Clark, 1997) including bowhead whales (*Balaena mysticetus*) (Mellinger and Clark, 2000), right whales (*Eubalaena*) (Munger *et al*., 2005; Urazghildiiev *et al*., 2009) and sei whales (*Balaenoptera borealis*) (Baumgartner *et al*., 2008). Methods based on sine wave modelling and Bayesian inference (Halkias and Ellis, 2006) and those built on tracking *time* × *frequency* peaks in a spectrogram (Roch *et al*., 2011b) have been used for odontocete whistles. Certain other types of *time* × *frequency* tracking methods have been used in the detection of mysticete tonal calls (e.g. Mellinger *et al*., 2011). Madhusudhana *et al*. (2009) combined frequency contour tracking with a rule-based expert system for the classification of blue whale calls. Energy- and frequency-based methods have been examined for odontocete echolocation clicks (Houser *et al*., 1999), and have been recently improved by Klinck and Mellinger (2011) in their Energy Ratio Mapping Algorithm (ERMA). Erbe and King (2008) utilised Shannon entropy measures in the frequency domain for the detection of marine mammal vocalisations in recordings. The Teager-Kaiser energy operator (Kaiser, 1990a) aids in the

detection of short-duration high-energy impulses in recordings, and is popularised by its widespread use in the detection of echolocation clicks, e.g. Kandia and Stylianou, 2006; Soldevilla *et al*., 2008; Roch *et al*., 2008; Gervaise *et al*., 2010. It is most dominantly used as a first-pass operation to pick out sound waveform sections that contain the signals of interest. Other techniques are also used as a first-pass or signal-conditioning operation for subsequent classification operations. Examples include the use of chirplet transform in blue whale call characterisation (Bahoura and Simard, 2008), Hilbert-Huang transform for killer whale (*Orincus orca*) calls (Adam, 2008) and wavelet transform for sperm whale calls (Adam *et al*., 2005). Image processing techniques such as edge detection and ridge detection have been employed for identifying spectrographic features of bioacoustic signals in spectrograms, e.g. Gillespie, 2004; Kershenbaum and Roch, 2013. Thode *et al*. (2012) employ thresholding and morphological processing of spectrograms in the detection phase of their approach for recognising bowhead whale sounds.

A relatively less studied area is the use of techniques successfully employed in human speech recognition, such as dynamic time warping (DTW) (Sakoe and Chiba, 1978), vector quantisation (VQ) (Gray, 1984), Gaussian mixture models (GMM) (Huang *et al*., 2001) and hidden Markov models (HMM) (Rabiner, 1989). The use of DTW was explored in Buck and Tyack (1993) for distinguishing the "signature whistles" of bottlenose dolphins (*Tursiops truncatus*). Weisburn *et al*. (1993) used HMMs for separating bowhead whale calls from white Gaussian noise. HMMs with cepstral features were used for the first time in the classification of different call types of a single marine mammal species, northern resident killer whales, in Brown and Smaragdis (2009). GMMs have been used in Roch *et al*. (2007) and Roch *et al*. (2011a) for the classification of echolocation clicks of a specific population of odontocetes, and in Mouy *et al*. (2008) for distinguishing bowhead whale calls from other biological sounds. With neural networks, almost all known implementations for classifying whale calls use raw spectrogram data as features (Thode *et al*., 2012) , e.g. Erbe (2000).

Automatic recognition of ship noise has in the past received much attention from defence organisations. Hence, there is relatively little apposite literature available in

the public domain. The dominant component of ship noise is propeller cavitation while the other components originate at the engines and other machinery (Urick, 1983). Chen *et al*. (2000) used a two-pass split-windows algorithm in the pre-processing stage and evaluated the relative performance of four types of neural network classifiers in the classification of ship sounds. Yang *et al*. (2002) compared the effectiveness of six different feature extraction methods, used with a Mahalanobis distance based classifier, in the classification of sounds from six ships. They have shown that a wavelet analysis and three different fractal-based methods perform better than traditional spectrum-based methods. Bao *et al*. (2010) proposed a novel approach based on a non-linear analysis of acoustical signals via empirical mode decomposition. Some researchers have narrowed the focus of their work on the identification of small vessels, e.g. Ogden *et al*., 2011; Sorensen *et al*., 2010. Relatively little work has been done on the automatic recognition of some of the other anthropogenic sounds such as those of underwater explosions, airgun surveys, pile driving, etc. In the method proposed by Woodman *et al*. (2004), sound pressure levels exceeding a pre-determined threshold value trigger detection of dynamite fishing events.

The automatic recognition of underwater sounds of geophysical origin (e.g. seismic, volcanic, ice cracking and breaking) presents a different problem altogether. Much of the existing research related to seismic sounds has been carried out by researchers in geophysics and related fields, with little or no emphasis on fully automatic recognition techniques (e.g. Fox *et al*., 2001). Much of the work was previously limited to terrestrial observations (e.g. Gledhill, 1985). Increasing interests towards fully understanding the earth's interior through measurements from global seismic networks have resulted in expanding the networks across ocean floors (Webb, 1998). Hanson *et al*. (2001) compared the ratios of short-term to long-term signal energy averages computed over a series of frequency bands against pre-set thresholds for the detection of high-intensity low-frequency signals received at the hydroacoustic stations of the International Monitoring System (IMS) of the Comprehensive nuclear-Test-Ban Treaty (CTBT). With a subsequent classification step, their work distinguished several types of anthropogenic and geophysical sounds. Sukhovich *et al*. (2011) used wavelet analysis with a similar "ratio of averages" approach for

detecting acoustic signals generated in water by teleseismic *P*-waves. With an added processing step in the form of Gradient Boosted Decision Trees (GBDT), a machine learning technique, Sukhovich *et al*. (2014) extended the method of Sukhovich *et al*. (2011) to automatically distinguish *T*- and *P*-waves. Polar ice cracking and breaking has been studied as an indicator of global climate change (e.g. Scambos *et al*., 2000) and hydroacoustic detection methods are well suited for remote monitoring of such events. Li (2010) described a method that utilises prior knowledge about propagation characteristics and a multivariate classification method for the detection of Antarctic ice events. Automatic recognition of other natural sounds generated from events such as rain and sea-surface waves are of importance for weather forecasting and oceanography as they enable remote detection and measurement of phenomena such as precipitation, sea-surface winds, sea state, etc. (Scrimger, 1985). The characteristics of sounds generated during such events have been widely studied over many decades (e.g. Heindsmann *et al*., 1955; Scrimger *et al*., 1987; Farmer and Vagle, 1988; Medwin *et al*., 1992; Prosperetti and Oguz, 1993). Consequently, passive acoustic detection and measurement of such events have also been quite popular (e.g. Shaw *et al*., 1978; McConnell, 1983; Bourassa, 1984; Nystuen *et al*., 2000; Manasseh *et al*., 2006).

1.4. Aim of the thesis

Given the innumerable variety of sounds occurring in underwater environments, it is not feasible to realise an automatic soundscape characterisation system by employing a limited set of the available recognition techniques. On the other hand, a characterisation system made up of many independent techniques would be computationally cumbersome and would therefore lose its utility in real-world applications. Another impediment in realising a comprehensive automatic characterisation system is the gap in our knowledge on the sources of a variety of observed sounds. However, by analysing an acoustic signal aurally and/or visually (waveform analysis, Fourier analysis, etc.), human analysts are generally able to say with some confidence whether a sound is of biological, anthropogenic or geophysical origin based on prior encounters of similar sounds and their recognisable features

and patterns. Acoustic signals that are separable from the background ambient noise are readily detected by analysts who are then able to associate a source with the detected signal. A schematic emulation of the two-step analysis process is shown in Figure 1.1.

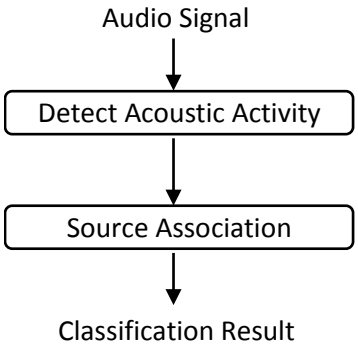

Figure 1.1. Schematic representation emulating the two-step process in the manual analysis of underwater soundscapes.

Acoustic signals can be considered to fall into a time and frequency continuum. In the frequency domain, signals occur either as *narrowband* (*tonal*) signals where the spectral energy is concentrated in very narrow frequency bands, or as *broadband* signals where the spectral energy is distributed over a relatively wider range of contiguous frequency bands. Examples of tonal signals include odontocete whistles, tonal components of ship noise and harmonic tremors emitted by icebergs. Examples of broadband signals include sounds of snapping shrimp, some fish choruses, earthquakes, underwater explosions and cavitation noise of ship propellers. In the time domain, signals occur in various durations from short pulsed signals (e.g. echolocation clicks and airgun discharges) to long continuous signals (e.g. baleen whale song, dredging and drilling noises). The spectro-temporal characteristics of acoustic signals discussed here are illustrated in Figure 1.2 using underwater recordings. Narrowband and broadband signals are shown in Figure 1.2(a) and Figure 1.2(b), respectively. They are both examples of long continuous signals. The contrasting nature (temporal) of short pulsed signals is shown in Figure 1.2(c).

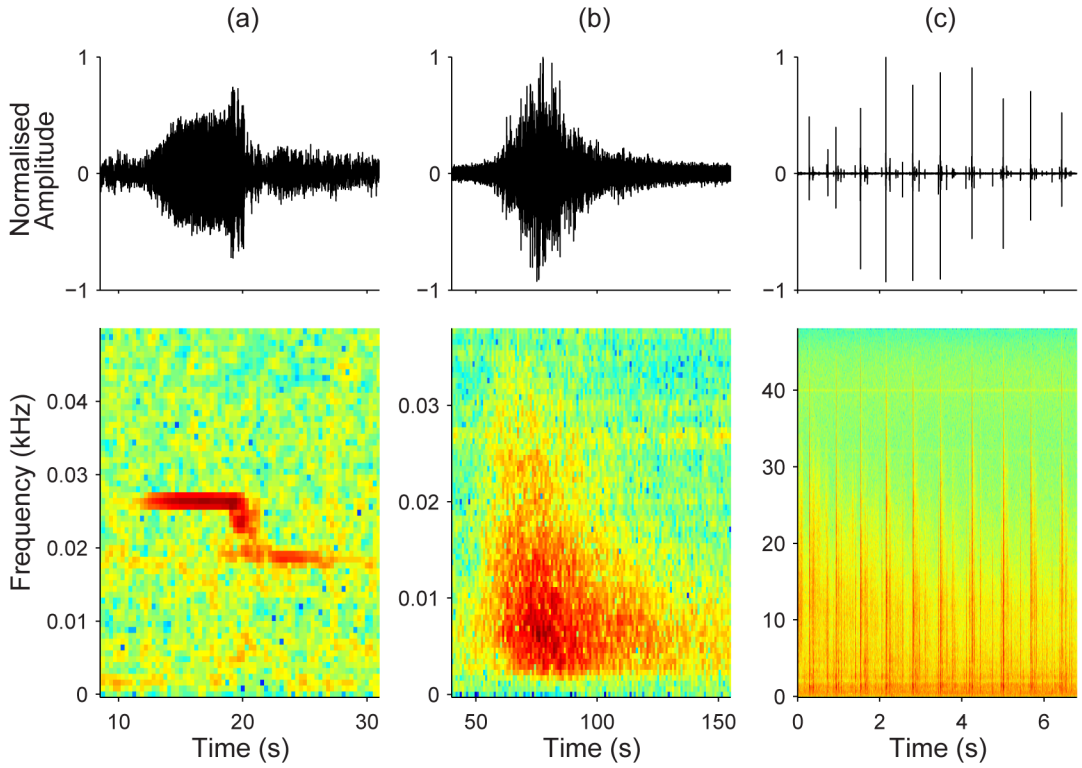

Figure 1.2. Waveforms (top row) and spectrograms (bottom row) of acoustic recordings containing (a) an instance of a Z-shaped call of an Antarctic blue whale (*Balaenoptera musculus intermedia*), (b) undersea earthquake and (c) sperm whale (*Physeter macrocephalus*) echolocation clicks.

The goal of this thesis has been to develop a system for the automatic detection of acoustic activity in underwater soundscapes. Three independent signal detectors were developed. Two of these cater to an extreme each in the time and frequency continuum. Very short and pulsed signals, henceforth referred to as *transients* for convenience, are handled by a transient signal detector. Note that the exact meaning of transients may differ in other literature where they are generally considered to be signals of limited duration. The durations considered for transients in the literature varies from tens of microseconds (e.g. Gillespie and Chappell, 2002; Gerard *et al*., 2009) and milliseconds (e.g. Baumgartner and Mussoline, 2011; Mellinger and Clark, 2000) to seconds (e.g. Helble *et al*., 2012) and minutes or longer (e.g. Li and Gavrilov, 2008; Hanson *et al*., 2001). Chapter 2 describes the functioning of the transient signal detector. A description of the transient signal detector along with detailed performance analysis was also presented in Madhusudhana *et al*., 2015. Tonal signals are handled by a tonal detector, which is described in Chapter 3.

Observed signals that occur between the two extremes generally appear as “blobs” in a *time* × *frequency* representation. This is particularly true in the case of signals from distant sources, because sound propagation affects the spectrum: acoustic energy at higher frequencies gets attenuated faster and multipath arrivals cause spreading of signals in time. Detection of such signals is handled by the broadband detector, which is described in Chapter 4.

Chapters 2, 3 and 4 start with a review of the relevant detection approaches available in the literature. These are followed by a description of the respective detectors and finally, analyses of the detectors’ performances are presented. Suggestions for integrating these detectors into a comprehensive, automatic underwater soundscape characterisation system are proposed in Chapter 5.

Chapter 2.
Transient Detection

Prior research has shown that echolocation clicks of several species of terrestrial and marine fauna can be modelled as Gabor-like functions. In this chapter, a system is proposed for the automatic detection of a variety of such signals. By means of mathematical formulation, it is shown that the output of the Teager-Kaiser Energy Operator (TKEO) applied to Gabor-like signals can be approximated by a Gaussian function. Based on those inferences, a detection algorithm involving the post-processing of the TKEO outputs is presented. The ratio of the outputs of two moving-average filters, a Gaussian and a rectangular filter, is shown to be an effective detection parameter. Detector performance is assessed using synthetic and real (taken from MobySound database) underwater acoustic recordings. The detection method is shown to work readily with a variety of echolocation clicks and in various recording scenarios. The system exhibits low computational complexity and operates several times faster than real-time. Performance comparisons are made to other publicly available detectors including PAMGuard.

2.1. Introduction

Odontocetes emit biosonar signals, commonly referred to as echolocation clicks, for navigation and foraging purposes. They emit short-duration (up to a few milliseconds) impulsive signals, called clicks, and interpret the received echoes to detect and identify objects and obstacles. Beaked whales, porpoises, sperm whales and some delphinids also produce clicks for communication purposes. The characteristics of echolocation clicks differ between different species by their duration, frequency content, inter-click intervals (ICI; temporal separation of successive clicks), inter-pulse intervals (IPI; temporal separation of successive pulses in a composite click), etc. Indicative echolocation clicks of a few species are shown in Figure 2.1. Notice the frequency modulation in the beaked whale click and the multi-pulsed nature of a sperm whale click.

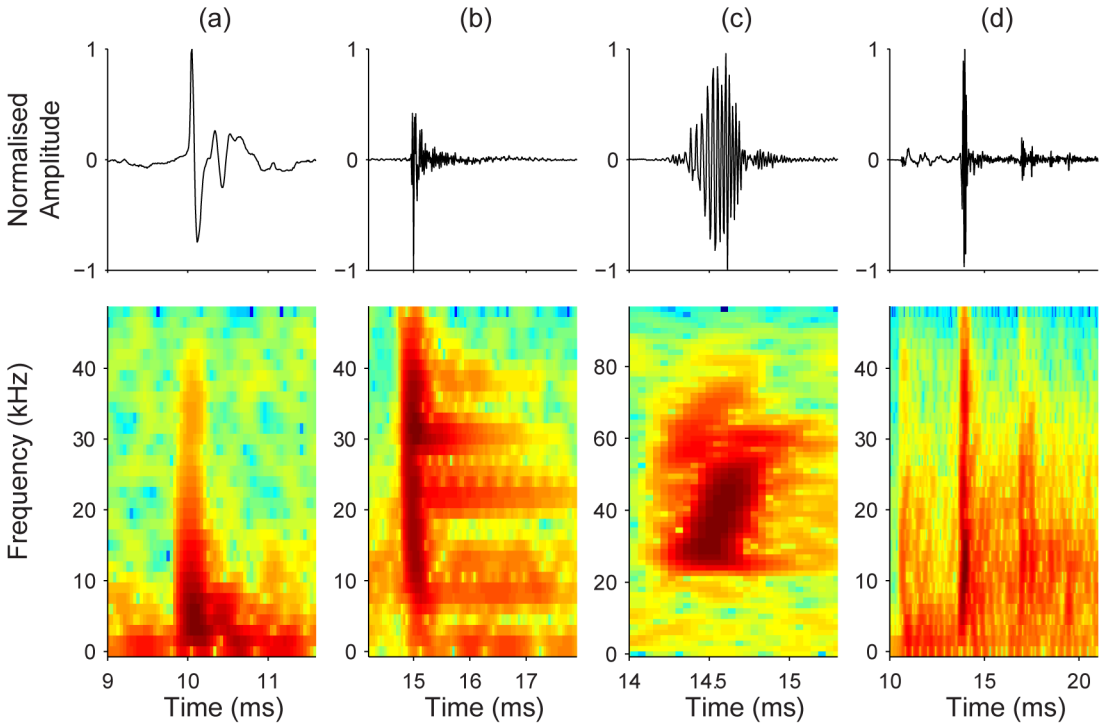

Figure 2.1. Waveforms (top row) and spectrograms (bottom row; FFT parameters: 500 μs Hanning window, 90% overlap) of indicative echolocation clicks of (a) rough toothed dolphin (*Steno bredanensis*), (b) Risso's dolphin (*Grampus griseus*), (c) Blainville's beaked whale (*Mesoplodon densirostris*) and (d) sperm whale (*Physeter macrocephalus*).

Clicks or impulsive signals from non-biological sources in underwater acoustic recordings include those from underwater sonar systems and impulsive noise artefacts arising from electronic glitches, mechanical impacts on the recording equipment, etc. It has been shown that echolocation clicks of several species of marine and terrestrial fauna can be approximated by Gabor-like functions (formulation presented in section 2.2). Examples include odontocetes (Kamminga and Beitsma, 1990; Kamminga *et al.*, 1996; Kamminga *et al.*, 1993; Kamminga and Stuart, 1995) and Egyptian fruit bats (Holland *et al.*, 2004). A Gabor function (Gabor, 1946) is a harmonic function localised by a Gaussian envelope. Several other studies, albeit without using the term "Gabor function" explicitly, acknowledge the presence of a Gaussian-like amplitude envelope resulting in small time-bandwidth products in the biosonar signals. Some of the species covered by these studies include Blainville's beaked whale (*Mesoplodon densirostris*) (Johnson *et al.*, 2006), finless porpoise (*Neophocaena phocaenoides*) (Goold and Jefferson, 2002), Hector's

dolphin (*Cephalorhynchus hectori*) (Thorpe and Dawson, 1991) and Mediterranean bottlenose dolphins (*Tursiops truncatus*) (Greco and Gini, 2006). A Gabor wavelet transform (Gabor, 1946) or a Gabor filter (Marčelja, 1980) applied to an acoustic time series could thus help to highlight the underlying clicks. In another study, van der Schaar *et al*. (2007) attempted identification of individual sperm whales based on modelling their clicks by Gabor functions. It will be shown how the application of the Teager-Kaiser Energy Operator (TKEO) (Kaiser, 1990a) to such signals simplifies and enhances their detectability with automatic detectors.

The TKEO has been used by several bioacousticians for automatic detection of underwater echolocation clicks (Kandia *et al*., 2006; Roch *et al*., 2008; Soldevilla *et al*., 2008; Roch *et al*., 2011a; Klinck and Mellinger, 2011). Several non-TKEO based methods have also been proposed, such as those based on kurtosis (Gervaise *et al*., 2010), on phase slopes (Kandia *et al*., 2008), on spectrogram correlation (Harland, 2008; Dobbins, 2009) and thresholding (Morrissey *et al*., 2006), on stochastic matched filtering (Caudal *et al*., 2008), on amplitude envelope levels (DeRuiter *et al*., 2009), on using noise-variable adaptive thresholds in an energy-based detector (Moretti *et al*., 2006; McCarthy *et al*., 2011) and on the use of support vector machines (Jarvis *et al*., 2008). Most of the existing click-detection algorithms based on the TKEO either use a simple moving-average filter comparing the outputs to a fixed threshold, rely on a noise floor that is pre-computed over a large time interval or perform some form of forward-backward peak selection operation within large audio segments (Kandia *et al*., 2006; Roch *et al*., 2008; Soldevilla *et al*., 2008; Roch *et al*., 2011a; Klinck and Mellinger, 2011). Some of the approaches that avoid the pitfalls of employing a fixed threshold perform multi-pass processing over large segments of recordings with an inherent assumption that spikes of echolocation clicks do not constitute a majority of the considered segment. The threshold is computed in an initial pass and then the spike locations corresponding to clicks in the segment are identified over one or more subsequent passes over the entire segment in consideration. The dependence of a detector on the assessment of certain signal statistics over long durations not only affects its response time, but also bears an impact on the consistency of its performance when employed in highly dynamic noise environments. Hence, such methods are not ideal for application in an online

scenario. They also run the risk of discarding weaker clicks in a temporal neighbourhood of multiple higher energy clicks. The method proposed by Kandia *et al*. (2006) is targeted at detecting sperm whale clicks and is based on measuring the deviation of the distribution of the TKEO output from a Gaussian shape. Analysis is performed iteratively on short successive frames. Barring the other elements meant for precisely locating the onset of a click, the algorithm would report detections when the deviation exceeds a pre-estimated skewness threshold. The method proposed by Roch *et al*. (2008) also performs operations frame-wise. The 40th percentile of the TKEO outputs in a frame is taken as the 'noise floor' and parts of the TKEO output that lie over 50 times this noise floor are considered to represent clicks. Similar approaches are employed in Roch *et al*. (2011a) and Soldevilla *et al*. (2008). Contrary to the usual practice of applying the TKEO directly to audio signals, Klinck and Mellinger (2011) apply the TKEO to the ratio of the outputs of two different band-pass filters and compare the result to a dynamic detection threshold. The threshold also relies on measurements from frames of 60 s duration.

In this chapter, I present an algorithm that employs two short moving-average filters to provide near-instantaneous spike detection in the TKEO output and that is well suited for processing continuous input audio samples.

The next section presents an analysis of applying the TKEO to Gabor signals. Then, the inferences made from the analysis are verified with a case study. The subsequent sections describe the detection algorithm and discuss its performance.

2.2. Applying the TKEO on a Gabor-like signal

2.2.1. Theoretical analysis

The TKEO output of an arbitrary continuous signal $x(t)$ is given by (Kaiser, 1990b)

$$\Psi_c[x(t)] = \dot{x}^2(t) - x(t)\ddot{x}(t), \tag{2.1a}$$

where the operators ˙ and ¨ denote the first and second derivatives, respectively. The TKEO output of an arbitrary discrete signal x_n is given by (Kaiser, 1990a)

$$\Psi_d[x_n] = x_n^2 - x_{n-1}x_{n+1}. \tag{2.1b}$$

For a Gabor function, there are several equivalent ways of mathematically expressing its Gaussian amplitude envelope (e.g., Kamminga *et al*., 1990; Holland *et al*., 2004). For ease of establishing a relationship with the width of a click, I chose the following representations for continuous and discrete Gabor signals:

$$G(t) = Ae^{\frac{-(t-t_0)^2}{2\sigma^2}} \cos\{\omega(t-t_0)+\phi\} \tag{2.2a}$$

$$G_n = Ae^{\frac{-(nT_s-t_0)^2}{2\sigma^2}} \cos\{\omega(nT_s-t_0)+\phi\}, \tag{2.2b}$$

where A is the signal amplitude, t_o and σ are the mid-epoch and standard deviation of the Gaussian envelope, respectively, and T_s is the sampling interval in the discrete case. The cosine term represents the carrier signal with phase ϕ and angular frequency $\omega = 2\pi/T_c$, where T_c is the period of the carrier wave.

Harmonic signals localised by a Gaussian envelope can be represented more generally as

$$G(t) = Ae^{\frac{-(t-t_0)^2}{2\sigma^2}} \cos\{\omega_t(t-t_0)+\phi\} \tag{2.3}$$

where ω_t describes the angular frequency as a function of time. Of particular interest to us are the cases with constant frequency carrier waves (CFCW) and those with linearly chirped carrier waves (LCCW), due to their similarity to commonly encountered echolocation clicks. An example of each case is shown in Figure 2.2.

The term 'Gabor-like' used in this chapter refers to these two types of signals. Signals of the latter form are commonly known as Gabor chirps (Mann and Haykin, 1991). The time dependence of their carrier frequency can be expressed as

$$\omega_t = \omega_0 + \dot{\omega}_t (t - t_0) . \tag{2.4}$$

Note that in this form, ω_o corresponds to the carrier wave's central frequency which is its instantaneous frequency at t_o. The carrier's instantaneous period corresponding to the central frequency will be denoted as T_o. For Gabor-like signals of CFCW type, $\dot{\omega}_t = 0$ in Eq. (2.4). The carrier wave's effective instantaneous frequency resulting from Eq. (2.4) must remain positive and finite within the full width of the Gaussian envelope, which can be defined as 6σ. This constrains the values of $\dot{\omega}_t$ to the range $0 \le |\dot{\omega}_t| < (\omega_0 / 3\sigma)$.

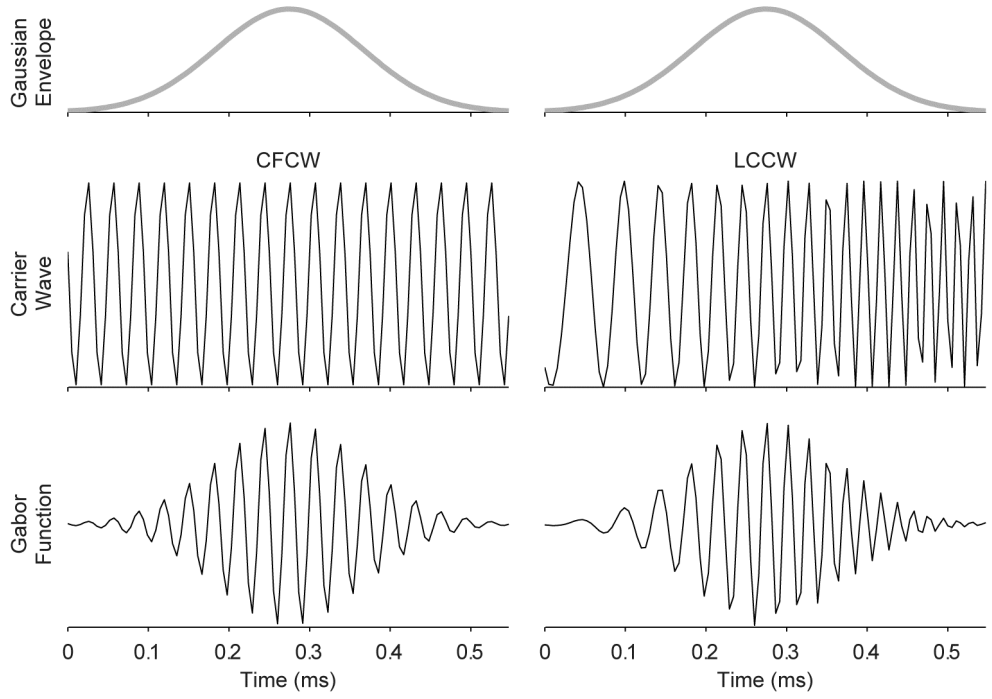

Figure 2.2. Gabor functions (bottom row) produced as per Eq. (2.3) with A = 1, σ = 0.091 ms and ϕ = 0, shown along with their constituent carrier waves (middle row) and Gaussian envelopes (top row). $\dot{\omega}_t$ is so chosen to yield a carrier frequency of 32 kHz for the CFCW type and a carrier frequency sweep from 24 kHz to 48 kHz (over the 6σ duration) for the LCCW type. The discrete signals were generated with a sampling frequency of 192 kHz.

Substituting $G(t)$ in Eq. (2.3) for $x(t)$ in Eq. (2.1) and simplifying the result using trigonometric identities, we arrive at the following form of the TKEO output for Gabor-like signals –

$$\Psi_c[G(t)] = A^2 e^{\frac{-(t-t_0)^2}{\sigma^2}} \left\{ [\omega_t + \dot{\omega}_t(t-t_0)]^2 + \frac{1}{2}[2\dot{\omega}_t + \ddot{\omega}_t(t-t_0)]\sin 2\theta + \frac{1}{\sigma^2}\cos^2\theta \right\}, \tag{2.5}$$
$$\theta = \omega_t(t-t_0) + \phi.$$

Ψ_c consists (in order of appearance) of a constant (A^2), a Gaussian component and a component comprising three additive terms that affect the shape of the Gaussian component. For convenience, I will refer to the three additive terms as T1, T2 and T3 in the order they appear in Eq. (2.5). By denoting the standard deviation of the Gaussian curve component in Ψ as σ_{TK}, we can express its relationship to the Gaussian envelope of $G(t)$ as

$$\sigma_{TK} = \frac{\sigma}{\sqrt{2}}. \tag{2.6}$$

Using Eq. (2.4), Eq. (2.5) can be rewritten for Gabor-like signals as

$$\Psi_c[G(t)] = A^2 e^{\frac{-(t-t_0)^2}{\sigma^2}} \left\{ [\omega_0 + 2\dot{\omega}_t(t-t_0)]^2 + \dot{\omega}_t \sin 2\theta + \frac{1}{\sigma^2}\cos^2\theta \right\}, \tag{2.7}$$
$$\theta = [\omega_0 + \dot{\omega}_t(t-t_0)](t-t_0) + \phi.$$

Let us consider separately the effect of T1, T2 and T3 on Ψ. The term T1 is a quadratic quantity and its minimum occurs at $-\omega_0/2\dot{\omega}_t$ relative to the Gaussian component's maximum. The magnitude of this temporal offset at its minimum is $3\sigma_{TK}/\sqrt{2}$ at the maximum $|\dot{\omega}_t| = \omega_o/3\sigma$ and it increases with decreasing $|\dot{\omega}_t|$. With its minimum occurring sufficiently away from t_o, the term T1 introduces a skew in the Gaussian component of Ψ. Notice that T1 is a constant ($\mathrm{T1} = \omega_o^2$) for Gabor-like

signals of CFCW type and, consequently, the Gaussian shape of Ψ is not skewed. The effects of T2 and T3 on Ψ can be examined by considering their values at the limits of $\dot{\omega}_t$. For the maximum value of $\dot{\omega}_t$, Eq. (2.7) can be rewritten as

$$\Psi_c\left[G(t)\right] = A^2\omega_o^2 e^{\frac{-(t-t_0)^2}{\sigma^2}}\left\{\left[1+\frac{2}{3\sigma}\left(t-t_0\right)\right]^2 + \frac{1}{\pi\lambda}\sin 2\theta + \frac{9}{\pi^2\lambda^2}\cos^2\theta\right\}, \tag{2.8}$$

where $\lambda = 6\sigma/T_o$ is the number of periods of the carrier wave's central frequency contained within the full width (6σ) of the Gaussian envelope of $G(t)$. The harmonic elements of T2 and T3 introduce distortions in an otherwise smooth curve of Ψ. The scaling of these distortions, viz. $1/\pi\lambda$ and $9/\pi^2\lambda^2$ (hereafter referred to as distortion scaling factors), are driven by λ. These terms are, however, small relative to unity when the Gabor-like signal is well-formed, i.e., contains at least a few periods of the carrier. Figure 2.3 shows the variation of the distortion scaling factors in T2 and T3 for a few values of λ at $\dot{\omega}_t = \omega_0/3\sigma$. Since T1 approaches unity at t_o in Eq. (2.8), the maximum cumulative distortion produced by T2 and T3 can be seen from Figure 2.3 as being small relative to T1 in the region around t_o for well-formed signals. For any

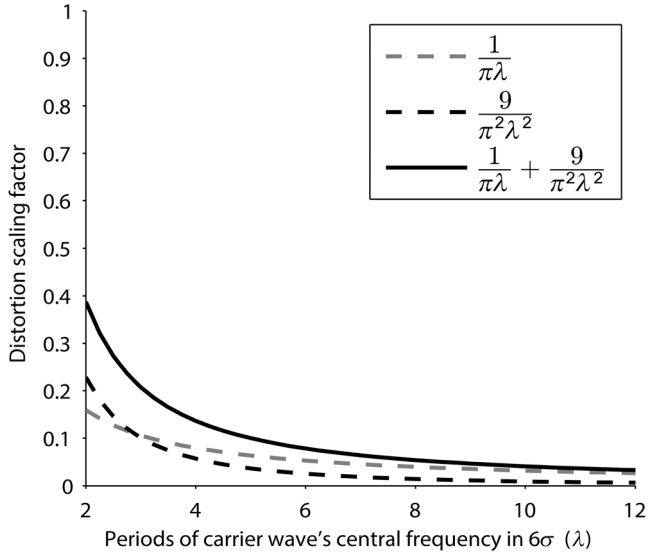

Figure 2.3. Scaling (dashed lines) of the distortion produced by the harmonic elements of T2 and T3 in Eq. (2.8), shown for a few values of λ. The solid line is indicative of the upper limit on the magnitude of distortion as a cumulative effect of T2 and T3.

particular value of λ, the maximum distortion of the Gaussian in Ψ occurs at maximum $\dot{\omega}_t$ and, as $\dot{\omega}_t$ approaches 0, the distortion results only from T3. So, it can be inferred in general that for well-formed Gabor-like signals, the magnitude of the distortions caused by T2 and T3 are small compared to the scaling and skewing caused by T1 over a significant extent of the Gaussian component of Ψ in the vicinity of t_o. Hence, the resulting nature of Ψ is largely dominated by a Gaussian. This is demonstrated in Figure 2.4 for a synthetic signal with a reasonably high rate of $\dot{\omega}_t$. Similarly high rates of frequency change in echolocation signals have been observed only in some subspecies of beaked whales (Zimmer *et al*., 2005; Rankin *et al*., 2011). Although the distortion of Ψ is visible at large $|\dot{\omega}_t|$, it is not significant compared to the non-skewed Gaussian output of the TKEO.

Thus far, I have shown that applying the TKEO to Gabor-like signals suppresses the harmonic component and that its output is well approximated by a scaled Gaussian impulse which is narrower than the amplitude envelope of the input signal by a factor of $1/\sqrt{2}$.

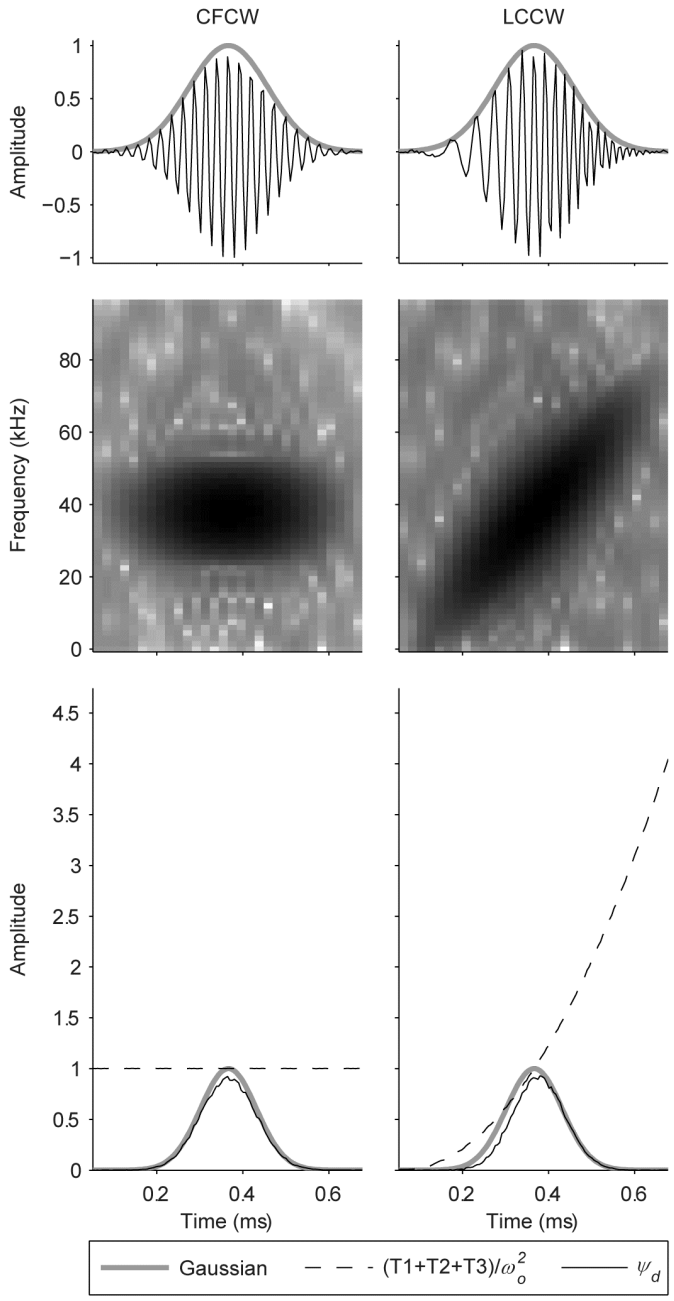

Figure 2.4. Waveforms (top row) and spectrograms (middle row) of synthetic Gabor functions produced with A = 1, σ = 0.091 ms and ϕ = 0. $\dot{\omega}_t$ is chosen as to yield a carrier frequency of 38 kHz in the CFCW case and a frequency sweep from 21 kHz to 55 kHz over the 6σ in the LCCW case. Grey overlays show the Gaussian envelopes. The bottom row plots show the corresponding Gaussian and quadratic-approximate (T1+T2+T3; scaled here, by $1/\omega_0^2$, to enable comparisons) components of the analytical TKEO output. Discrete TKEO output is overlaid over the pure Gaussian. The discrete TKEO output in the LCCW case indicates the introduced skew causing a forward shift of ~0.01 ms in its peak.

2.2.2. Case study

In order to verify the findings from the above analysis for real echolocation clicks, I performed a curve-fitting exercise on 200 handpicked killer whale clicks from a recording made during May 2011 over the Australian Northwest Shelf using an omnidirectional hydrophone (Teledyne Reson TC 4033, bandwidth 1 Hz–140 kHz, receiving sensitivity −202 dB re 1 V/µPa) and a CMST underwater sound recorder (http://cmst.curtin.edu.au/products/underwater-sound-recorder/; accessed on Sept 13, 2016). The recorded audio was sampled at 192 kHz. Gabor curves were fitted to the waveforms of each click, and Gaussians fitted to their corresponding TKEO outputs (see Figure 2.5). The Levenberg-Marquardt (LM) algorithm (Gill *et al*., 1981) is known to perform well in non-linear curve-fitting tasks and hence it was chosen for this analysis. The averages of the estimated parameters of the individual curve-fits were considered in producing the overlaid (dark) Gabor and Gaussian curves. The Gabor fitting of the waveforms yielded parameter estimates of σ = 0.0116 ms and T_o = 0.0324 ms resulting in $\lambda \approx 2.15$. A σ_{TK} estimate of 0.0079 ms supports the relationship expressed in Eq. (2.6). For the Gaussian fit of the TKEO outputs, an average *Summed Square of Errors/Residuals* value of 0.01 and a *Root Mean Squared Error* value of 0.03 confirmed the usefulness of the model for fitting purposes, and an average *Adjusted* R^2 value of 0.98 indicated a 'good fit'.

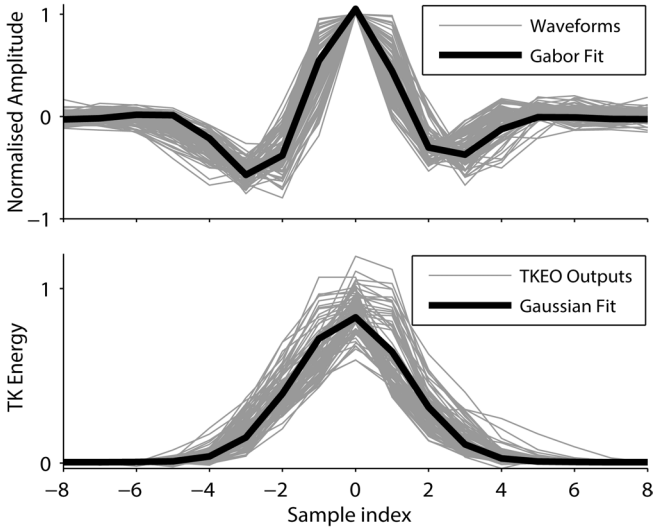

Figure 2.5. Curve fitting of killer whale click waveforms with a Gabor function (top) and of their corresponding TKEO outputs with a Gaussian curve (bottom). Grey lines show clicks' waveforms and their corresponding TKEO outputs in respective plots.

2.3. Automatic detection

So far, I have shown that, for signals that can be modelled as Gabor-like functions (e.g., underwater echolocation clicks), the corresponding TKEO values tend to approach a Gaussian shape. Based on these inferences, I will now describe a simple system for the detection of Gabor-like clicks in acoustic recordings.

2.3.1. Detector design

A short rectangular moving-average filter produces an averaging or smoothing effect on an input signal. Since the outputs of the TKEO are predominantly non-negative, a longer moving-average filter produces a flattening effect on the TKEO outputs. In contrast, a bell-shaped averaging filter (e.g., Hamming, Hanning or Gaussian function) has the potential of highlighting short-duration energy surges in TKEO outputs while flattening non-spiked high-energy sections. I chose a scaled Gaussian function for our first moving-average filter (MAF1) as it allows for easy control of the acuteness of the bell shape. Convolution operation with MAF1 can be expressed as

$$h_{MAF1}(n) = \frac{T_s}{\sigma_G\sqrt{2\pi}} \sum_{i=-N}^{N} e^{\frac{-(iT_s)^2}{2\sigma_G^2}} x_{n+i} \tag{2.9}$$

for a filter of length $2N + 1$, where n is the sample index and σ_G is the standard deviation of the Gaussian function. The factor $(T_s / \sigma_G\sqrt{2\pi})$ ensures that the filter gain (area under the curve) approaches unity. The acuteness of the Gaussian can be controlled with σ_G. The choice of values for σ_G and N is discussed in the next sub-section.

Consider a second moving-average filter (MAF2) – a rectangular averaging filter of the same length as MAF1. The amplitude of the filter is chosen such that the filter gains of MAF1 and MAF2 are the same. Similar gains allow for fair comparisons to be made of the two filters' outputs.

For an input unit impulse, $h_{MAF1}(n)$ peaks at the point corresponding to the non-zero element of the impulse and falls off on either side of it. In contrast, the response of MAF2 ($h_{MAF2}(n)$) is flat. The proposed detection algorithm exploits this difference in characteristics of the responses of the two filters. Consider the difference [$h_{MAF1}(n)$ - $h_{MAF2}(n)$] expressed as a fraction of $h_{MAF1}(n)$. This quantity will be denoted as Filter Difference Ratio (FDR) which is a normalised measure of the extent of $h_{MAF1}(n)$ over $h_{MAF2}(n)$.

$$FDR(n) = \frac{h_{MAF1}(n) - h_{MAF2}(n)}{h_{MAF1}(n)} \tag{2.10}$$

Impulse responses of typical filters and the ensuing FDR are shown in Figure 2.6. The dotted horizontal line in the FDR plot highlights the maximum value of FDR (FDR_{peak}). For a chosen combination of MAF1 and MAF2, there are four noteworthy properties of FDR –

i) The FDR curve and FDR_{peak} remain the same for input impulses of any given amplitude scaling.
ii) The difference [$h_{MAF1}(n)$ - $h_{MAF2}(n)$] and the ensuing FDR are maximum when the impulse is at the centre of the filters.
iii) The value of the numerator never exceeds the denominator. Hence, the resulting ratio is less than 1.
iv) $h_{MAF1}(n)$ is smaller than $h_{MAF2}(n)$ at input samples sufficiently away (in time) from the non-zero element of the impulse. The numerator and hence the ensuing FDR are negative for such points.

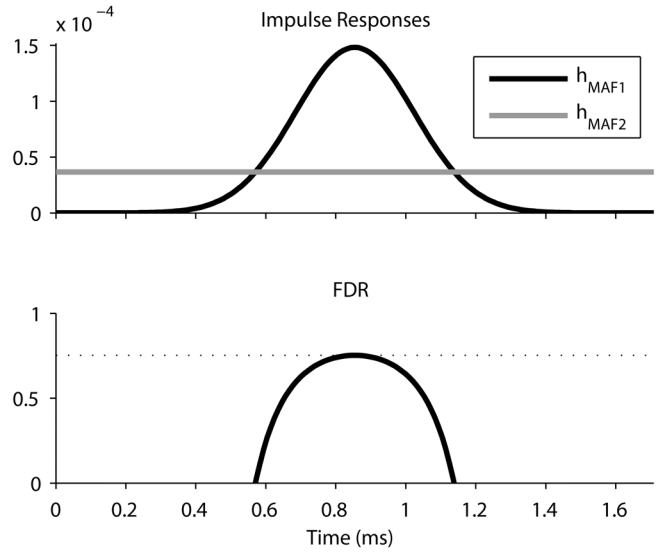

Figure 2.6. Impulse responses (top) of filters MAF1 (σ_G = 0.169 ms) and MAF2 and the corresponding FDR (bottom). FDR plot restricted to the range [0, 1]. Dotted line in the FDR plot indicates the peak FDR value.

Similar to a unit impulse, acute Gaussian curves also have a steep rise followed by a steep fall. We can see from Eq. (2.6) that the Gaussian-like outputs (hereafter referred to as spike) obtained from applying the TKEO to Gabor-like signals also have an acute profile. When the outputs of the TKEO applied to audio recordings containing Gabor-like signals are convolved with MAF1 and MAF2, and the FDR is determined, we can expect to see curves similar to those in Figure 2.6 at locations corresponding to clicks in the original audio. As with unit impulses of different amplitudes, the FDR curve would remain similar for clicks with different intensities. Hence, I chose to set the detector threshold to be a function of FDR_{peak} for the chosen combination of MAF1 and MAF2. However, TKEO outputs of real clicks differ from a unit impulse in two ways. Firstly, a combination of factors (like noise and choice of sampling rate) results in a possibility of bearing small negative values in the neighbourhood of the energy pinnacle of the TKEO output corresponding to a Gabor-like signal. Secondly, the width of the spike is wider than a unit impulse. As a result of these two factors, the tip of the FDR corresponding to a click would be lower than the FDR_{peak} computed for the chosen filters. Hence, the detection threshold can be set as a fraction of the employed filters' FDR_{peak}. Figure 2.7 demonstrates the outcome of filtering and FDR computation for synthetic data imitating TKEO outputs with different amplitudes. Notice how a fixed threshold, that is 85% of the FDR_{peak}, can serve as a reasonable cut-off for detecting spikes.

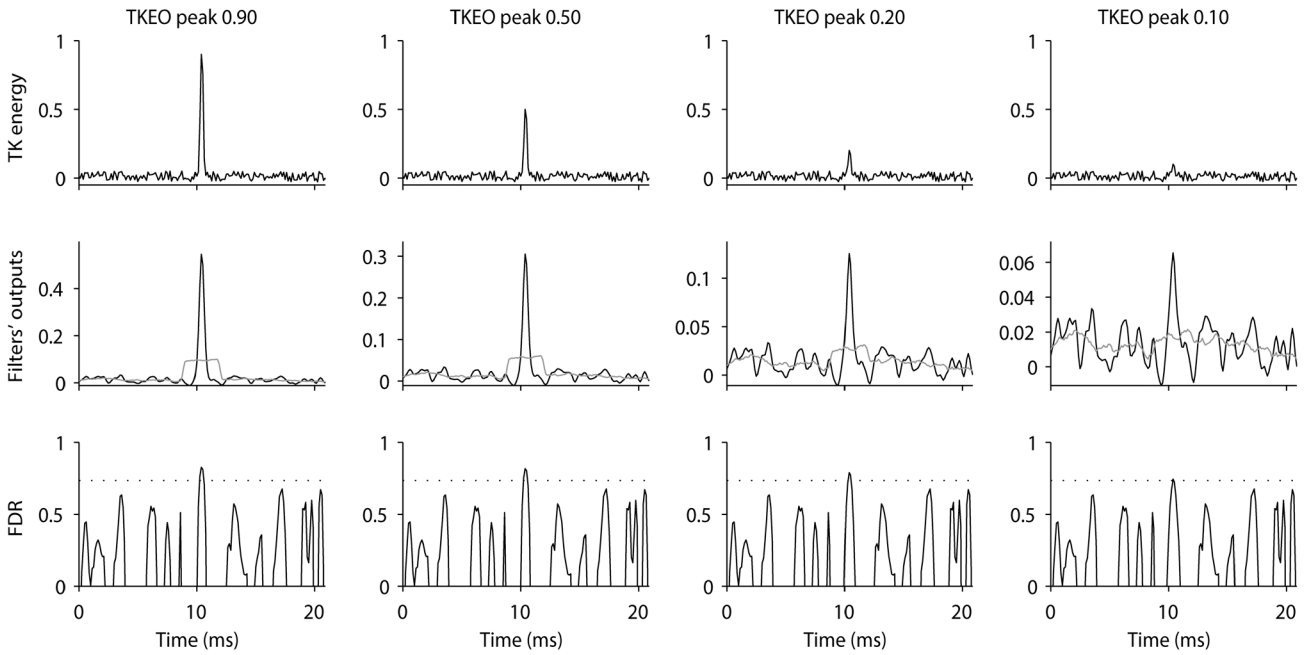

Figure 2.7. Demonstration of filtering and FDR computation for synthetic TKEO values with varying strengths for transient surges. First row shows the synthetic TKEO values with spikes ranging from 0.10 to 0.90. Second row shows the result of filtering the TKEO values with MAF1 (black curves) and MAF2 (grey curves). The third row shows the FDR (solid line) and the threshold (dashed line) set as 85% of FDR_{peak}.

Thus far, I have established that the output of MAF1 remains high for TKEO values corresponding to Gabor-like signals and in turn the FDR value produces a local maximum. However, the TKEO may produce non-positive outputs for sections of input audio that do not correspond to clicks. Depending on the length of MAF1 (and MAF2) and the negative strength of the TKEO output, this may sometimes translate to non-positive outputs from MAF1 and MAF2. This, in turn, would yield FDR values that are not meaningful for our application (e.g., $\pm\infty$). In certain implementations, FDR computation with such values may even cause undesirable exceptions (e.g., divide-by-zero exception). Since we know that a non-positive value in either filters' output does not indicate the presence of a spike in the TKEO output, we can safely bypass calculation of FDR for such values. Considering property (iv) of the FDR, we also bypass computation of FDR when $h_{MAF1}(n) \not> h_{MAF2}(n)$.

Considering property (iii) of FDR and the constraints described above ($h_{MAF1}(n) > 0$; $h_{MAF2}(n) > 0$ and $h_{MAF1}(n) > h_{MAF2}(n)$) for the computation of meaningful FDR values, we can see that the useable range of FDR values is effectively reduced to [0, 1]. Further, FDR values that are beyond the threshold value (fraction of FDR_{peak}) indicate the presence of Gaussian-like spikes in the TKEO outputs, in turn indicating the presence of Gabor-like signals in the input audio.

2.3.2. Implementation

The width of a Gaussian at half its peak value, commonly known as full width at half maximum ($FWHM = 2\sqrt{2\ln(2)}\sigma \approx 2.355\sigma$) provides a better feel for the width of the Gaussian pulse in visual observations. I will denote the FWHM and the standard deviation of the Gaussian envelope in the target click as $FWHM_{EC}$ and σ_{EC}, respectively. The standard deviation, σ_{TK}, of the Gaussian curve resulting from applying the TKEO to Gabor-like signals can be derived using Eq. (2.6) as

$$\sigma_{TK} = \frac{\sigma_{EC}}{\sqrt{2}} = \frac{FWHM_{EC}}{4\sqrt{\ln(2)}}. \tag{2.11}$$

The value of σ_{TK} obtained using estimates of $FWHM_{EC}$ made from visual observations of representative clicks' waveforms can be used as a guide in designing the needed filters. We can set the standard deviation of the Gaussian in MAF1 to be the same as σ_{TK} where it would function as a matched filter. We know that 99.7% of the area under a Gaussian curve is contained within a distance of 3σ on either side of its mean. Setting the length of the filter to $6\sigma_G$ would account for contributions only from the bulk of a spike without consideration for the points in its immediate neighbourhood. Extending the filter length would not only weigh the high energy regions, but also appropriately penalise low energy regions, thereby enabling only those sections to stand out that correspond to actual spikes in the TKEO output. However, a very long averaging filter stands the risk of clubbing close lying spikes. This causes smearing in the output thereby affecting their detectability with the FDR. Figure 2.8 demonstrates the effect N has on FDR and on the subsequent detection. Let us consider the faint pulse occurring at ~10.4 ms. As the energy of the pulse is not significant compared to background noise, a shorter MAF2 produces a larger output resulting in smaller FDR values as compared to the corresponding FDR_{peak}. For the same pulse, the FDR curves corresponding to different N show that larger N yields larger FDR. While increasing N is beneficial for pulses that are temporally well-separated from other high-energy signals, the resulting larger MAF2 increases the risk of accounting for energy from neighbouring signals (including other pulses) for pulses that are not temporally well-isolated. For the pulse occurring at ~8 ms, notice that its FDR is influenced by the preceding pulse for $N = \lceil 6\sigma_G/T_s \rceil$ and is influenced on both sides for $N = \lceil 7\sigma_G/T_s \rceil$. Based on such observations made from a few dozen instances, I have empirically arrived at a value of $N = \lceil 5\sigma_G/T_s \rceil$ for MAF1 (and in turn, for MAF2). Note here that all σ_G values are expressed in time units and may bear non-integer values and hence rounding N up to the next higher integer is necessary. Considering the widths of the different types of echolocation clicks commonly encountered, this value of N does not make the full filter length ($2N + 1$) unwieldy and at the same time enables fair weighting of points both on and in the neighbourhood of a spike. Once the values for σ_G and N are identified as described, MAF1 can be realised as

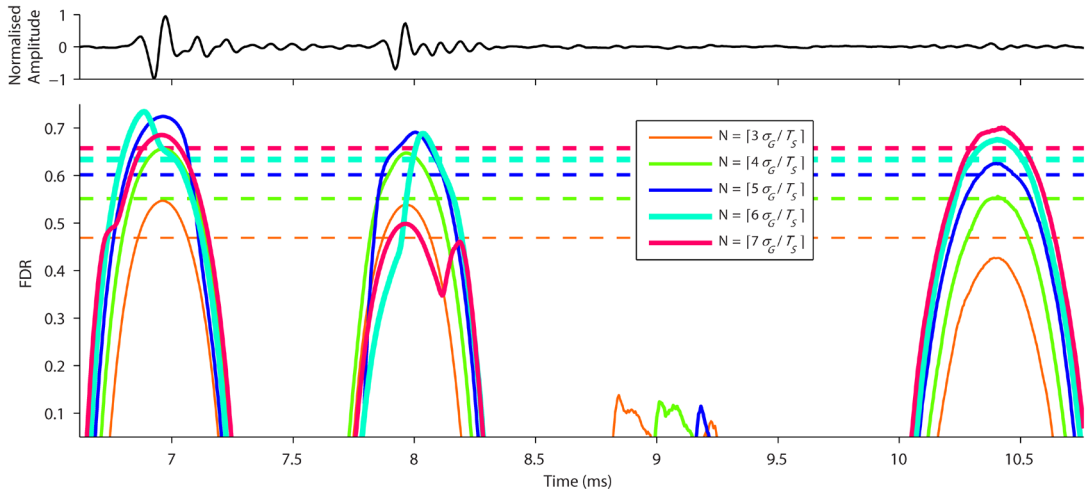

Figure 2.8. Demonstration of the effect of N on click detection using a segment of underwater acoustic recording (sampled at 192 kHz) containing sperm whale clicks. The top panel shows the waveform of the recording consisting of three distinct pulses. The bottom panel shows the corresponding FDR for different values of N. The range of y-axis values is restricted to enable clarity. A detection threshold of 80% of the resulting FDR_{peak} is also shown as dashed lines for each value of N.

$$MAF1(n) = \frac{T_s}{\sigma_G \sqrt{2\pi}} e^{\frac{-(nT_s)^2}{2\sigma_G^2}} \tag{2.12}$$

where n = - N, …, -3, -2 -1, 0, 1, 2, 3, …, N is the index of the sampled point in the filter. MAF2 can be realised as

$$MAF2(n) = \frac{\sum_{m=-N}^{N} MAF1(m)}{2N+1}. \tag{2.13}$$

The value of FDR_{peak} for the combination of MAF1 and MAF2 can be obtained by setting n = 0 in Eq. (2.12) and Eq. (2.13) and substituting the resulting values in Eq. (2.10). The product of the obtained FDR_{peak} and a user-controlled value (in the range 0-1) becomes the detection threshold for the system. A schematic of the proposed detection system is presented in Figure 2.9.

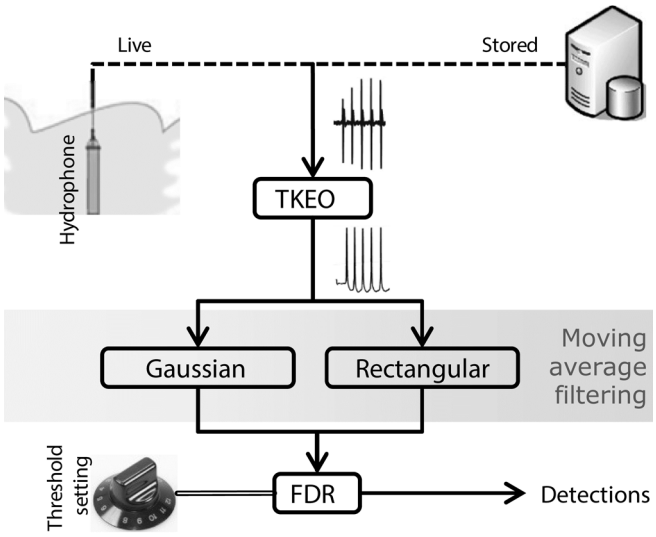

Figure 2.9. Schematic of the proposed click-detection system. Dashed lines are used to indicate that the input could either be pre-recorded audio or live real-time inputs.

2.4. Performance evaluation

The performance of the system was evaluated using both synthesised data and real audio recordings. For the latter, publicly available underwater audio recordings from MobySound archive (Heimlich *et al*., 2011) were used. The recording sets used are listed in Table 2.1. Synthetic data were generated using pieces of real underwater recordings. A 28s long audio fragment of ambient sea noise free of echolocation clicks was handpicked to serve as background noise. Two sets of twenty short audio clips containing single echolocation clicks were extracted from underwater sound recordings. Clips with sperm whale clicks, representing the CFCW type, constituted one set and clips with beaked whale clicks, representing the LCCW type, constituted the other. Two hundred instances of clicks were randomly drawn (with repetition) from one set and then superimposed at uniformly distributed random points in time across the ambient sea noise recording. The amplitude of each superimposed click was altered to yield a particular Signal-to-Noise Ratio (SNR) value. The SNR values chosen were uniformly distributed within the range from 5 dB to 30 dB. The SNR value was defined from the energy of the click being superimposed and the energy of background noise, both values determined within the frequency band of interest (3 – 30 kHz for sperm whales and 20 – 80 kHz for beaked whales) and integrated over the

Table 2.1. Datasets obtained from MobySound for testing the proposed detector. The last column shows the number of clicks that were in the annotations. Note that the number of clicks occurring in the recordings may be higher.

Species	Dataset identifier and audio file(s)	Sampling Rate (Hz)	Duration (s)	No. of clicks
Rough Toothed Dolphins	**RoughToothed_Marianas(MISTC)-Annotated** MISTCS070316-113000.wav	96000	412.7	57
Risso's Dolphins	**Rissos-SCORE-annot** Set1-A2-H17-081406-0000-0030-1225-1255loc.wav	96000	1800	172
Beaked Whales	**Mesoplodon_CanaryIsles-Annotated** md05_294a10590-11850.wav	192000	1260	1037
Sperm Whales	**Sperm whales_Bahamas(AUTEC)-Annotated** SpermWh_A2_030306-H16_short.wav	96000	870.2	2270
Spotted Dolphins	**SpottedDolphin_Bahamas(AUTEC)-Annotated** Set3_A4_042705_CH5_H40_A0600-0630.wav	96000	1800	408
Striped Dolphins	**StripedDolphin_Marianas(MISTC)-Annotated** MISTCS070309-092000.wav MISTCS070309-083000.wav	96000 96000	600 426.5	40 31

time interval containing 90% of the click energy. The noise fragment along with the superimposed clicks constitutes a synthetic test input. The start and end times of each superimposition were recorded for later comparison with detection results. Synthesis was repeated 1000 times for each species while generating different insertion points, different clip permutations and different SNR values at each repetition. In order to emulate the diversity in click characteristics prevalent in real underwater audio, a certain level of click dissimilarity was ensured within each clip set based on a "by eye" assessment.

The FWHM (and in turn σ_G) of MAF1 can be tuned as described in the previous section to achieve optimal performance in each of the aforementioned tests, i.e. for each species. However, I chose to use a single setting for all the tests in order to be able to show that the algorithm is capable of performing detection regardless of the species producing the clicks. The chosen value of FWHM = 0.40 ms translates to a filter length of 329 points for a sampling rate of 192 kHz, and 165 points for a sampling rate of 96 kHz.

For comparative performance analysis, tests with synthesised data were repeated with two other detectors - PAMGuard (http://www.pamguard.org/) and a TKEO-

based detector described in Roch *et al*. (2011a). PAMGuard is a publicly available software program that provides automatic detection/classification capabilities. The default "Click Detector" module was employed. It is a non-TKEO based detector which works by comparing signal levels to estimated background noise levels. The detector's various parameters were set as shown in Table 2.2. The latest version of PAMGuard available at the time of this work, viz. v1.13.02 BETA, was used. For testing the method of Roch *et al.* (2011a), a MATLAB based implementation was employed. The implementation used is available as a part of the *Silbido* (Roch *et al*., 2011b) package at http://roch.sdsu.edu/software/silbido_JASA2011baseline.zip (accessed on Dec 13, 2014). The detector's parameters were set as shown in Table 2.3. While some of the parameter values given in Table 2.2 and Table 2.3 were chosen based on *a priori* knowledge, others were arrived at following short trials using a small subset of the synthesised test data. While results better than those shown here may be possible for the compared methods, determining the optimal combination of parameter values is a non-trivial task and is beyond the scope of this study.

Table 2.2. Parameter settings used to configure the click detector module in PAMGuard for tests with synthesised data.

Parameter	**Sperm Whale**	**Beaked Whale**
Pre-Filter	High Pass: 200 Hz	High Pass: 10 kHz
Trigger Filter	Band Pass: 3-30 kHz	Band Pass: 20-80 kHz
Long Filter	0.00001	0.00001
Long Filter 2	0.000001	0.000001
Short Filter	0.1	0.1
Min. Click Separation	100 samples	100 samples
Max. Click Length	1024 samples	1024 samples
Pre Sample	40 samples	40 samples
Post Sample	0 samples	0 samples

Table 2.3. Parameter settings used to configure the click detector of Roch *et al*. (2011a).

Parameter	Sperm Whale	Beaked Whale
Ranges	3 – 30 kHz	20 – 80 kHz
MinClickSaturation	1.5 kHz	10 kHz
MaxClickSaturation	30 kHz	60 kHz
MeanAve_s	3 s	3 s
TransitionBand	0.2 – 3 kHz	3 – 20 kHz
FrameLength_s	0.01 s	0.01 s
ClickPad_s	0.0075 s	0.0075 s
MinClickSep_s	0.5 s	0.5 s
ClipThreshold	(disabled)	(disabled)

Tests with synthetic data were repeated for different sensitivity settings for all three methods. For the proposed detector, the threshold settings were varied from 0.4 to 1. In PAMGuard, the Trigger Threshold parameter of the click detector module was varied from 7 dB to 14 dB. The method described in Roch *et al*. (2011a) uses different thresholds in the two stages of the detection algorithm. The stage 1 threshold parameter was varied from 2 dB to 16 dB with the stage 2 threshold set at 5, 10, 25 and 50. Testing was repeated for the proposed detector, with pre-filtered inputs, where the synthesised data were bandpass filtered (using a Butterworth filter; passbands of 3 – 30 kHz for sperm whales and 20 – 80 kHz for beaked whales) before being fed to the detector.

With all three methods reporting detections as intervals (start and end times), a click present in input data (real or synthesised) is considered 'detected' if any of the following are true –

- The known/recorded interval of the click in the input audio completely envelops the intervals of any reported detections.
- A reported detection's interval completely envelops the known/recorded interval of the click.
- The temporal overlap with any reported detection is at least 60% of the known/recorded duration of the click.

In the case of synthesised data, a significant portion of each click occurs around the midpoint of the containing clip. Therefore, 60% overlap ensures that the click is appropriately accounted for by any partially overlapping detection. Reported detections that enable any of the above three conditions to be satisfied are considered to be 'true detections'. With these definitions of 'detected' clicks and 'true detections', performance metric 'recall' can be defined as the ratio of the number of 'detected' clicks to the number of clicks present in the test inputs, and the metric 'precision' can be defined as the ratio of the number of 'true detections' to the number of reported detections. Figure 2.10 shows the precision-recall (PR) trade-off characteristics for the three detectors. The various curves in the middle row plots show the PR characteristics for the different stage 2 threshold settings considered. Threshold settings that produced optimal PR trade-off values were identified from Figure 2.10 for the three detectors and the variation of the detectors' recall as a function of clicks' SNR were assessed at these thresholds. The corresponding results

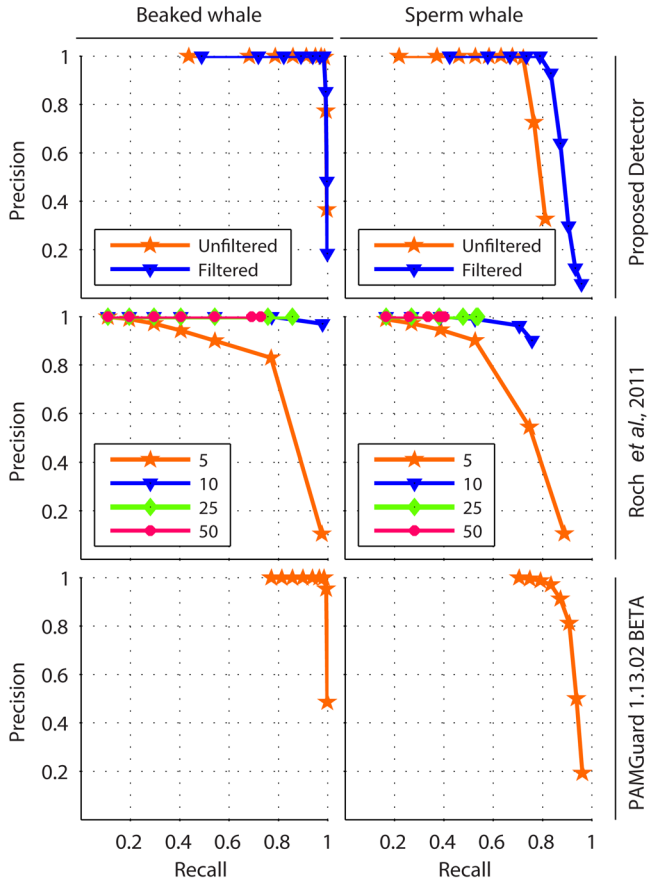

Figure 2.10. Detector performance on synthesised data – Precision-Recall trade-off curves.

are shown in Figure 2.11. Figure 2.12 summarises the detector's performance in capturing the pre-annotated clicks of different species in real underwater audio recordings.

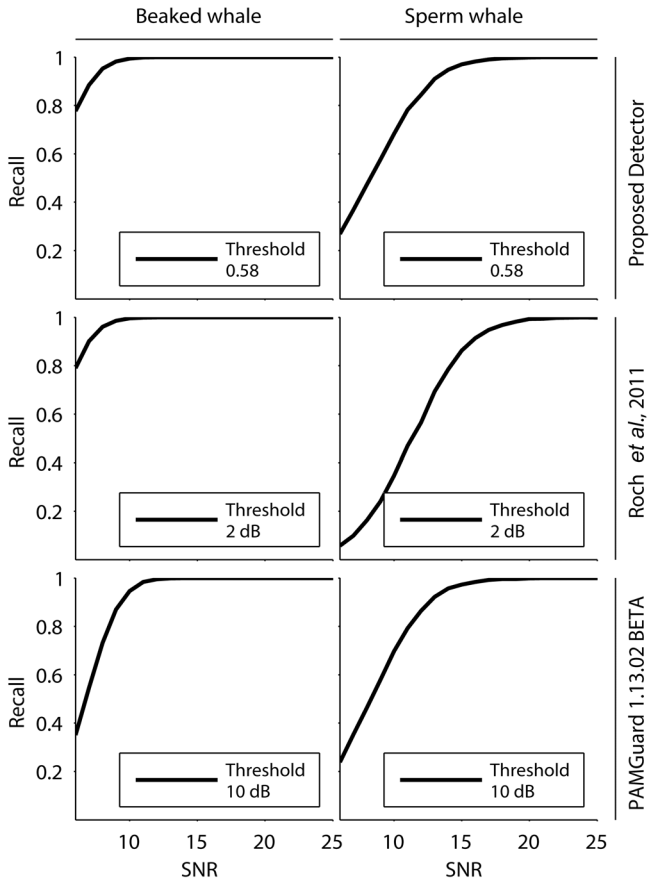

Figure 2.11. Detector performance on synthesised data – Recall vs. SNR. Results for the proposed detector are shown for tests performed with bandpass-filtered inputs. Results for the detector of Roch *et al*. (2011a) are shown for tests performed with a stage 2 threshold of 10 and the plot legend indicates the stage 1 threshold.

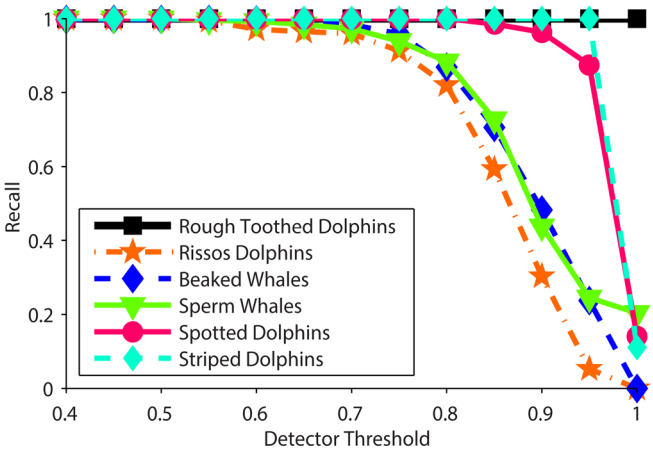

Figure 2.12. Detector recall (as a function of threshold) on real underwater recordings containing echolocation clicks of different species.

For the proposed method, comparing the PR curves for filtered and unfiltered inputs, we can see that improvements in performance can be achieved with appropriate filtering of the input signals. The apparent performance inconsistency for different species, as seen in Figure 2.12, may be attributed to the use of a fixed width MAF1 across all tests. With an appropriate tuning of FWHM (or σ_G) in MAF1, more consistent results may be achieved in the case of tests with real audio and further performance improvements may be possible in the case of tests with synthesised data. However, this is a subject for further investigation.

The real-time factor of a detection/classification system is an indicator of its speed/throughput and is defined as the ratio of the time taken by the system for processing a given input to the duration of the input. Smaller the real-time factor, faster is the system. When tested on a desktop computer with an Intel® i7 CPU and 16 GB of RAM (running Microsoft® Windows 7), a MATLAB implementation of the proposed detector exhibited an average (over different thresholds) real-time factor of 0.019 for 192 kHz audio and 0.007 for 96 kHz audio. For the optimal threshold setting identified from Figure 2.10, the real-time factor was 0.019 as well. When run on the same computer, PAMGuard processed the synthesised data with an average real-time factor of 0.058 at the threshold setting of 10 dB. Meaningful real-time factors could not be determined for the implementation of the method of Roch *et al*. (2011a) owing to the serialisation and the subsequent reloading of intermediate results across stages.

2.5. Discussion

An automatic detector of Gabor-like signals was suggested and tested in this study. As shown with the mathematical formulation, the carrier frequency component of a Gabor-like signal virtually disappears in its TKEO output when the carrier frequency is either constant or varying nearly linearly with time. An additional benefit of this property is that it makes an implementation of the detector immune to species' calling behaviour variations that would affect the clicks' frequency content (Au, 1993, p. 121). This was validated by the performance of the detector on a variety of recorded clicks with no changes in detector settings. The robustness of the system with varying SNRs was demonstrated in the tests with synthesised data. The evaluation with the audio procured from MobySound also showed that the detector worked well with different recording scenarios. The audio recordings were obtained from different geographical locations while the data collection in each set was performed with different recording equipment configurations. The detector exhibited consistency in performance across all recordings used in the tests. In the tests using synthesised data, all the three detectors compared in this analysis exhibited good PR characteristics (satisfactorily high area under the PR curves in Figure 2.10) and, the proposed detector showed improved performance with pre-filtered data. From Figure 2.11 we can see that, at lower SNRs, the proposed detector offered higher recall than PAMGuard for the beaked whale set and higher recall than the method of Roch *et al*., (2011a) for the sperm whale set. The proposed detector also offers significant throughput improvements over PAMGuard – improvements of 67% and 88% for audio sampled at 192 kHz and 96 kHz, respectively. Consequently, the proposed detector can also be used for targeted species' click detection with significant gain in processing speed. Coupled with the simplicity in detector settings, the indicated performance improvements make the proposed detector an attractive choice for various applications.

The angle between the direction of a click's direct propagation path to a receiver and the orientation of the individual producing the click has been shown (e.g.: Au, 1993; Møhl *et al.*, 2003; Au and Würsig, 2004; Madsen *et al.*, 2004; Au *et al.*, 2012) to have an impact on the waveform of the recorded clicks. While it can be argued that

the theoretical signals considered may closely represent on-axis (having little or no relative angles) recorded clicks (Johnson *et al*., 2006), it can be safely assumed that a majority of the clicks captured in open water recordings were off-axis (having high relative angles). Together, the theoretical proof and the experimental validation show that the detector performs well regardless of the calling species' orientation with respect to the recording equipment. A formal analysis of this sub-topic is a subject for further investigation.

The high processing speed and its simple control-flow make the proposed system feasible for pipelined hardware implementations. The few basic mathematical and logical operations that make up the system would take little processing time on modern hardware. Although, there is already noticeable difference in the throughput as compared to PAMGuard (see real-time factors above), an implementation of the proposed system in C/C++ or Java has potential in yielding much higher speeds. Also, the response latency of the system is very small involving a one sample delay caused by the TKEO computation followed by a filter group delay of $\lceil 5\sigma_G/T_s \rceil+1$, resulting in $(N + 2)$ samples. Assuming that an implementation performs the two averaging/filtering operations in a parallel fashion, for the settings considered in the above tests, it can be shown that the maximum delay in reporting detections would be within ~0.8 ms of the occurrence of the clicks.

Chapter 3.
Tonal Detection

Among various types of sounds observed underwater, narrowband acoustic signals occur prominently and are commonly used to characterise the source producing it. Disparate systems have been developed for the detection or recognition of several forms of narrowband signals produced by specific sources. In this chapter I present a generic system, based on post-processing of spectrograms, for the automatic extraction of time-frequency contours of narrowband signals produced by any source – biological, anthropogenic or geophysical. A two-phase approach is proposed where the first phase is based on an image-processing technique for detecting intensity ridges and the second phase is a Bayesian filtering approach for tracing the trajectory of detected ridge apices. The rationale for the various conditionals and choice of system parameters are geared to result in a generic (non-targeted) system and the theoretical motivation for the same are detailed. The performance of the system is tested with real underwater audio containing odontocete whistles and is compared to one of the existing methods.

3.1. Introduction

Numerous approaches are available for the automatic detection of specific types of narrowband (tonal) sounds in underwater audio recordings. Examples of underwater tonal signals include odontocete whistles, mysticete vocalisations, vessel noise and quasi-tonal low-frequency sounds from ice events, such as harmonic tremors from icebergs (Li and Gavrilov, 2008; Talandier *et al*., 2006). Figure 3.1 shows examples of tonal signals from underwater acoustic recordings. The spectral and temporal characteristics of tonal signals in underwater soundscape vary widely. In the time domain, tonal signals may vary from a few milliseconds to minutes or hours. Based on spectral content, tonal signals may occur with constant frequency or varying frequencies over time. Commonly, tonal signals occur with one or more harmonics. Received tonal signals often have amplitude variations which may be the result of

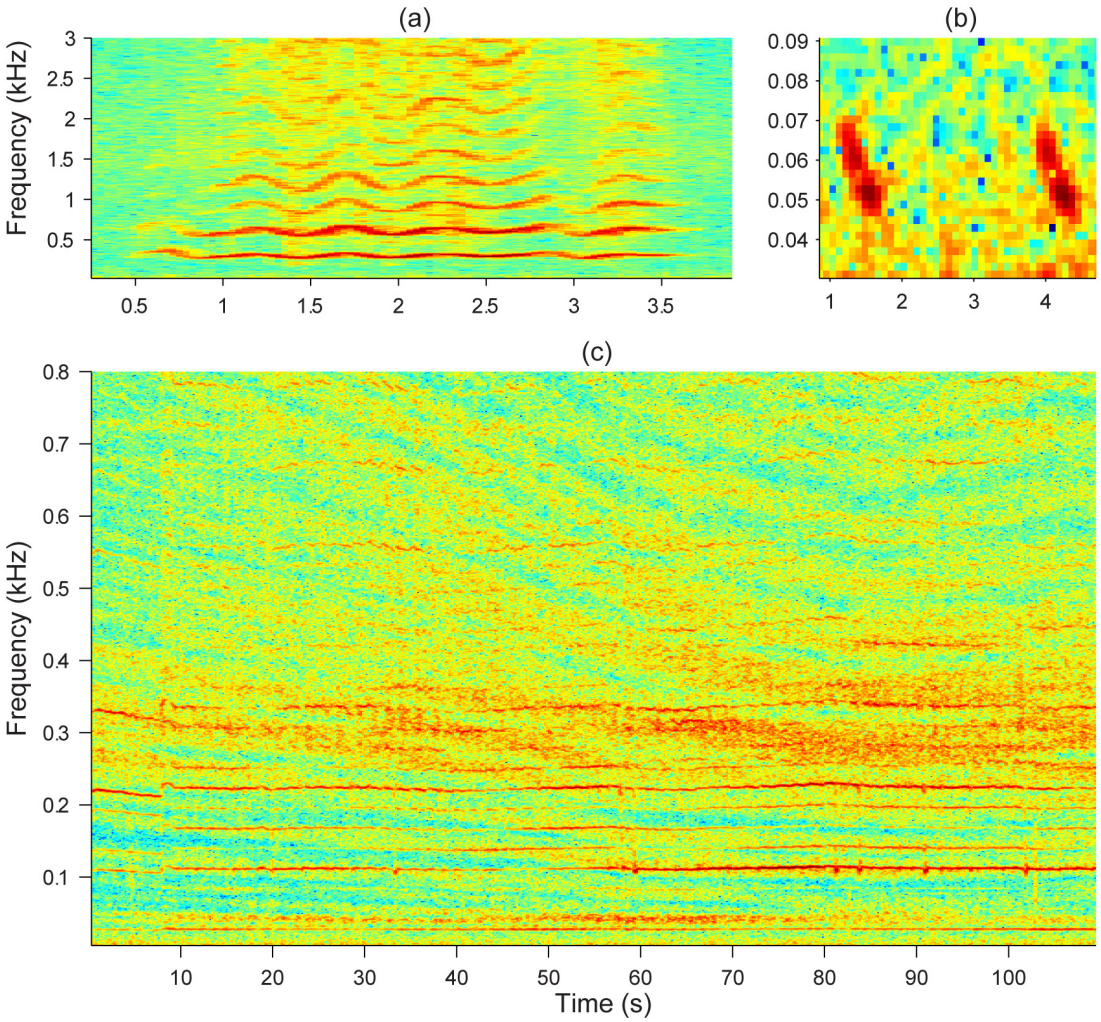

Figure 3.1. Spectrograms of underwater acoustic recordings showing (a) a component of a humpback whale (*Megaptera novaeangliae*) song, (b) rapidly downsweeping tonals from an unidentified source and (c) long-duration sounds from a passing vessel. The diffuse regions of higher energies in (c), besides the sharp tonal components, are a result of interference effects (striations) changing with time due to ship's motion relative to the hydrophone and local bathymetry.

change in source behaviour and orientation, acoustic interferences, properties of the local bathymetry, etc. The range of frequencies that a single tonal signal spans varies for different sources. Characteristics of tonal signals produced by marine mammals varies with species and, for some species (e.g. bottlenose dolphins), varies from one individual to another. Some species, such as humpback whales (*Megaptera novaeangliae*) and blue whales, form songs by repeating patterns of tonal call units and both, the songs and independent units, may change over time. A majority of the available recognition approaches are targeted at identifying one or more specific

types of tonal sounds and some of these approaches generally excel in specific recording scenarios. A recording scenario involves the specific marine environment, its ambient soundscape, the recording and mooring equipment, etc. – all of which can affect the performance of a signal detector. Some of the earlier attempts in the automatic detection of targeted underwater tonal sounds involved the use of simple matched filtering methods for the detection of blue whale calls (Stafford *et al*., 1998; Stafford *et al*., 1994). Spectrogram correlation has been effectively and widely used for many mysticete calls (Mellinger and Clark, 1997) including bowhead whales (*Balaena mysticetus*) (Mellinger and Clark, 2000), right whales (*Eubalaena*) (Munger *et al*., 2005; Urazghildiiev *et al*., 2009) and sei whales (*Balaenoptera borealis*) (Baumgartner *et al*., 2008). Li (2010) described a method for the detection of remote Antarctic ice breakup events based on multivariate classification of frequency dispersed signals from such events. Ogden *et al*. (2011) used autocorrelation in conjunction with comb-like filters and discrete Kalman filtering (Kalman, 1960) for the extraction of harmonic components from the sounds of small boats. Sorensen *et al*. (2010) used wavelet decomposition coupled with filter-banks for initial detection followed by comparative correlation with known reference signals for the subsequent extraction of harmonically related components from sounds of small vessels.

My work is an attempt to realise a generalised system for the automatic extraction of various forms of tonal signals present in underwater acoustic data procured under different recording scenarios. The systems proposed by Baumgartner *et al*. (2011) and Roch *et al*. (2011b) are attempts towards achieving a generalised system and are shown to perform well with the extraction of baleen whale tonal calls and odontocetes whistles, respectively. Mellinger *et al*. (2011) also proposed a generalised tonal signal detector and showed the method to work well with minke whale (*Balaenoptera acutorostrata*) "boing" sounds (Thompson and Friedl, 1982). A majority of signal pre-conditioning operations (e.g. spectral means normalisation) normally employed in the existing systems are targeted at enhancing the detectability of the tonal signals of interest by suppressing unwanted and interfering signals. Contrary to the considerations made in the existing methods, in this study only non-tonal signals (e.g. short pulsed signals and long-lasting broadband sounds) are

considered as "noise." Since the goal in the current approach is simply to detect and trace as many tonal signals as there are in the inputs without regard to the sources producing the underlying sounds, application-specific de-noising of the inputs will not be considered within the algorithm. As will be shown later, one of the components of the proposed system is inherently capable of ignoring a majority of non-tonal signals. The reader must note that the proposed system performs detection alone. Subsequent classification of the detected tonal signals may be task-specific and is beyond the scope of this study.

Spectrograms based on Fourier analysis provide visual representations of the frequency content of audio signals, enabling quick and easy assimilation by human observers in both onsite and offsite monitoring applications. For convenience in dealing with the wide range of spectral power levels P, they are commonly expressed in a logarithmic (dB) scale as $10\log_{10}(P/P_o)$ where P_o is a reference level. Several alternative spectro-temporal representations are available for the analysis of acoustic signals. Examples include chirplet analysis (Mann and Haykin, 1991), cepstral analysis (Oppenheim and Schafer, 1975) and Hilbert analysis (Huang *et al*., 1998). Some recognition techniques based on other spectral analysis methods have been shown to yield better performance over using Fourier analysis in certain targeted automatic passive acoustic monitoring (PAM) applications (e.g. Yang *et al*., 2002; Ioana *et al*., 2006). However, given the availability of highly efficient implementations of fast Fourier transform (FFT) and the widespread use of spectrograms as a means of human-aided analysis, spectrograms are still a very attractive choice as a tool in generalised automatic analysis of underwater audio recordings. Some of the existing automatic tonal detection approaches based on post-processing of spectrograms involve the use of image-processing operations such as image-thresholding and edge-detection (e.g. Datta and Sturtivant, 2002; Gillespie, 2004; Ou *et al*., 2013; Thode *et al*., 2012). Esfahanian *et al*. (2014) employed a family of 32 two-dimensional (2D) Gabor kernels as spatial filters in their approach targeted at dolphin whistles. The approaches described in Halkias and Ellis, 2006; Madhusudhana *et al*., 2008; Roch *et al*., 2011b and Mellinger *et al*., 2011 tackle the contour extraction problem in two phases where prominent spectral peaks in individual frames (spectra) of a spectrogram are identified in the first phase

following which closely lying peaks (in the frequency axis) from neighbouring frames are "connected" to eventually trace the underlying tonal signals' contours. The specifics of the peak-picking and the subsequent contour-tracing sub-processes vary in those approaches. Carevic (2013) proposed a system based on the use of Bayesian modelling and particle filter. However, the efficacy of the system on real audio signals remains yet unassessed. The method described in this chapter is closely related to the approach of Kershenbaum and Roch (2013) in treating spectrogram regions corresponding to tonal signals as intensity ridges. In the approaches of Kershenbaum and Roch (2013) and this study, emphasis lies on utilising additional information which is inherently available in spectrograms within the immediate spectro-temporal neighbourhood of putative tonal signals. The dissimilarities in the two approaches will be highlighted in the following sections.

Ridge-detection is a common technique widely employed in image-processing and computer vision operations (Szeliski, 2010) such as automatic feature selection and image segmentation. In a topographical sense, a ridge is a narrow elongated region of high elevation between two regions of relatively lower elevations. Spectrograms are analogous to 2D greyscale images where spectral intensities at time-frequency (TF) points in the spectrogram correspond to image's pixel intensity. Tonal signals result in narrow ridge-like regions of relatively higher intensities in a spectrogram. Hence, it can be argued that a method based on ridge-detection is well suited for the purpose of extraction of TF contour tracks in spectrograms.

The following section describes the proposed system and its various components. The subsequent sections present tests of the system using real underwater data and provide a discussion on the system's performance.

3.2. Algorithm

The proposed approach employs a modified form of the numerical method developed by Lindeberg (1998a) for the automatic extraction of ridges in images. The suggested modifications are geared specifically to handle spectrograms. Lindeberg (1998a) has

also proposed a method based on a nearest-neighbour approach for tracing the contours formed by ridge apices. Lindeberg's method assumes that putative ridge-like structures are fully contained within an input image. This makes it infeasible for on-site PAM applications where spectrogram frames are continuously generated from input audio. Also, disambiguation of intersecting TF contours requires additional processing and the problem complexity increases when several contours are involved. Kershenbaum and Roch (2013) also employ a similar search scheme within a 3 × 3 pixel neighbourhood. In contrast to these, a tracking approach based on Bayesian filtering is used for "connecting" detected ridge apices across successive frames of a spectrogram. The benefits of this hybrid scheme will become evident in the rest of the chapter.

3.2.1. Input preparation

Spectrograms present a simple means to visually distinguish tonal signals amongst other sounds. This nature of spectrograms could be easily exploited for the purposes of automatic extraction of tonal signals using image-processing based methods. The choice of parameters for spectrogram computation is application specific and is beyond the scope of this chapter. Mellinger *et al*. (2011) proposed an iterative 'Parameter Optimisation Procedure' for choosing an optimal combination of all algorithm parameters which also include spectrogram parameters. Such an elaborate approach may be exhaustive and often unnecessary. As indicated in Kershenbaum and Roch, 2013, for successful application of image-processing based methods, the time resolution Δt and frequency resolution Δf resulting from the chosen spectrogram parameters must allow for spectral and temporal variations in tonal signals to be discernible in a visual sense. On the other hand, choosing fine-grained time or frequency resolutions increases the overall processing time and may not always result in improved detection performance. On the choice of analysis windows, Hamming and Hanning windows, being well suited for the spectral analysis of continuous (non-transient) sounds (Svend and Herlufsen, 1987), are commonly used in the measurement of random noises. With a significantly higher side-lobe falloff rate, the Hanning window offers better separation for weak signals in the spectral vicinity of stronger signals (Harris, 1978). Given the vast variety of sounds possible

in underwater recordings, this benefit of the Hanning window outweighs the marginal loss in frequency resolution as compared to the Hamming window and, hence, could be a favourable choice. The uncertainty principle resulting in a trade-off between time and frequency resolutions dictates that a single set of spectrogram parameters cannot yield similar distinguishing capability across wide ranges in the frequency domain. Spectrogram parameters can be chosen to yield a desired level of distinguishability in the range of frequencies that are of interest to a specific application and the input to the proposed system could be a frequency band-limited portion of the spectrogram. The proposed system is generic in its functionality or control-flow which would remain the same for the processing of any arbitrary band-limited portion of a spectrogram.

3.2.2. Detection of ridge points

The spectrogram of an acoustic signal is analogous to a surface where the elevation at any point on the surface corresponds to spectral intensity at the corresponding TF point. Denoting the first and second derivatives of spectral intensity along the T- and F-axes as ∂T, ∂F, ∂TT, ∂FF and ∂TF, eigenvalues of the Hessian matrix of second derivatives at each TF point are given as

$$\lambda_{(t,f)} = \frac{1}{2}\left(\partial TT_{(t,f)} + \partial FF_{(t,f)} \pm \sqrt{\left(\partial TT_{(t,f)} - \partial FF_{(t,f)}\right)^2 + 4\partial TF^2_{(t,f)}} \right) \tag{3.1}$$

and the orientations of the eigenvectors are given as

$$\beta_{(t,f)} = \arctan\left(\frac{2\partial TF_{(t,f)}}{\partial TT_{(t,f)} + \partial FF_{(t,f)} \pm \sqrt{\left(\partial TT_{(t,f)} - \partial FF_{(t,f)}\right)^2 + 4\partial TF^2_{(t,f)}}} \right). \tag{3.2}$$

By definition, the two eigenvectors determined at each TF point are mutually orthogonal. Local directional derivatives can be defined at each TF point with the introduction of a local (p, q)- system that is aligned with the corresponding eigenvectors. For simplicity, the p- direction β_p shall be considered to correspond to the direction of the eigenvector with larger absolute magnitude, i.e. $|\lambda_p| > |\lambda_q|$. As

such, for TF points that correspond to a ridge-top, β_q aligns with the direction of the contour formed by the ridge-top. Local directional derivatives of intensity in the (p, q)- system can be expressed in terms of the surface derivatives as

$$\begin{aligned} \partial p &= \cos(\beta_q)\partial F - \sin(\beta_q)\partial T, \\ \partial q &= \sin(\beta_q)\partial F + \cos(\beta_q)\partial T. \end{aligned} \tag{3.3}$$

The intensity is locally maximal in the direction of the dominant eigenvector when

$$\begin{aligned} \partial p &= 0 \\ \partial pp &< 0. \end{aligned} \tag{3.4}$$

Note that ∂pp (and ∂qq) is equivalent to λ_p (and λ_q). An additional constraint

$$\partial pp << \partial qq \tag{3.5}$$

would hold only for narrow elongated features such as ridges. As such, the conditions of Eq. (3.4) and Eq. (3.5) suffice in finding points corresponding to tonal signals in a spectrogram. The notion of intensity ridges is demonstrated in Figure 3.2 using a section of a spectrogram of real underwater audio containing bottlenose dolphin (*Tursiops truncatus*) echolocation clicks and whistles. Some detection outcomes corresponding to the ridges in the spectrogram are also shown.

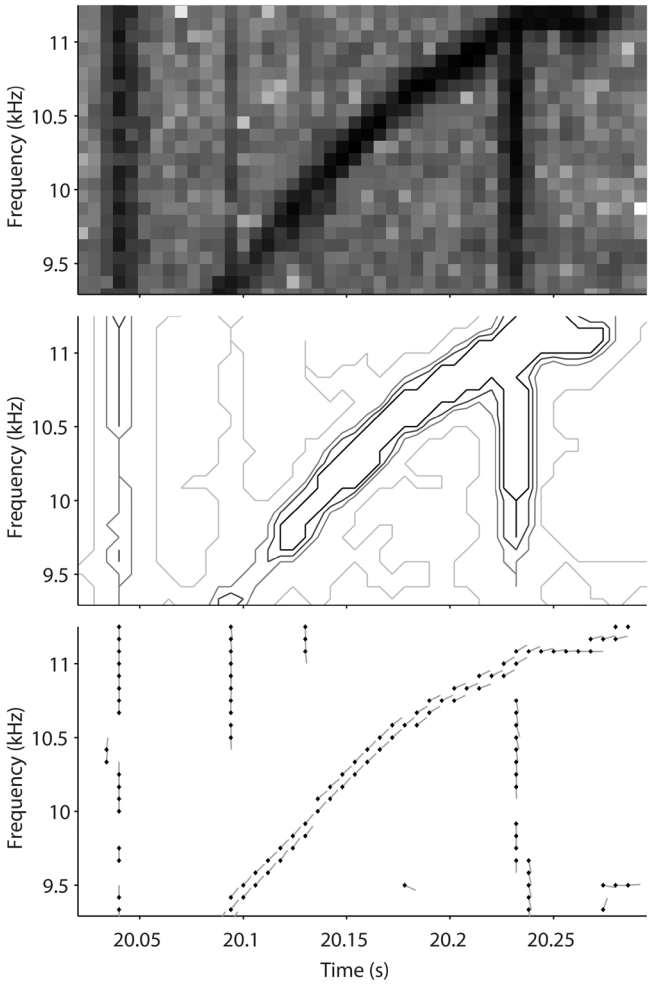

Figure 3.2. Approximated intensity profile curves (middle panel) and some ridge-detection results (bottom panel) corresponding to a sample spectrogram (top panel). In the middle panel, differences in intensity magnitudes are shown using different greyscale colours for different levels - darker curves represent higher intensities. In the bottom panel, TF points where Eqs. (3.4) and (3.5) hold are indicated with black diamond markers. The grey lines emerging out of such points indicate the estimated angle β_q that the non-dominant eigenvector makes with the horizontal axis.

Performing spatial smoothing of the spectrogram prior to estimation of surface derivatives improves ridge detection by suppressing high-frequency noises on the intensity surface. Spatial smoothing of an intensity surface χ using a 2D Gaussian kernel

$$g(x, y; \sigma_G) = \frac{1}{2\pi\sigma_G^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_G^2}\right)$$
$$x = ..., -2\Delta t, -\Delta t, 0, \Delta t, 2\Delta t, ... \quad (3.6)$$
$$y = ..., -2\Delta f, -\Delta f, 0, \Delta f, 2\Delta f, ...$$

is defined by the convolution

$$L(\cdot; \sigma_G) = g(\cdot; \sigma_G) * \chi(\cdot), \quad (3.7)$$

where σ_G is the root mean square (RMS) width of the Gaussian kernel and L is the scale-space representation of the surface at scale σ_G. Spatial smoothing is applied to a linear-scale spectrogram rather than a spectrogram in a logarithmic (decibel) scale since the 2D Gaussian kernels are more akin to the ridge-like features in a linear scale spectrogram. The similarity of ridge-like features in a linear scale spectrogram to Gaussian kernels is shown using a spectrum (a spectrogram frame) as an analogy in Figure 3.3. This contrasts from the approach of Kershenbaum and Roch (2013).

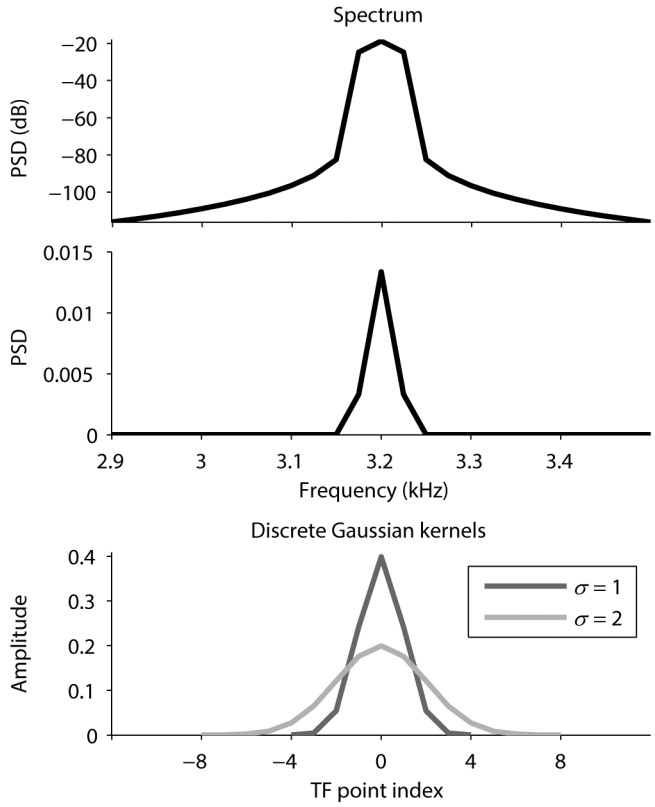

Figure 3.3. Discrete power spectra of a pure sinusoidal signal in logarithmic scale (top panel) and linear scale (middle panel) showing differences in their similarities to 1-dimensional Gaussian kernels (bottom panel).

Also, a linear-scale spectrogram offers better separation of closely-lying ridges. For notational convenience, derivatives ∂u and ∂uv (where u and v are one of T, F, p or q) obtained post spatial smoothing shall be represented as L_u and L_{uv}, respectively. Ridge-defining Eqs. (3.4) and (3.5) can now be expressed in terms of scale-space representation as

$$\begin{aligned} L_p &= 0 \\ L_{pp} &< 0 \end{aligned} \tag{3.8}$$

and

$$L_{pp} << L_{qq}, \tag{3.9}$$

respectively. The quantity $|L_{pp}|$ is indicative of the sharpness of a ridge and hence can be used as a measure of ridge strength. Alternative measures of ridge strength that are more appropriate for spectrograms are presented later. Spectral leakage in FFT computation can dilate the acuteness of intensity ridges corresponding to tonal signals. The dilation becomes more prominent for tonal signals having moderate-to-high absolute rates of frequency modulation (FM), i.e. when $| \beta_q | > \pi/4$. Spatial smoothing using a single kernel may not yield optimal results when applied to spectrograms containing intensity ridges of different widths. Spatial smoothing at multiple scales (using kernels with different σ_G) allows for narrow and coarse ridge-like features in spectrograms to be detected at appropriate scales. For effective spatial smoothing, the width of the kernel must be comparable to the width of the ridge along its dominant curvature. A convenient way to compare ridge and kernel widths is to consider the spectrum's half-power bandwidth (-3dB bandwidth) and the smoothing kernel's "full width at half-maximum" (FWHM). The FWHM of a Gaussian kernel equals $2\sqrt{2\log(2)}\sigma_G \approx 2.355\sigma_G$. The half-power bandwidth for some of the commonly used spectral analysis windows is smaller than 2.355 frequency bins (e.g. 1.54 bins for a Hanning window and 1.30 bins for a Hamming window) (Harris, 1978). Therefore, a smoothing kernel with $\sigma_G \geq (2.355^{-1} \times$ half-power bandwidth) becomes a natural choice. Lindeberg (1998a) proposed a way of choosing multiple scales such as to enable meaningful comparisons of ridge strengths

determined across different scales. Using the notion of a *γ-parameterised normalised derivative* (Lindeberg, 1998b)

$$\partial x_{\gamma-norm} = \sigma_G^{\gamma} \partial x, \tag{3.10}$$

Lindeberg (1998a) expressed the *γ-normalised* second directional derivatives as

$$\begin{aligned} L_{pp,\gamma-norm} &= \sigma_G^{2\gamma} L_{pp} \\ L_{qq,\gamma-norm} &= \sigma_G^{2\gamma} L_{qq} \end{aligned}, \tag{3.11}$$

and showed that at any scale, the scale-space representation approximates the width of an underlying ridge when γ = ¾. Having better correspondence to true ridge widths, $|L_{pp,\gamma\text{-}norm}|$ offers a better alternative over $|L_{pp}|$ for making ridge strength comparisons across scales. Selection of ridge points detected in different scale-spaces is based on their strengths forming local maxima across successive scale-spaces, i.e.

$$\begin{aligned} &\partial \sigma_G \left(\left| L_{pp,\gamma-norm}(t,f;\sigma_G) \right| \right) = 0, \\ &\partial \sigma_G \sigma_G \left(\left| L_{pp,\gamma-norm}(t,f;\sigma_G) \right| \right) < 0. \end{aligned} \tag{3.12}$$

Empirical analysis with a few spectrograms containing tonal signals showed that spatial smoothing at two scales, σ_G = 1 and σ_G = 2, sufficed in successfully extracting the corresponding ridges.

Both $|L_{pp}|$ and $|L_{pp,\gamma\text{-}norm}|$ values are sensitive to the height of a ridge as well as the signal-to-noise ratio (SNR) resulting from the background levels and surrounding noises in the ridge's TF neighbourhood. Comparing $|L_{pp}|$ or $|L_{pp,\gamma\text{-}norm}|$ against fixed thresholds would manifest a bias to strong signals. To overcome this, the *γ-normalised* second directional derivatives are further normalised with the height of the intensity ridge in the corresponding scale space, i.e.

$$\begin{aligned} L_{pp,scale-norm}(t,f;\sigma_G) &= \frac{L_{pp,\gamma-norm}(t,f;\sigma_G)}{L(t,f;\sigma_G)} \\ L_{qq,scale-norm}(t,f;\sigma_G) &= \frac{L_{qq,\gamma-norm}(t,f;\sigma_G)}{L(t,f;\sigma_G)} \end{aligned}. \tag{3.13}$$

This allows us to express ridge strength measurements as relative values, thereby making comparisons against predefined thresholds more convenient. In the ridge-tracing procedure described in Section 3.2.3, use of $|L_{pp,scale\text{-}norm}|$ as the preferred ridge strength measure better aids the tracing of ridges corresponding to amplitude modulated tonal signals and ridges spanning time intervals with varying noise levels.

The utilization of the ridge detection criteria described above is demonstrated here using synthetic tonal signals. Three sinusoidal signals having frequencies of 6 kHz, 12 kHz and 18 kHz and peak amplitudes of 1.6, 0.5 and 0.27, respectively, were superimposed to produce a composite signal. Gaussian noise with an RMS amplitude of 0.7 was added to the composite signal. This resulted in average signal-to-noise ratio (SNR) values of 20 dB, 10 dB and 5 dB for the three tonal signals. Spectrograms were calculated from both the clean composite signal and its noise-added variant following which spatial smoothing was applied at scales $\sigma_G = 1$ and $\sigma_G = 2$. The resulting $|L_{pp,\gamma\text{-}norm}|$ and $|L_{pp,scale\text{-}norm}|$ values are shown in Figure 3.4. The values of $|L_{pp,\gamma\text{-}norm}|$ are higher for scale $\sigma_G = 1$ than scale $\sigma_G = 2$ indicating the suitability of smaller scale for narrow ridge-like structures as per Eq. (3.12). High sensitivity of $|L_{pp,\gamma\text{-}norm}|$ to signal intensity and SNR is evident at both scales. In the case of non-noisy signals, the similarities in the peak $|L_{pp,scale\text{-}norm}|$ values indicate that the additional normalisation eliminates sensitivity to signal intensity. In the case of added noise, drops of 10 dB and 15 dB in SNR correspond to lowering of $|L_{pp,\gamma\text{-}norm}|$ by 90% and 97%, respectively. In contrast, the respective drops observed in $|L_{pp,scale\text{-}norm}|$ were 13% and 34% for scale $\sigma_G = 1$, and 22% and 46% for scale $\sigma_G = 2$. Closely lying unrelated local spectral peaks in TF space sometimes produce ridge-like structures in scale-space representations and result in moderate $|L_{pp,scale\text{-}norm}|$ values at non-ridge TF points. Examples of this can be seen in Figure 3.4 for the noise-added case. Detection of such spectral structures, which would otherwise result in false-positives, can be avoided by ignoring TF points having $|L_{pp,scale\text{-}norm}|$ values smaller than a limit. In the noise-added variant, a limit of 0.25 allowed for the acceptance of all ridge points corresponding to the high and mid intensity signals. Over all frames of the spectrogram, the mid intensity signal's effective SNR in each frame varied in the range 4 – 15 dB. In the case of the weakest signal, its effective per-frame SNR ranged from -12 dB to 13 dB and its $|L_{pp,scale\text{-}norm}|$

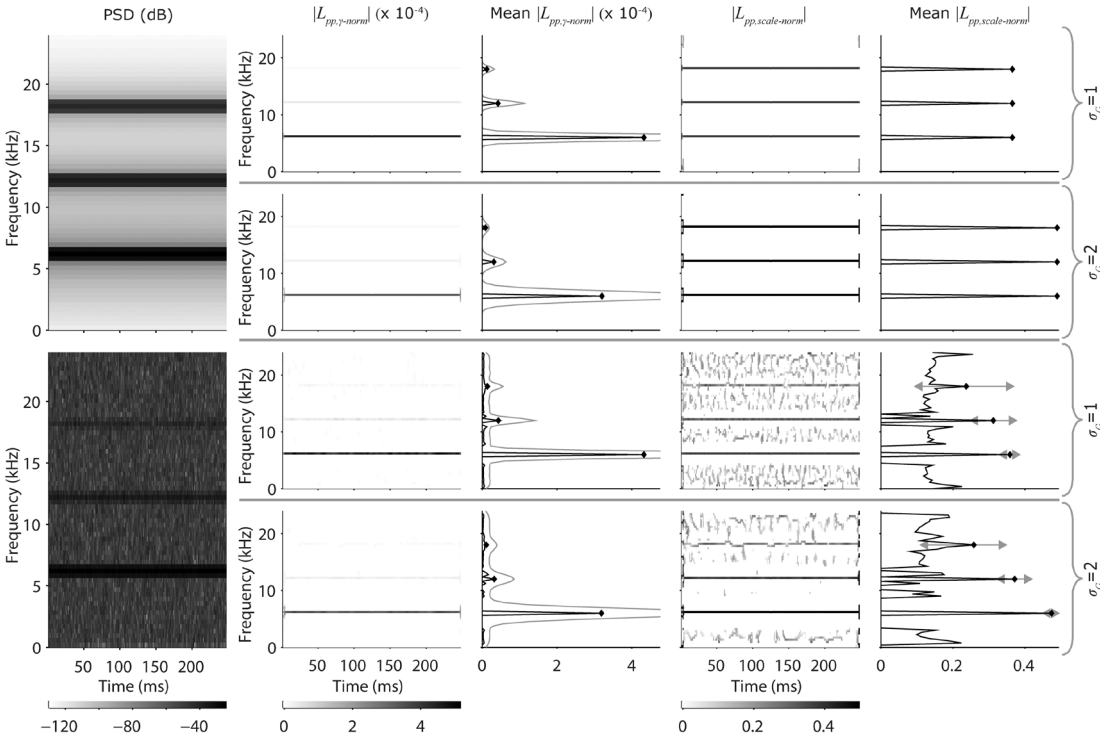

Figure 3.4. Comparison of normalised ridge strength measures $|L_{pp,\gamma\text{-}norm}|$ and $|L_{pp,scale\text{-}norm}|$ shown for synthesised data (see text for details). The first column shows spectrograms (parameters: 2.6 ms Hanning window, 1.3 ms overlap) produced using clean (top) and noisy (bottom) signals. The corresponding $|L_{pp,\gamma\text{-}norm}|$ and $|L_{pp,scale\text{-}norm}|$ are shown in the second and fourth columns, respectively, for the considered scale-spaces. For TF points that satisfy Eq. (3.8) (i.e. ridge points), $|L_{pp,\gamma\text{-}norm}|$ and $|L_{pp,scale\text{-}norm}|$ values are averaged over time and the averages in each frequency bin are shown with black lines in the third and fifth columns, respectively. The averages corresponding to the three tonal signals are highlighted with diamond markers. Time averaged scale-space values L at each frequency bin are shown as grey lines in the third column (clipped for overall clarity). In the case of added noise, the full range of $|L_{pp,scale\text{-}norm}|$ values corresponding to all ridge points are shown with grey arrows.

values exceeded 0.25 in nearly 60% of all spectrogram frames. Since the $|L_{pp,scale\text{-}norm}|$ values are nearly independent of signal SNR and intensity, it can be argued that the check

$$\left|L_{pp,scale-norm}\right| > 0.25 \tag{3.14}$$

would have similar effect for other well-generated spectrograms as well.

For discrete systems, the first of the conditions in Eq. (3.8) is realised by determining zero-crossings in L_p between neighbouring TF points. The condition in Eq. (3.9) is realised by checking that the difference $|L_{pp,scale\text{-}norm} - L_{qq,scale\text{-}norm}|$ is not small. As seen in Figure 3.4, even for free-of-noise tonal signals, $|L_{pp,scale\text{-}norm}|$ remains less than 1. In comparison to possible range of values that $|L_{pp,scale\text{-}norm}|$ could have, i.e. 0.25 – 1, a difference of 0.1 is a significant quantity. As such, the check in Eq. (3.9) is achieved as

$$\left(L_{pp,scale-norm} - L_{qq,scale-norm}\right)^2 > 0.1^2 . \tag{3.15}$$

Note that the left-hand side quantities in Eqs. (3.14) and (3.15) are analogous to the ridge strength measures $\mathcal{M}_{\gamma\text{-}norm}L$ and $\mathcal{A}_{\gamma\text{-}norm}L$, respectively, as defined in Lindeberg, 1998a. Sometimes, it may be appropriate for certain analyses to ignore weak tonal signals having energy lower than some pre-defined value. Identified ridge points are discarded by the algorithm where the smoothed spectral intensities L at the corresponding TF points are lower than `min_intensity` [1].

At this point, I would like to highlight three characteristics of the ridge detection component. First, the form of spatial smoothing employed and the subsequent computation of intensity derivatives ensures that only local information from immediate spectral and temporal vicinity is factored into decision making. Such an approach is well suited for acoustically dynamic environments. Second, Eq. (3.9) ensures that TF points corresponding to non-ridge-like structures, such as blobs, are rejected although they may correspond to local maxima in spectral intensity. Third, ridge points for which $|\beta_q|$ approaches 90° correspond to vertical ridges in a spectrogram that are likely caused by short broadband signals and hence can be ignored.

The ridge strength measure $|L_{pp,scale\text{-}norm}|$ and the orientation of the ridge-top β_q are included in the ridge-detector outputs for each detected ridge point. In computing the angle β_q that L_{qq} makes with the time axis, restricting the arctangent calculations to just the first and fourth quadrants ensures that β_q is always in the direction of

[1] User-settable algorithm parameters will be denoted using monospaced font.

increasing time. Note that in the determination of angles β_p and β_q, unit values considered along the T- and F-axes are the input spectrogram's time and frequency resolutions, respectively. Appropriate scaling is necessary wherever the angles are used. For convenience with the use of subscripts through the remainder of the chapter, the quantity $|L_{pp,scale\text{-}norm}|$ and the angle β_q shall be represented with the symbols τ and ψ, respectively.

3.2.3. Tracing ridge contours

Tracing of a TF contour along increasing time in a spectrogram is analogous to tracing the path of a point object moving in a 2D space with zero acceleration along one of the dimensions. Bayesian filtering approaches are a popular choice in solving 2D and 3D object tracking problems. Correct modelling of the underlying process driving the object's motion enables successful tracking of the object. The process models proposed in Mallawaarachchi, 2008; Roch *et al.*, 2011b and Kershenbaum and Roch, 2013 for TF contour tracing or smoothing only utilise a TF contour's peak frequency and its derivatives. The spectral power (ρ) of narrowband signals is generally overlooked in TF contour tracing operations; however, its consideration may aid with the disambiguation of intersecting contours and also help with the separation of closely spaced contours. The ridge detection phase (Section 3.2.2) makes available additional local information at each ridge point – ridge strength τ and orientation ψ. The frequency f at a detected ridge point will be referred to as peak frequency in this section. Including f, ρ, τ and ψ together in the list of contour features enable a more complete stochastic model for the purposes of predictive contour tracing.

Changes in the peak frequency f of a narrowband signal across successive frames in a spectrogram can be formally expressed as

$$f_{i+1} = f_i + \dot{f}_i \Delta t \,, \tag{3.16}$$

where f_i and $\dot{f}_i$ are the underlying signal's peak frequency and FM rate respectively, at the i^{th} frame. The underlying process driving the frequency modulation of

narrowband signals is dependent on the source and cannot be modelled. Although it can be measured *post hoc*, prediction of $\dot{f}_i$ with acceptable levels of accuracy is problematic. Since ψ provides an indication of the orientation of a TF contour's ridge at each frame, it can be considered as the contour's "steering" factor at the corresponding frame. Therefore, $\dot{f}_i$ can be determined at each frame as

$$\dot{f}_i = \tan(\psi_i)\frac{\Delta f}{\Delta t}. \tag{3.17}$$

This is contrary to the process model used in Mallawaarachchi, 2008, where $\dot{f}_i$ is predicted at each step from ongoing state updates in the Kalman filtering process. Note that the estimation of ψ in Section 3.2.2 considered unit lengths along time and frequency axes to be Δt and Δf, respectively. Hence the trigonometric tangent value in Eq. (3.17) is scaled with the quantity (Δf / Δt). The measured spectral power and estimates of the ridge strength of narrowband signals can be affected by several factors such as amplitude modulation at the source, multipath sound propagation in the underwater sound channel resulting in constructive and destructive interferences, presence of overlapping sounds and varying background noise levels. As such, ρ and τ of a TF contour can vary considerably over its duration. However, I shall treat them as temporally invariant quantities –

$$\begin{aligned} \rho_{i+1} &= \rho_i \\ \tau_{i+1} &= \tau_i \end{aligned}; \tag{3.18}$$

and allow for the intra-contour variations in their magnitudes to be addressed using variance estimates obtained *a priori* using training data, which will be discussed later in this section.

Given that the observable quantities f, ρ and τ can be predicted as per Eqs. (3.16) and (3.18), the state of a contour at frame i shall be defined using these quantities as [2]

[2] Labelling of vectors and matrices in the Kalman filtering steps follows common convention. Vectors shall be represented using lowercase boldface letters and matrices shall be represented using capital letters.

$$\boldsymbol{x}_i = \begin{bmatrix} f_i \\ \rho_i \\ \tau_i \end{bmatrix}. \tag{3.19}$$

A linear time invariant (LTI) stochastic state-space model for the prediction of a contour's new state (at the next frame) can be defined as

$$\begin{aligned} \hat{\boldsymbol{x}}_{i+1} &= A\boldsymbol{x}_i + B\boldsymbol{u}_i \\ A &= I_3 \\ B &= \begin{bmatrix} \Delta t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \end{aligned} \tag{3.20}$$

where $\hat{}$ indicates that the vector is a prediction, A is the *state transition* matrix, B is the *control* matrix, I_3 is a 3×3 identity matrix and $\boldsymbol{u}_i = \begin{bmatrix} \dot{f}_i & 0 & 0 \end{bmatrix}^T$ is the TF contour's control vector at the i^{th} frame which drives its instantaneous frequency changes.

Kalman filtering methods include two types of errors or noises in modelling a problem – process noise and measurement noise. These will be briefly explained here as they apply to the problem of TF contour tracing. The estimation of f, ρ, τ and ψ suffers from noises from multiple stages of signal processing. Since the tracing operation only considers data available from a spectrogram, the noises in f and ρ estimations caused by PAM hardware and the spectrogram generation process itself (artefacts and spectral leakage) can be conveniently ignored. An undesired effect of the spatial smoothing performed in Section 3.2.2 is the apparent shifting of a TF contour's ridge points towards high energy regions in the immediate TF vicinity of the underlying tonal signal. As a result, the estimation of f, ρ, τ and ψ as described in Section 3.2.2 are bound to have errors. These errors constitute the tracing process' "measurement" noise. In contrast to measurement noise, process noise quantifies the differences in a contour's observed and predicted states. With the process model defined in Eq. (3.20), an assumption that the estimates of ψ are free of measurement noise allows us to treat the underlying estimation errors as process errors in f. As such, characteristics of process errors in f, ρ and τ can be quantified using hand-

traced contours from training data as the process error covariance Q (see Appendix A.1). Determining the measurement error covariance R is, however, not a straightforward process. As will be described later, employing an alternate view on R will enable us to choose meaningful values for the independent measurement error variances for f, ρ and τ.

The process errors can be assumed to be Gaussian distributed based on the analysis presented using training data in Appendix A.1. Taking this into consideration and given the LTI system presented in Eq. (3.20), Kalman filtering becomes a natural choice for solving the tracking problem. In Kalman filtering, the *state prediction error* [3] covariance is estimated as

$$\hat{P}_{i+1} = AP_iA^T + Q, \tag{3.21}$$

and subsequently, the *innovation* covariance S_{i+1} and the *Kalman gain* K_{i+1} are determined as

$$S_{i+1} = H\hat{P}_{i+1}H^T + R, \text{ and} \tag{3.22}$$

$$K_{i+1} = \hat{P}_{i+1}H^T S_{i+1}^{-1}, \tag{3.23}$$

where $H = I_3$ is the *observation* matrix.

Multiple ridge points Γ_n (n = 1, 2, …, N) may be detected at any frame. The process of choosing the most suitable ridge point for extending a contour is described later in the section. The f, ρ and τ values of a detected ridge point Γ_n make up its *measurement* vector $\boldsymbol{z}_n$. Due to the occurrence of process and measurement errors, a contour's predicted state $\hat{\boldsymbol{x}}$ may not align with the chosen ridge point's *measurement vector*. Extension of a contour using the chosen ridge point Γ_n is brought about via the *innovation*, *state update* and *prediction error covariance update* steps described respectively in Eqs. (3.24), (3.25) and (3.26) below

$$\tilde{\boldsymbol{y}} = \boldsymbol{z}_n - H\hat{\boldsymbol{x}}_{i+1} \tag{3.24}$$

[3] Kalman filtering elements are denoted using italicised font in this section.

$$\boldsymbol{x}_{i+1} = \hat{\boldsymbol{x}}_{i+1} + K\tilde{\boldsymbol{y}} \tag{3.25}$$

$$P_{i+1} = (I - KH)\hat{P}_{i+1} \tag{3.26}$$

During the frame-wise processing of a spectrogram, multiple contours (m = 1, 2, …, M) may be active at any frame. At frame i + 1, the frontier of an active contour m is described by the f, ρ and τ values in its predicted state $\hat{\boldsymbol{x}}_{i+1|m}$ and its orientation $\tilde{\psi}_m$ at frame i. The orientation $\tilde{\psi}_m$ is determined using the ridge point Γ_n which had extended the contour at frame i and the difference in f values of $\boldsymbol{x}_{i|m}$ and $\boldsymbol{z}_n$ as described in Figure 3.5. For each of the $M \times N$ pairwise combinations of active contours and ridge points available at frame i + 1, a cost function $C(m, n)$ is determined based on the 4-dimensional (f, ρ, τ and ψ) Mahalanobis distance (Mahalanobis, 1936) between Γ_n and contours' frontiers. Contrary to the consideration of only f, ρ and τ in state and measurement vectors, inclusion of ψ in the cost function allows for penalising candidate extensions that may otherwise lead the tracing astray.

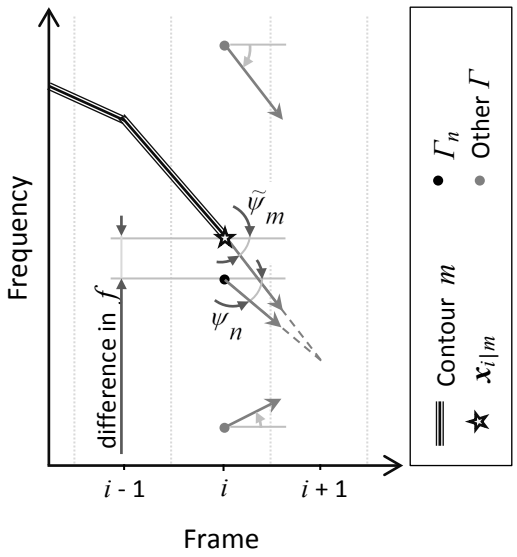

Figure 3.5. Elements required for the computation of a contour's orientation, shown using synthetic data. The grey arrows emerging from the ridge points indicate their orientations. The "difference in f" and ψ_n are used to determine $\tilde{\psi}_m$ such that the contour's frontier and Γ_n point to the same location in the succeding frame as indicated with broken lines.

Considering good candidate extensions to be those that lie within three standard deviations of the respective means along each of the four dimensions, an upper limit on $C(m, n)$ can be defined as $\sqrt{3^2+3^2+3^2+3^2}$. Sudden and extreme directional changes in ridge contours are uncommon in spectrograms. An upper limit on the disparity in orientations of an active contour and a candidate extension ridge point was empirically chosen to be 100°. Further, ridge points occurring considerably away from a contour's frontier, along the frequency axis, are not considered extension candidates. The acceptable range of frequencies for a contour is determined from its $\tilde{\psi}_m$ and its frequency at the previous frame $f_{i|m}$. Possible dubious pairings are discarded where any of the below checks fail:

i. $C(m, n) \leq \sqrt{3^2 \times 4}$,

ii. $|\tilde{\psi}_m - \psi_n| \leq 100°$,

iii. $\text{f\_min}(f_{i|m}, \widetilde{\psi}_m) \leq f_n \leq \text{f\_max}(f_{i|m}, \widetilde{\psi}_m)$ [4],

iv. $\arctan\left(\overline{\text{F}}^{-}(f_n) \cdot \Delta t/\Delta f\right) < \psi_n < \arctan\left(\overline{\text{F}}^{+}(f_n) \cdot \Delta t/\Delta f\right)$ [5].

The last of the above checks discards candidate pairs containing ridge points that likely correspond to vertical ridges. From the set of candidate pairings that pass the above checks, a subset is to be chosen such that a one-to-one mapping of contours to ridge points exists in the subset and that the resulting sum of the costs C of the pairings in the subset is a minimum. This is achieved using the Munkres algorithm (Munkres, 1957), which solves the assignment problem with a computational complexity of $O(n^3)$. The measurement vectors corresponding to the ridge points so chosen for each contour are used in the extension of the corresponding contours as described with Eqs. (3.24) through (3.26).

At any frame, if no suitable Γ_n is available for the extension of a traced contour, the contour is extended using the predicted state, i.e., $\boldsymbol{x}_{i+1}$ is set to $\hat{\boldsymbol{x}}_{i+1}$. This is helpful with tonal signals having short temporal discontinuities (apparent or real) in a spectrogram. Since ψ values are unavailable through the discontinuities, the control vector $\boldsymbol{u}_i$ remains unaltered through the predicted extensions. Making extensions using predicted states is only permitted for up to a few successive frames and tracing of contours will cease when the limit, controlled by the parameter `max_contour_inactivity`, is reached. Contours whose tracing have ceased are reported if their durations without the trailing predicted extensions exceed `min_contour_length` and are otherwise discarded. Also, while a contour being traced is shorter than `min_contour_length` and has 40% or more predicted extensions, then it is immediately discarded.

At any frame, detected ridge points that passed check (iv) above are gathered into groups of contiguous points. If any group has local minima in ridge strength, the corresponding valley points are eliminated and the group becomes further divided. In

[4] To enable clarity in algorithm flow, descriptions of functions f_min(·) and f_max(·) have been provided in Appendix A.2.

[5] The functions $\overline{\text{F}}^{+}(\cdot)$ and $\overline{\text{F}}^{-}(\cdot)$ are defined in Appendix A.1.

each of the resulting groupings, if one or more contained ridge points were used in contour extensions, the corresponding group is marked as used. Of the groups not marked as used, ridge points corresponding to local maxima (in τ) within the groups are identified and are used for starting new contour traces. Allowing non-maximal ridge points to also be considered for extensions is beneficial in the case of intersecting contours. Restricting the starting of new traces to just the strongest ridge points suppresses concurrent non-dominant traces.

The noise in measurements of f, ρ and τ caused by spatial smoothing may be characterised using training data for a specific combination of spectrogram and smoothing kernel parameters. Obtaining such characteristics for a general purpose solution would be impractical. R influences how the chosen observation $\boldsymbol{z}_n$ and the state prediction $\hat{\boldsymbol{x}}_{i+1}$ are relatively weighted in determining the state update $\boldsymbol{x}_{i+1}$. Small values in R are indicative of higher confidence in measurement, and result in $\boldsymbol{x}_{i+1}$ occurring closer to $\boldsymbol{z}_n$ than $\hat{\boldsymbol{x}}_{i+1}$. Smoothing of the traced contour in both f and ρ dimensions can be achieved by increasing the respective measurement error variances in R. Smoothing in the τ dimension is not of much concern as the ridge strength measure is not part of algorithm outputs. Setting the corresponding value in R to zero simply forces strict adherence to measured values of τ. Spatial smoothing performed in the ridge detection phase (Section 3.2.2) may shift ridge apices by one or more frequency bins. Since frequency values in $\boldsymbol{z}_n$ are multiples of Δf, choosing a standard deviation value considerably smaller than Δf for the measurement error in the f dimension may result in sharp inflections in tracing results. Preliminary trials showed that a measurement error covariance of

$$R = \begin{bmatrix} \Delta f^2 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.27}$$

resulted in satisfactory smoothing of the traced contours in the f dimension. The degree of smoothing in the ρ domain was not evaluated, and choosing an appropriate value for ρ measurement error variance is a subject for further investigation. The current value chosen to be 10, corresponds to a 5 dB standard deviation in the measurement error of ρ.

3.2.4. A note on algorithm parameters

In the algorithm presented in this chapter, the values suggested in almost all of the heuristics checks are expected to yield similar performance for different inputs, and the reasoning for the values chosen have been provided wherever appropriate. However, choosing globally optimal values for the parameters `min_intensity`, `min_contour_length` and `max_contour_inactivity` is not a trivial task. It is, sometimes, a subjective matter. For example, some applications may choose to ignore very weak or very short signals; the notion of whether closely spaced successive ridges in a spectrogram are to be considered as fragments of a single tonal signal or as different individual signals varies from one analyst to another. As such, it is up to an end user to choose values for these three parameters as may be appropriate for an application, while all other numeric criteria in the algorithm can be retained unchanged.

3.3. Testing

The performance of the proposed algorithm was evaluated using a MATLAB based implementation. For a general purpose TF contour tracker, such as the one proposed in this study, developing systematic tests to cover a variety of testing conditions (analysis bandwidths, noise levels, overlapping contours, etc.) is problematic. A performance analysis approach that yields five quality metrics was proposed in Roch *et al*., 2011b. An implementation of that analysis approach, available at http://roch.sdsu.edu/software/silbido-1.1-beta2.zip (accessed on August 24, 2015), was employed for assessing performance of the proposed algorithm. The definitions of the different metrics are briefly reiterated here. A reported detection is considered to correspond to a ground truth TF contour when the average frequency deviation across the duration of their temporal overlap is less than 350 Hz. The metric "deviation" indicates the average disparity in frequency. Reported detections that correspond to any ground truth TF contours are considered valid detections, otherwise they are considered as false positives. A ground truth TF contour that has correspondence to one or more reported detections is said to be matched. The metric

"recall" is defined as the ratio of the number of matched TF contours to the total number of ground truth TF contours. The metric "precision" is defined as the ratio of the number of valid detections to the total number of reported detections. Multiple temporally-disjoint detections may correspond to a single ground truth TF contour. The metric "fragmentation" is the average number of reported detections per ground truth TF contour. The metric "coverage" is the average measure of the percentage of ground truth TF contours' durations covered by reported detections.

The study presented in Roch *et al*., 2011b demonstrates the performance of two approaches tested with real world underwater audio recordings available from the MobySound archive (Heimlich *et al*., 2011). The proposed algorithm was tested using the same dataset, restricting the tests to only those audio files for which annotations made by a trained analyst were available. The audio data were downsampled to 96 kHz after bandpass filtering to suppress acoustic energies below 5 kHz and above the resulting Nyquist frequency of 48 kHz. Most of the annotated signals were known to occur within this frequency range (Roch *et al*., 2011b). Using MATLAB's `spectrogram` function, linear magnitude spectrograms were calculated using Hanning-windowed frames of 8 ms duration (resulting in Δf = 125 Hz) with 50% overlap (resulting in Δt = 4 ms). As with the tests described in Roch *et al*., 2011b, only those annotated contours that are longer than 150 ms and having SNR >= 10 dB for at least one third of their duration were considered to represent ground truth detections. A total of 2686 annotated TF contours passed the chosen criteria. The list of audio files and the corresponding test results are presented in Table 3.1.

Table 3.1. Performance analysis of the proposed algorithm with test data from the MobySound archive containing tonal calls from four species of the delphinidae family – bottlenose dolphins (*Tursiops truncatus*), melon-headed whales (*Peponocephala electra*) and long-beaked and short-beaked common dolphins (*Delphinus capensis* and *Delphinus delphis*, respectively).

	Audio File	Ground Truth Signals	Precision %	Recall %	Deviation ± σ Hz	Coverage ± σ %	Fragments
Bottlenose dolphin	Qx-Tt-SCI0608-N1-060814-121518.wav	89	81.9	89.9	117 ± 47	79.0 ± 24.3	2.1
	palmyra092007FS192-070924-205305.wav	143	93.4	99.3	102 ± 43	83.3 ± 20.1	1.8
	palmyra092007FS192-070924-205730.wav	122	90.3	97.5	104 ± 47	79.5 ± 19.1	1.8
	All	354	89.0	96.3	106 ± 45	81.0 ± 20.9	1.8
Long-beaked common dolphin	Qx-Dc-CC0411-TAT11-CH2-041114-154040-s.wav	300	81.2	88.7	89 ± 50	67.1 ± 26.3	1.5
	Qx-Dc-SC03-TAT09-060516-171606.wav	10	73.9	100.0	53 ± 14	68.8 ± 27.9	1.7
	QX-Dc-FLIP0610-VLA-061015-165000.wav	247	92.2	88.3	89 ± 41	72.7 ± 23.6	1.6
	All	557	85.6	88.7	88 ± 46	69.5 ± 25.3	1.5
Melon-headed whale	palmyra092007FS192-070925-023000.wav	90	70.2	87.8	80 ± 37	61.7 ± 24.6	1.4
	palmyra092007FS192-071004-032342.wav	78	90.8	93.6	88 ± 39	80.6 ± 20.9	1.5
	palmyra102006-061020-204327_4.wav	170	92.1	85.9	80 ± 54	66.4 ± 25.7	1.4
	All	338	84.7	88.2	82 ± 47	68.6 ± 25.3	1.4
Short-beaked common dolphin	Qx-Dd-SCI0608-N1-060815-100318.wav	92	88.4	94.6	83 ± 39	81.7 ± 22.1	1.6
	Qx-Dd-SCI0608-Ziph-060817-100219.wav	1112	50.4	70.1	113 ± 61	70.3 ± 26.8	1.5
	Qx-Dd-SCI0608-Ziph-060817-125009.wav	233	79.3	89.7	92 ± 43	74.6 ± 24.2	1.6
	All	1437	56.9	74.9	107 ± 57	72.1 ± 26.2	1.5

Since no calibration data was available, all spectral intensity levels mentioned in this section are relative to that of the maximum amplitude of the recordings in normalised units. Annotated whistles having peak spectral levels lower than -115 dB re 1 unit^2/Hz were observed and the lower extremes differed for each file. For effective detection of low-intensity ridge points, the `min_intensity` parameter had to be set appropriately for each file. Setting a high value resulted in many missed detections (low recall). Setting a very low value resulted in faint signals being reported as well. As the available annotations do not include such low intensity signals, this results in an apparent drop in precision. Values were chosen to achieve a balance between precision and recall rates. The `min_intensity` values chosen for each file are listed in Table 3.2. When calibration data is available, these values need to be adjusted appropriately. The other algorithm parameters, chosen empirically, remained the same across all testing and are listed in Table 3.3.

Table 3.2. Test-specific settings of the parameter min_intensity. The values of intensity levels are relative to the highest intensity in the respective audio files.

Audio File	**`min_intensity`**
Qx-Tt-SCI0608-N1-060814-121518.wav	-105 dB
palmyra092007FS192-070924-205305.wav	-75 dB
palmyra092007FS192-070924-205730.wav	-65 dB
Qx-Dc-CC0411-TAT11-CH2-041114-154040-s.wav	-115 dB
Qx-Dc-SC03-TAT09-060516-171606.wav	-103 dB
QX-Dc-FLIP0610-VLA-061015-165000.wav	-87 dB
palmyra092007FS192-070925-023000.wav	-78 dB
palmyra092007FS192-071004-032342.wav	-90 dB
palmyra102006-061020-204327_4.wav	-100 dB
Qx-Dd-SCI0608-N1-060815-100318.wav	-85 dB
Qx-Dd-SCI0608-Ziph-060817-100219.wav	-105 dB
Qx-Dd-SCI0608-Ziph-060817-125009.wav	-109 dB

Table 3.3. Parameter settings chosen for the proposed algorithm across all test inputs.

Parameter	**Value**
`min_contour_length`	50 ms
`max_contour_inactivity`	25 ms

At the time of this study, there were no publicly available pre-annotated datasets containing tonal signals from non-biological sources. Hence, performance of the proposed detector for non-biological tonal signals is demonstrated with indicative examples. Figure 3.6 and Figure 3.7 show the results of tonal extraction for sounds of anthropogenic and physical origins, respectively. In both examples, the detector has successfully extracted most of the human-discernible tonal signals. Detailed analysis showed that there were few false positives overall. Some of the wrongly reported detections were a result of "tonal hijacking" (e.g. TF contour jump from about 760 Hz to 920 Hz immediately past 4 s in Figure 3.6; TF contour jump from about 185 Hz to 135 Hz around 60 s in Figure 3.7) which will be discussed in more detail in the following section.

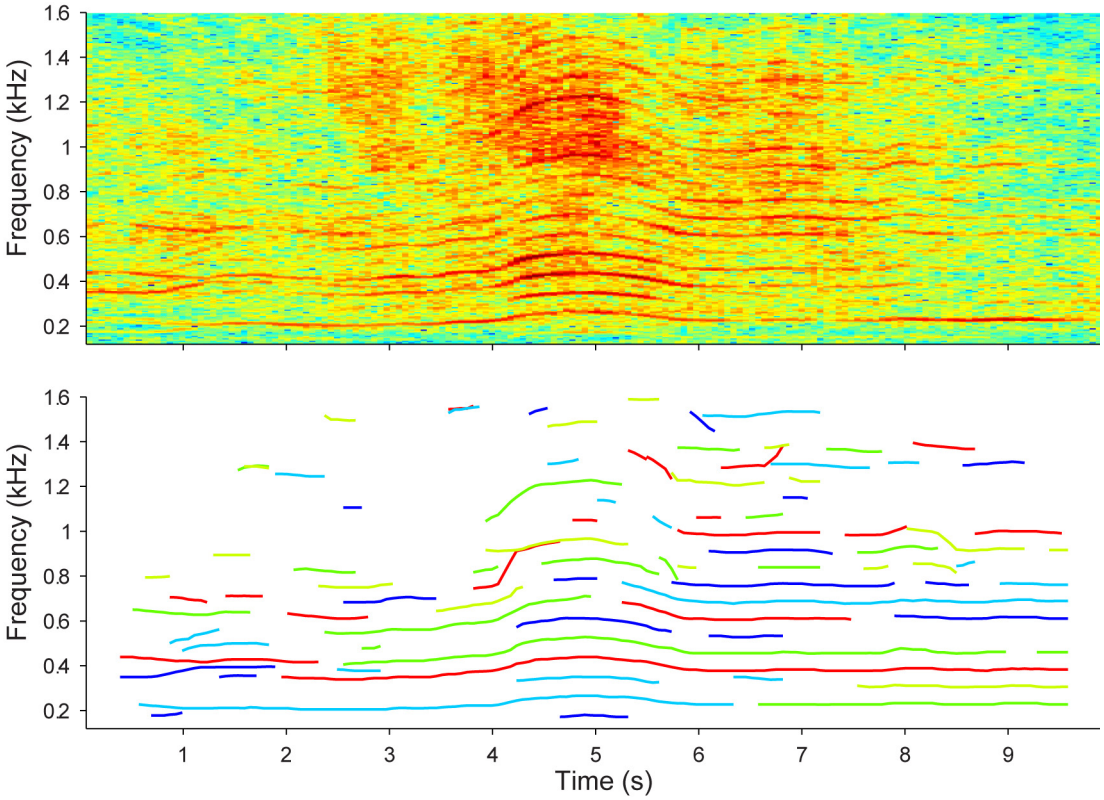

Figure 3.6. Extracted TF contours (bottom panel) from a spectrogram (top panel; FFT parameters: 180 ms Hanning window, 120 ms overlap) of a recording containing tonal sounds of a jetski. Traced contours (or fragments) are coloured differently to ease disambiguation. Algorithm parameters: `min_intensity` = -65 dB, `min_contour_length` = 240 ms, `max_contour_inactivity` = 180 ms.

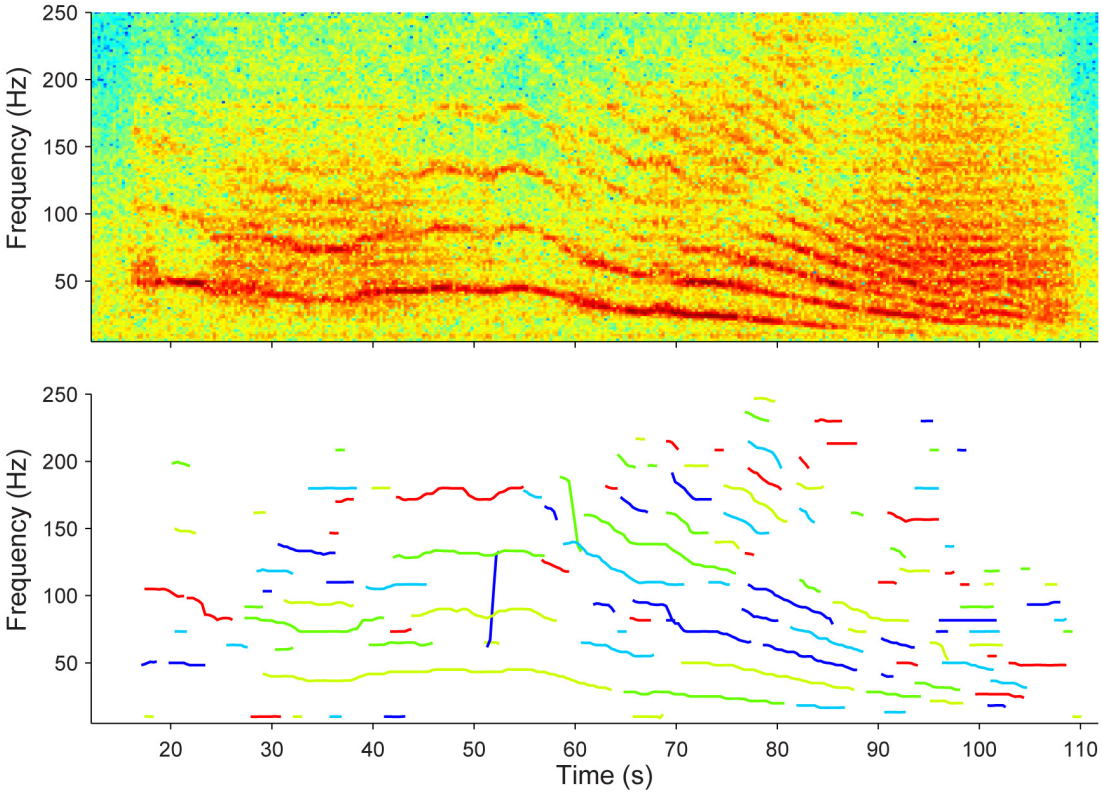

Figure 3.7. Extracted TF contours (bottom panel) from a spectrogram (top panel; FFT parameters: 600 ms Hanning window, 300 ms overlap) of a recording containing Antarctic iceberg harmonic tremors. Traced contours (or fragments) are coloured differently to ease disambiguation. Algorithm parameters: `min_intensity` = -50 dB, `min_contour_length` = 1.275 s, `max_contour_inactivity` = 975 ms.

3.4. Performance analysis and discussion

A method has been presented for the tracing of arbitrary TF contours in spectrograms of underwater acoustic recordings. The core of the proposed system is a 2D ridge-detection component whose outputs feed into a tracking subsystem based on Kalman filtering. The choice of criteria and parameter values of the ridge-detection component were driven by theoretical rationales. Most of the stochastic parameters of the tracing subsystem were obtained using training data while others were chosen heuristically. An implementation of the proposed method was tested using real-world underwater acoustic recordings. Of the different performance metrics considered, recall and precision rates are used predominantly in automatic detector analyses. The

results presented in Table 3.1 show a general improvement in precision and recall over those resulting for one of the existing methods (see Table III in Roch *et al*., 2011b).

Treating TF contours as intensity ridges enables the effective extraction of not just the signals' spectral peaks but also ridge orientation and strength related to TF contours. These features guide the subsequent tracing process. The ability of the system to inherently reject spectral peaks corresponding to non-ridge-like structures in a spectrogram leads to higher precision in overall performance. The benefit of this ability is demonstrated by a noticeable improvement in performance compared to that of the method presented in Roch *et al*., 2011b. For instance, with the file Qx-Dc-CC0411-TAT11-CH2-041114-154040-s.wav which contains sounds from long-beaked common dolphins, improvements of over 350% and 237% were seen in precision and recall rates, respectively, over the results of Roch *et al*., 2011b. A sample clip from the file is shown in Figure 3.8 and the tracing results presented may be compared to those shown in Fig. 9 in Roch *et al*., 2011b.

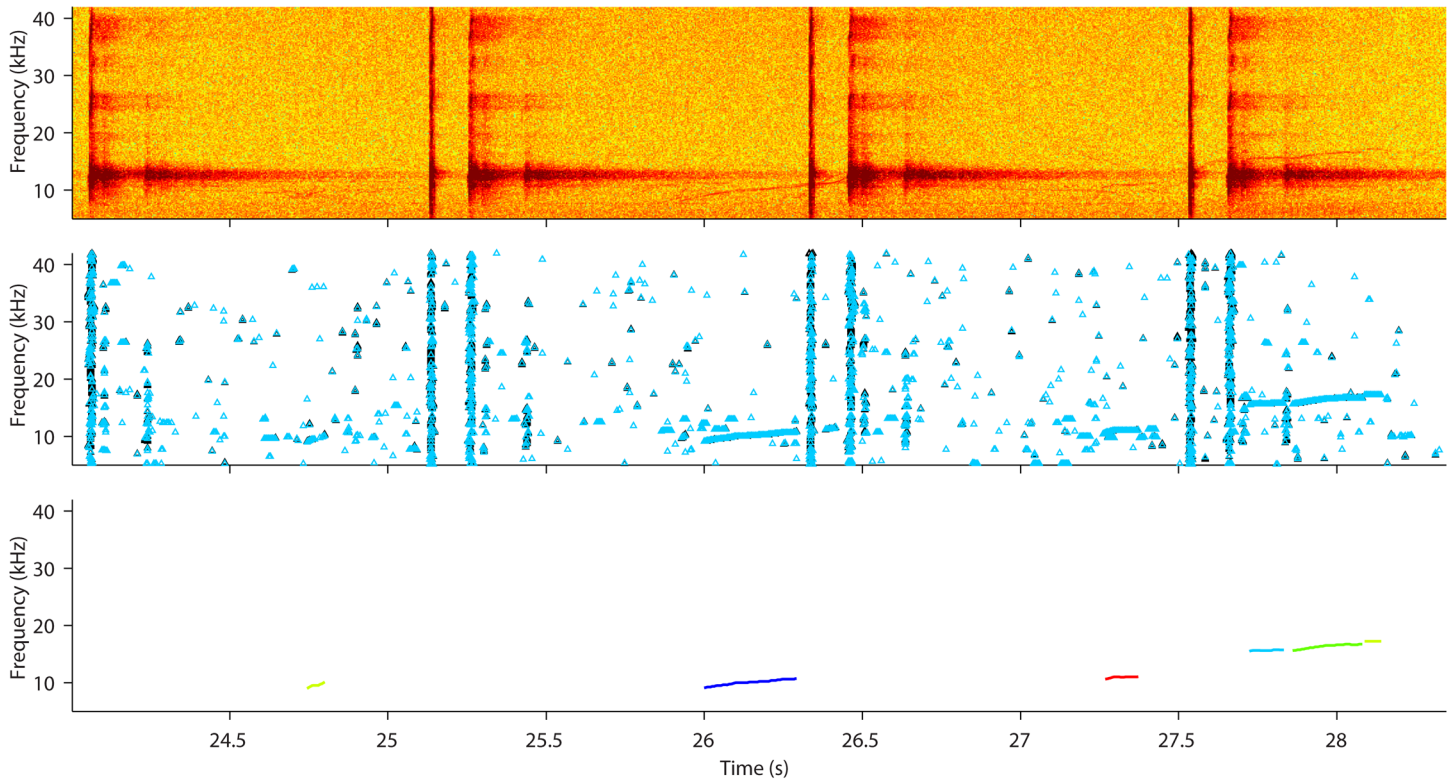

Figure 3.8. Performance of the proposed detector on an audio fragment from file Qx-Dc-CC0411-TAT11-CH2-041114-154040-s.wav. Top panel shows the spectrogram of the audio segment (downsampled to 96 kHz; FFT parameters: 8 ms Hanning window, 4 ms overlap), center panel shows the detected ridge points and the bottom panel shows the traced TF contours. Local maxima in the ridge-point detector outputs in each frame are shown as cyan (light) triangles and all other detected ridge-points are shown as black (dark) triangles. Traced contours (or fragments) are coloured differently to enable easy disambiguation.

The ridge point detection mechanism yields best results with TF contours that are free of clutter from other ridge-like TF structures. At the intersection of two or more ridge-like structures, the mathematical definition of a ridge point as per Eq. (3.5) or Eq. (3.9) may not hold true. Depending on the time and frequency resolutions of a spectrogram, no ridge points may be reported at or near the intersection. Examples of this can be seen in Figure 3.9a. Two TF contours cross each other at 16.54 s and three ridge-like structures (two tonal signals and an echolocation click) occur in close proximity between 16.70 s and 16.77 s. The detection of ridge points is affected for one or more signals in both examples.

The continued tracing of TF contours through temporal discontinuities in detected ridge points is expected to be handled by the mechanism of predicted extensions proposed in Section 3.2.3. The success of such "bridging" relies on the value chosen for the parameter `max_contour_inactivity`. Smaller values may be ineffective with bridging contours' fragments whereas larger values may run the risk of joining successive contours or fragments from different contours. In the example shown in Figure 3.9a, the chosen value of 25 ms for `max_contour_inactivity` suffices in bridging of the first gap occurring between 16.52 s and 16.54 s while the same is ineffective for the second discontinuity occurring between 16.68 s and 16.78 s. Some tonal signals of biological origin exhibit sudden jumps in frequency and are commonly referred to as stepped whistles (e.g. Oswald *et al*., 2003, Roch *et al*., 2011b). Discontinuities in TF contours caused by frequency steps are not expected to be identified or bridged by the proposed algorithm and portions of a whistle on either side of a step are likely to be reported as independent detections. Handling of steps is better suited in a post-processing stage (for example, in a classifier system as indicated in Section 5.2).

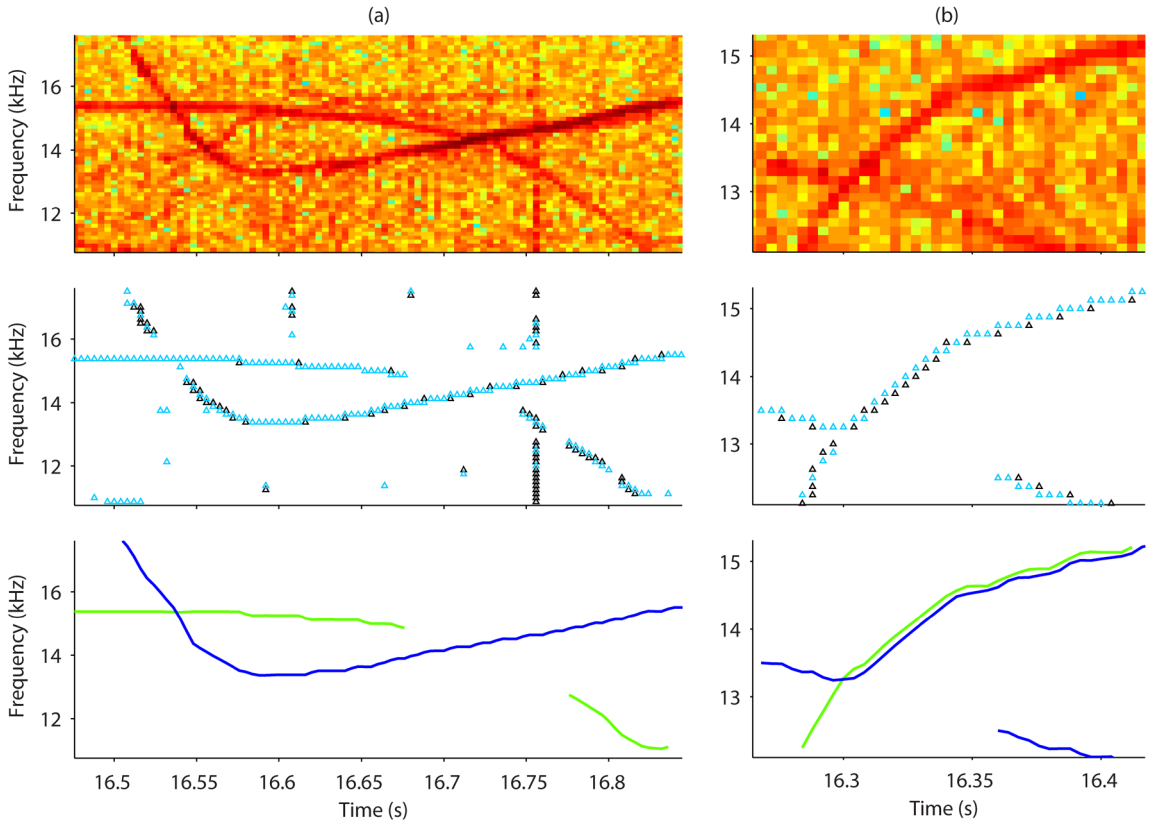

Figure 3.9. Outputs of ridge-point detection (middle row) and ridge tracing (bottom row) operations for spectrograms (top row) of audio clips from palmyra092007FS192-071004-032342.wav.

Sections of TF contours having moderate-to-high $|\psi|$ produce multiple closely-occurring ridge points per frame. When this occurs over a series of frames, it may result in multiple parallel and/or leapfrogging contours to be reported by the tracing component. The example in Figure 3.9b shows a pair of parallel traces beyond 16.3 s. The example is an interesting case deserving closer analysis. Beyond the intersection at 16.3 s, there is a noticeable drop in the downswept tonal's amplitude. Coupled with the ensuing clutter from the upswept tonal, this results in no ridge points being reported immediately past the intersection for the downswept tonal. Both tonal signals exhibit similar f, ρ and τ values prior to the intersection. Past the intersection, the absence of pertinent ridge points and the availability of "favourable" excess points from the upswept tonal signal result in the downswept tonal signal to be wrongly traced along the direction of the upswept tonal signal. Furthermore, over a few frames past the intersection, the ridge points reported for the upswept tonal signal seem more favourable to the wrongly traced contour. In such cases, a wrongly traced contour may "hijack" the available ridge points (past 16.35 s in Figure 3.9b). This may result in both reported contours being incorrect. A fraction of the losses in the reported precision and recall rates are attributed to such duplication and hijacking occurrences. In particular, sections of the spectrogram of the recording in Qx-Dd-SCI0608-Ziph-060817-100219.wav suffer heavy ridge-clutter due to intersecting whistles and echolocation clicks. A sample segment from the recording is shown in Figure 3.10. A pre-processing procedure that removes vertical ridge-like structures, such as the "short-duration transient suppression" method proposed in Mallawaarachchi, 2008, has the potential to improve performance for such recordings.

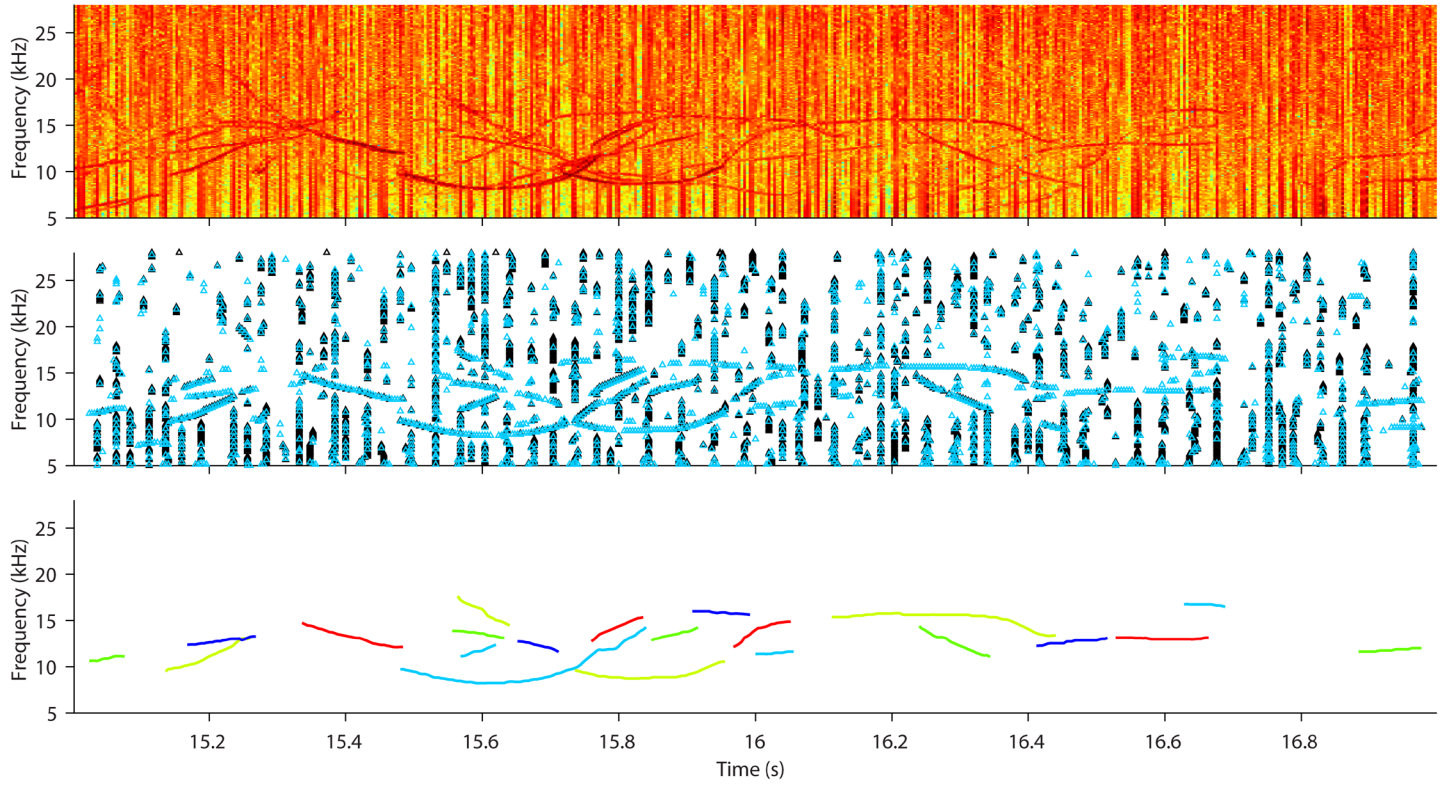

Figure 3.10. An indicative segment of recording from file Qx-Dd-SCI0608-Ziph-060817-100219.wav showing a high level of clutter caused by overlapping whistles and short broadband signals from echolocation clicks. The frequency axis is limited in order to highlight the level of clutter. The top, middle and bottom rows show, respectively, the spectrogram, outputs of ridge-point detector and outputs of ridge tracing component.

Spectral means subtraction is usually employed in existing methods as a pre-processing step for the purpose of noise-suppression. For example, in the methods described in Roch *et al.*, 2011b and Mellinger *et al.*, 2011, spectral subtraction is performed along each frequency bin using running averages obtained over certain predefined durations. The method in Madhusudhana *et al.*, 2008 performs similar spectral subtraction using averages obtained from the extremities of a chosen spectrogram segment. In testing my method, no means-subtraction or normalisation techniques were employed. A direct benefit of this is in the reduction of the system's response latency, thereby making it more suitable for application in an on-site scenario. A more significant benefit is that the system's inertia in adapting to changing ambient noise levels is considerably lowered. The audio in the file palmyra092007FS192-071004-032342.wav (melon-headed whale) suffers from intermittent broadband noise in the 5-25 kHz range due to hydrophone towing (Roch *et al.*, 2011b) and the corresponding results in Table 3.1 highlight my method's robustness applied to dynamic noise conditions.

Processing speed of a detection algorithm is usually quantified by its real-time factor which is defined as the ratio of the duration of the input signal to the time taken to process the input. When tested on a desktop computer with an Intel® i7 CPU and 16 GB of RAM (running Microsoft® Windows 7), a MATLAB based streaming-mode implementation of the proposed approach processed the input band-limited (5-48 kHz) spectrograms with an average real-time factor of 0.14.

Another performance metric commonly considered, particularly in the case of on-site monitoring applications, is a system's response latency. With the choice of using two smoothing kernels, the larger kernel ($\sigma = 2$) results in a spatial filter width of $[2 \times (4 \times \sigma)] + 1 = 17$ spectrogram frames, and a filter delay of 9 frames. The gradient estimation and zero-crossing determination steps add further 5 frames of latency. As suggested, normalisation based on long-duration measurements is not necessary with this approach. Assuming that an implementation of the proposed system is capable of operating at faster-than real-time processing speeds, reporting of the TF contours being traced can occur with a response latency of $14 \times \Delta t$ seconds. For the reported tests, this translates to a response latency of 56 ms.

A spectrogram denoising mechanism that removes vertical ridge-like structures, such as the one proposed in Mallawaarachchi, 2008, can be incorporated into the existing streaming-mode implementation without hindering its robustness or affecting its responsivity adversely. Such signal pre-conditioning has the potential of improving its performance when applied to heavily ridge-cluttered spectrograms. The approach described in Section 3.2.3 for making predicted extensions through short discontinuities in traced contours currently only allows for extensions of constant frequency slopes. The value of ψ can instead be updated iteratively through the discontinuities using successive estimates from fitting polynomial curves to trailing segments of the contour being traced. This could result in a more accurate prediction for signals with temporally varying rates of frequency modulation.

Chapter 4.
Detection of Broadband Signals

A spectrogram-based approach is presented for the detection of broadband signals in underwater audio. The approach employs an iterative 1-dimensional variant of a 2-dimensional multi-scale blob-detection technique commonly used in image processing. In contrast to the referenced 2-dimensional technique subject to an inherent bias for circular features, the iterative 1-dimensional approach enables detection of features of arbitrary bandwidth and duration. The iterative nature (of processing successive frames) makes it an attractive choice for in-situ streaming-mode applications. The algorithm automatically chooses values for several parameters based on the input spectrogram's frequency bounds and hence is capable of being readily employed for a variety of applications. When used with spectrograms, the technique's applications include detection of broadband signals of interest, e.g. Bryde's whale (*Balaenoptera brydei*) calls, underwater earthquakes, explosions. With long-term spectral averages (LTSA), the technique may be used in identifying long-lasting sounds contributing to ambient noise, e.g. fish choruses, sounds of wind and rain. Systematic testing and performance analysis of the method is yet pending and only representative examples of its performance are currently provided.

4.1. Introduction

A variety of methods have been developed for the automatic detection or recognition of broadband signals in underwater audio. In broadband signals, spectral energy is spread over a range of contiguous frequency bands. The appearance of broadband transient signals in their spectrograms depends on the choice of spectral analysis parameters. The detector proposed in this chapter is developed for signals (and spectral analysis settings) that results in broad maxima of spectrogram levels along both frequency and time axes. Underwater sources that produce long-duration broadband sounds include

- anthropogenic sources such as explosions,
- biological sources such as humpback whale barks and bellows, fish choruses, and
- geophysical events such as earthquakes and volcanic eruptions.

Most of the existing approaches are usually targeted for the detection of specific signals. Woodman *et al*. (2004) proposed a simple method for the detection of dynamite fishing events based on comparing sound pressure levels against a pre-determined threshold. Methods based on comparing the ratios of short-term to long-term signal energy averages computed over a series of frequency bands against pre-set thresholds have been used (Hanson *et al*., 2001; Sukhovich *et al*., 2011) for the detection of several types of high-intensity sounds occurring at low frequencies such as the sounds of undersea volcanoes and seismic activities, polar ice calving, etc.

Use of spectrograms for the analysis of acoustic signals is a popular choice as spectrograms provide visual representations of the frequency content of audio signals and its variations over time. Spectrograms are used in the manual analysis of underwater audio in both onsite and offsite monitoring applications. Fourier transforms of broadband signals produce continuous spectra. In spectrograms, long-duration broadband signals appear as bounded 2-dimensional (2D) regions having higher spectral intensities than the regions' immediate surroundings. Such regions are referred to as blobs in this chapter. Broadband signals are differentiable from narrowband tonal signals in spectrograms in that the bounds of tonal signals along the frequency axis are rather narrow (see Figure 1.2).

Several methods are available for the detection of underwater acoustic signals based on post-processing of spectrograms, e.g. Mellinger *et al*., 2011, Madhusudhana *et al*., 2008, Erbe and King, 2008, Lourens, 1990, Fox *et al*., 1995. Some approaches have used image-processing techniques, e.g. Gillespie, 2004; Kershenbaum and Roch, 2013; Thode *et al*., 2012. The method proposed in this chapter is based on ideas derived from an image-processing technique for detecting blobs in digital images. The method described here is meant for the general detection of broadband signals in spectrograms without regard to the sources producing the sounds.

4.1.1. Blob-detection in image-processing

Lindeberg (1993) proposed a method to detect circular blob-like features in images using the Laplacian operator ∇^2. Application of the Laplacian operator to a 2D grey-level image I is described as

$$\nabla^2\left[I_{(x,y)}\right]=\frac{\partial^2 I_{(x,y)}}{\partial X^2}+\frac{\partial^2 I_{(x,y)}}{\partial Y^2}. \tag{4.1}$$

An image I is first convolved with a 2D Gaussian kernel $G(x,y;\sigma)=\left(1/2\pi\sigma^2\right)\exp\left(-(x^2+y^2)/2\sigma^2\right)$ to achieve spatial smoothing at the desired scale σ^2 –

$$L(\sigma)_{(x,y)}=I_{(x,y)}*G(\cdot;\sigma), \tag{4.2}$$

following which the Laplacian operator ∇^2 is applied to the resulting smoothed representation L –

$$\nabla^2 L=\nabla^2\left[I_{(x,y)}*G(\cdot;\sigma)\right]. \tag{4.3}$$

The response of $\nabla^2 L$ is high for circular blobs in I that have their radii approaching $\sigma\sqrt{2}$ (Lindeberg, 1993). Considering convolution associativity and the derivative rule for convolutions, the sequence of operations on the right-hand side of Eq. (4.3) can be rearranged as $\nabla^2\left[G(\cdot;\sigma)\right]*I_{(x,y)}$. The term $\nabla^2\left[G(\cdot;\sigma)\right]$ is commonly referred to as the Laplacian of Gaussian (LoG) operator. The LoG operator is more popular for its use in edge-detection in images and such a technique was first proposed by Marr and Hildreth (1980). Blobs of different sizes are detected by considering scale-space representations at multiple scales. Feature salience across scales is established by choosing the largest of scale-normalised responses ($\sigma^2\times$LoG)*I. An example of a 2D Gaussian kernel and its corresponding LoG is shown in Figure 4.1.

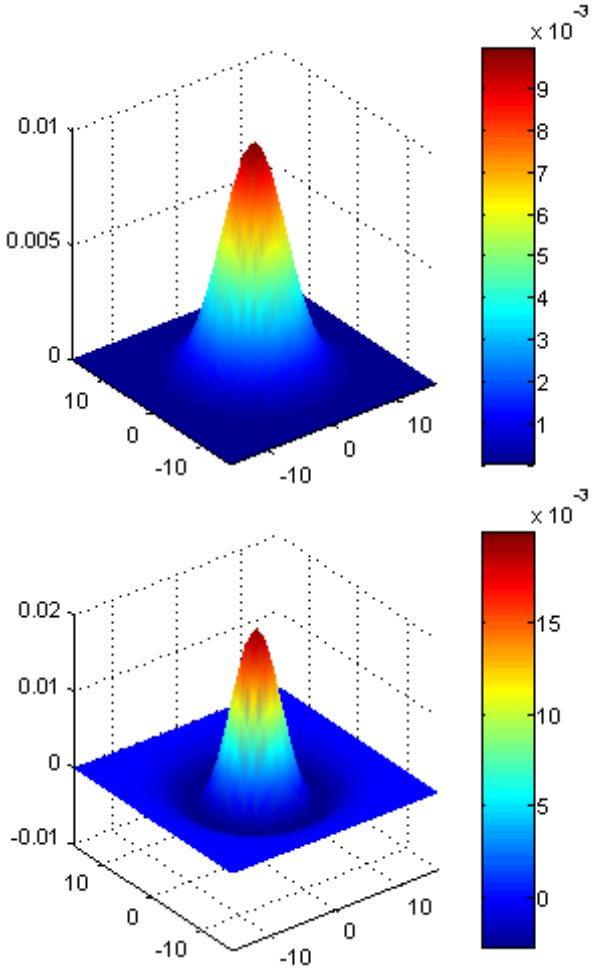

Figure 4.1. A 2D Gaussian kernel (top) and its corresponding scale-normalised LoG (bottom) for a scale of $\sigma = 4$.

4.2. Algorithm

A spectrogram of an acoustic signal is analogous to a 2D greyscale image where spectral intensities at time-frequency (TF) points correspond to grey-level values of pixels in the image. A blob-like feature in a spectrogram is a bound region formed by contiguous TF points having similar intensity levels and occurring with noticeably higher intensities than the TF points surrounding the region. Their bounds are finite along the frequency axis and, for well-formed spectrograms, the bounds are generally contained within the spectrogram's bandwidth. Such an assumption cannot be made about the bounds along the time axis as their temporal extents are defined by the duration of the underlying signal or the variations in its signal-to-noise ratio (SNR).

As blobs in spectrograms are of arbitrary shapes, the method summarised in Section 4.1.1 is not readily suitable for blob-detection in spectrograms. Instead, blob-detection is tackled here by breaking up the single 2D operation into two successive 1D operations. First, broad regions of relatively higher intensities within individual frames of the spectrogram are identified using a multi-scale 1D LoG approach (Section 4.2.2). For convenience, regions of relatively higher intensities spanning a considerable frequency range will be referred to as "plateaus" in this chapter. Intensity plateaus identified in successive frames are subsequently "joined" in tracing the evolution of the blob over time (Section 4.2.3).

4.2.1. Input preparation

Spectrograms present a simple means to visually distinguish continuous broadband signals amongst other sounds. This nature of spectrograms could be easily exploited for the purpose of automatic extraction of broadband signals using methods based on image-processing techniques. The choice of parameters for generating spectrograms is application specific and is beyond the scope of this study. General guidelines are provided here for calculating spectrograms that enable improved performance of the proposed detector. Hamming and Hanning windows are well-suited for the analysis of arbitrary continuous (non-impulsive) sounds (Svend and Herlufsen, 1987) and hence could be favourable choices. For successful application of image-processing based methods for extracting blob-like spectral structures, the time resolution Δt and frequency resolution Δf resulting from the chosen spectrogram parameters must allow for spectral and temporal variations in broadband signals to be discernible in a visual sense. Choosing fine-grained time or frequency resolutions increases the overall processing time and may not always result in improved detection performance. The spectrogram frame widths and Δt must be smaller than the duration of the shortest continuous broadband signal expected to be detected. Having moderate-duration broadband signals span over multiple frames readily allows for the avoidance of short impulsive signals from being detected. The input to the proposed system could be a frequency band-limited portion of the spectrogram or a full-bandwidth spectrogram. The proposed system is generic in its functionality or

control-flow which would remain the same for the processing of any arbitrary spectrogram.

4.2.2. Detection of 1D plateaus

In typical image-processing blob-detection approaches, high-frequency noises in an input image are suppressed by applying 2D Gaussian blurring (spatial smoothing using a Gaussian kernel) prior to application of a Laplacian operator. In the approach proposed here, application of the 1D LoG operators along the frequency axis (Section 4.2.2) inherently achieves spatial smoothing along the vertical dimension. In order to suppress sudden fluctuations in spectral intensity across successive frames, a "horizontal blur" operation is performed by convolving a 1D Gaussian kernel

$$G(x;\sigma) = \left(1/\sigma\sqrt{2\pi}\right)\exp\left(-x^2/2\sigma^2\right) \tag{4.4}$$

along each frequency bin of the spectrogram.

Spectral intensity plateaus in individual frames of a spectrogram are identified using a 1D LoG operator. The Laplacian of a 1D Gaussian function is

$$\nabla^2\left[G(x;\sigma)\right] = \frac{-1}{\sigma^5\sqrt{2\pi}}\left(\sigma^2 - x^2\right)\exp\left(\frac{-x^2}{2\sigma^2}\right). \tag{4.5}$$

Multiplication of the LoG operator by σ^2 is a common practice in image-processing applications. Normalisation by σ^2 renders convolution responses scale-invariant, thereby enabling fair comparisons to be made across scales. One-dimensional Gaussian functions and their corresponding normalised LoG operators, $\nabla^2_{norm}G$, are shown for a few different scales in Figure 4.2.

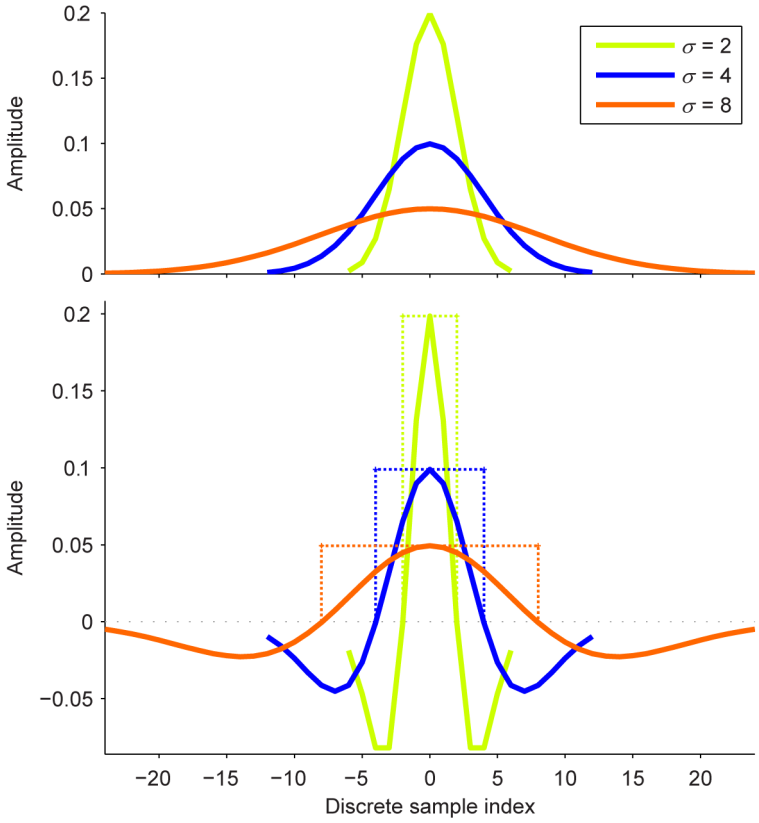

Figure 4.2. One-dimensional Gaussian functions (top) and their corresponding $\nabla^2_{norm}G$ (bottom), shown for three different scales. The dotted rectangles in the bottom plot indicate the optimal widths of high-intensity regions for which the responses of respective LoGs would be maximal. The widths of the Gaussian and LoG curves are restricted here to the range [-3σ, 3σ].

Convolution of a spectrogram frame with an LoG operator (normalised or otherwise) produces high responses for intensity plateaus whose widths are close to 2σ frequency bins. For an ideal 1D plateau such as a boxcar function (von Seggern, 1993) of width 2σ, the apex of the response occurs at its mid-epoch and the response reduces to zero near the edge of the plateau. The behaviour of $\nabla^2_{norm}G$ is demonstrated using synthetic 1D plateaus of different widths and heights in Figure 4.3. As can be seen in the first column of Figure 4.3, the response of $\nabla^2_{norm}G$ $(\sigma = 4)$ is maximum for a plateau width of 8 and it peaks at the centre of the plateau. The disparities between plateau edges and the points where the respective $\nabla^2_{norm}G$ responses reach zero is smallest for the plateau of width 8. As such, the estimates of

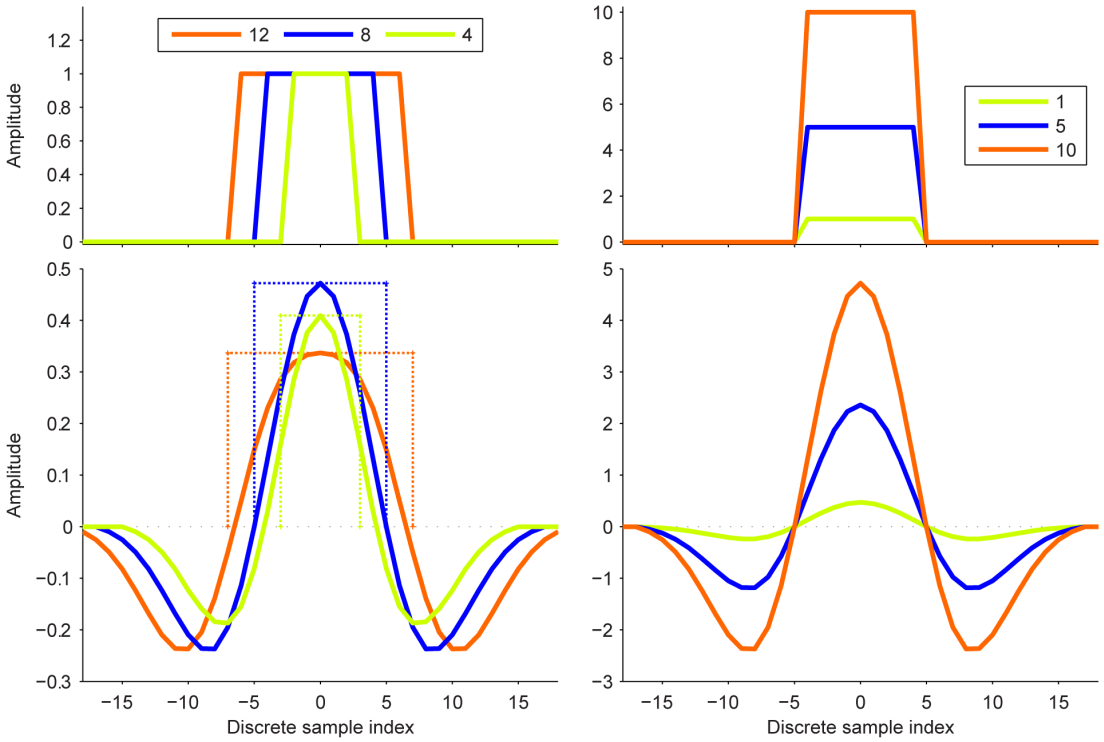

Figure 4.3. Demonstration of $\nabla^2_{norm}G$ responses (bottom row) to a variety of synthetic inputs (top row). The form of the synthetic inputs were chosen to roughly imitate high-intensity 1D plateaus in spectrogram frames. $\nabla^2_{norm}G$ operator with scale $\sigma = 4$ was used throughout. The first column demonstrates $\nabla^2_{norm}G$ responses for inputs of different widths. The dashed lines indicate the outer edges of the inputs. The second column demonstrates $\nabla^2_{norm}G$ responses for inputs of different heights.

plateau widths made from responses of $\nabla^2_{norm}G$ are more accurate when the scale of the LoG operator is chosen appropriately. In such width-matched conditions, the height of a plateau has little or no effect on the width estimates. This is evident from the plots in the second column of Figure 4.3.

In the case of width-matched $\nabla^2_{norm}G$ operators and ideal 1D plateaus as considered in Figure 4.3, the value at the apex of the response is given by $h\lambda$, where h is the height of the plateau and

$$\lambda = \sum_{-\sigma}^{\sigma} \nabla^2_{norm}G\,. \tag{4.6}$$

Scaling the operator by λ^{-1} prior to convolution makes the apex of the response equal the height of the plateau. Such a scaled, normalised operator will be denoted as $\nabla^2_{\lambda-norm}G$.

The absolute heights of synthetic plateaus considered in Figure 4.3 are analogous to the relative elevations of spectral plateaus compared to their immediate neighbourhoods in the frequency axis. This relative elevation is a crude estimate of the underlying signal's SNR. By employing $\nabla^2_{\lambda-norm}G$, per-frame SNR estimates of broadband signals can be obtained directly.

In summary, at a chosen scale, spectral intensity plateaus in a spectrogram frame are said to be detected where the response of $\nabla^2_{\lambda-norm}G$ is positive. The value at a local maximum in the response provides an estimate of the SNR of the underlying signal and the position of the response apex along the frequency axis is an indicator of the position of intensity "centroid" of the spectral plateau.

4.2.2.1. Choice of scales

The frequency bounds of spectrogram blobs differ for different blobs and often vary considerably over the duration of the underlying signal. High-intensity plateaus of different frequency bounds can be captured by employing $\nabla^2_{\lambda-norm}G$ operators at multiple scales σ_n (n = 1, 2, 3, …).

Considering 3σ extents on either side of the underlying Gaussian function's mean, the full width of a discrete $\nabla^2_{\lambda-norm}G$ operator is $[(2 \times 3\sigma) +1] = (6\sigma + 1)$ points. For meaningful convolution with a spectrogram frame, this width must be smaller than the number of frequency bins in the spectrogram. This defines a limit on how large a scale σ can be chosen.

Spectrogram blobs having considerably narrow frequency bounds are generally produced by narrowband tonal signals. Intuitively, the smallest scale $\nabla^2_{\lambda-norm}G$

operator would be chosen such that it would prevent falsely reporting any ridge-like features as blobs. However, choosing the smallest scale to be larger than typical ridge widths does not guarantee that ridges will always be discarded. Instead, the smallest scale is chosen such that ridge-like features are indeed captured, but will be explicitly discarded in a later step (see Section 4.2.2.3). This provides better confidence in ensuring that ridge-like features do not get falsely reported as blobs. The half-power bandwidths or -3 dB bandwidths for common windowing functions are smaller than 2 bins, e.g. 1.54 bins for a Hanning window and 1.30 bins for a Hamming window (Harris, 1978). A value of 2 for the smallest scale σ_1 (width of 4 frequency bins) suffices in capturing most ridge-like spectrographic features for later rejection.

Successive scale values are chosen to form a geometric sequence within the limits considered above. For an input spectrogram having N frequency bins, the range of scales σ_n considered are defined by:

$$\begin{aligned} &\sigma_n = \sigma_1 2^{(n-1)} \\ &n = 1, 2, 3, ..., \lfloor \log_2 [(N-1)/(6\sigma_1)] + 1 \rfloor. \end{aligned} \tag{4.7}$$

Note that N influences not only the value of the largest scale, but also the number of different scales considered. Furthermore, the width of the widest 1D plateau that can be captured with considerable confidence is limited by N as $2 \times \sigma_1 2^{\lfloor \log_2 [(N-1)/6\sigma_1] + 1 \rfloor}$. This limitation can be overcome by simply increasing the number of Fourier transform points considered in computing spectrograms.

4.2.2.2. Automatic selection of salient scales

Spectral plateaus of different widths can be detected by employing multiple $\nabla^2_{\lambda-norm} G$ operators, as per Section 4.2.2.1. However, a single spectral plateau may be detected at different scales. Salience of intensity plateaus detected at different scales is established based on $\nabla^2_{\lambda-norm} G$ responses forming local maxima across successive scales, i.e.

$$\frac{\partial\left(\nabla^2_{\lambda-norm} G * S\right)}{\partial\sigma} = 0$$
$$\frac{\partial^2\left(\nabla^2_{\lambda-norm} G * S\right)}{\partial\sigma^2} < 0 \ , \tag{4.8}$$

where S is a spectrogram frame. In the 2D grid resulting from the tessellation of operator responses at successive scales, local maxima are identified as points where the response values are larger than or equal to the values in their 9 neighbouring points. Figure 4.4 illustrates an example of the scale selection described here. The 2D grid is shown in the second row plot. The response apices occurring near 17.5 Hz and 26 Hz for the scales σ = 8 and σ = 4, respectively, are discarded as they are not

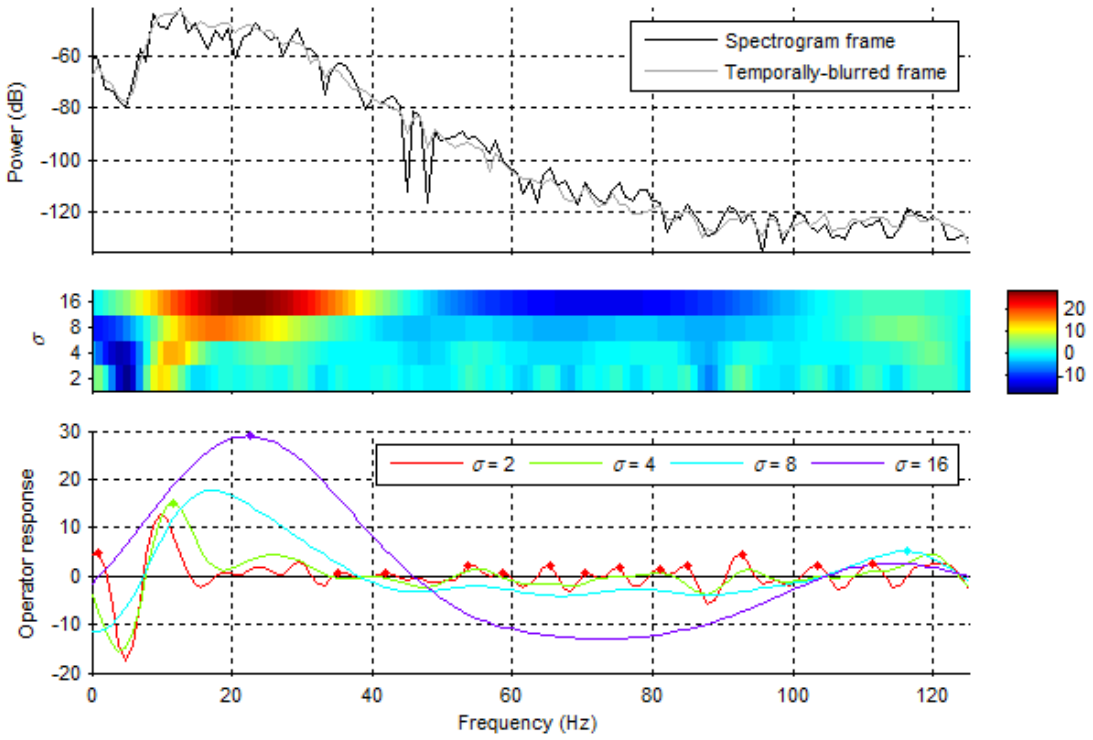

Figure 4.4. A demonstration of scale selection presented using a frame (top row) from a spectrogram (see Figure 4.7) of real underwater audio. The high intensity region between 5 Hz and 45 Hz corresponds to sound from an underwater earthquake recorded at a long distance. The middle and bottom rows show, in two different formats, the operator responses at various scales considered. Each point (f, σ) in the middle row plot indicates the operator response at the scale σ for frequency bin f. The local maxima over scales are indicated by diamond-shaped markers in the bottom row plot.

maximal points across scales. The response corresponding to the widest high-intensity region in the spectrogram forms a local maximum at around 23.5 Hz at and in scale $\sigma = 16$. Such a 2D local maxima gets considered for further processing.

4.2.2.3. Trimming & winnowing

Spectral plateaus detected at the finest scale will not be considered for further processing. Discarding these detections only after examining the maxima across scales ensures that small plateaus do not get detected at coarse scales.

Sometimes, it may be appropriate for certain analyses to ignore weak broadband signals having relative levels lower than some pre-defined value. A user-controllable quantity `SNR_threshold` [6] is defined for this purpose and any detected plateaus are discarded by the algorithm when their corresponding operator response apices are lower than `SNR_threshold`.

The estimates of detected plateau edges may not always be valid. The presence of other signals of high relative levels in the neighbourhood can result in additional local maxima before the operator response reaches zero. In such cases, the local minimum between two neighbouring response apices presents a good separation point for the corresponding plateaus. In general, the algorithm defines the frequency bounds of a detected plateau by searching on either side of its response apex for either a valley point or a zero-crossing, whichever occurs first.

Detection of salient plateaus and their edges at different scales is demonstrated in Figure 4.5. Notice that the ridge-like feature occurring at 60 Hz caused by a narrowband tonal signal is captured at the finest scale. As discussed above, these detections are not considered for further processing. The remaining blob-like features are captured at appropriate scales. The horizontal blurring of the spectrogram prior to applying $\nabla^2_{\lambda-norm} G$ suppresses some of the short-duration (transient) noises such as those occurring at higher frequencies around 78 s, 82 s and 116 s.

[6] User-settable algorithm parameters will be denoted using monospaced font.

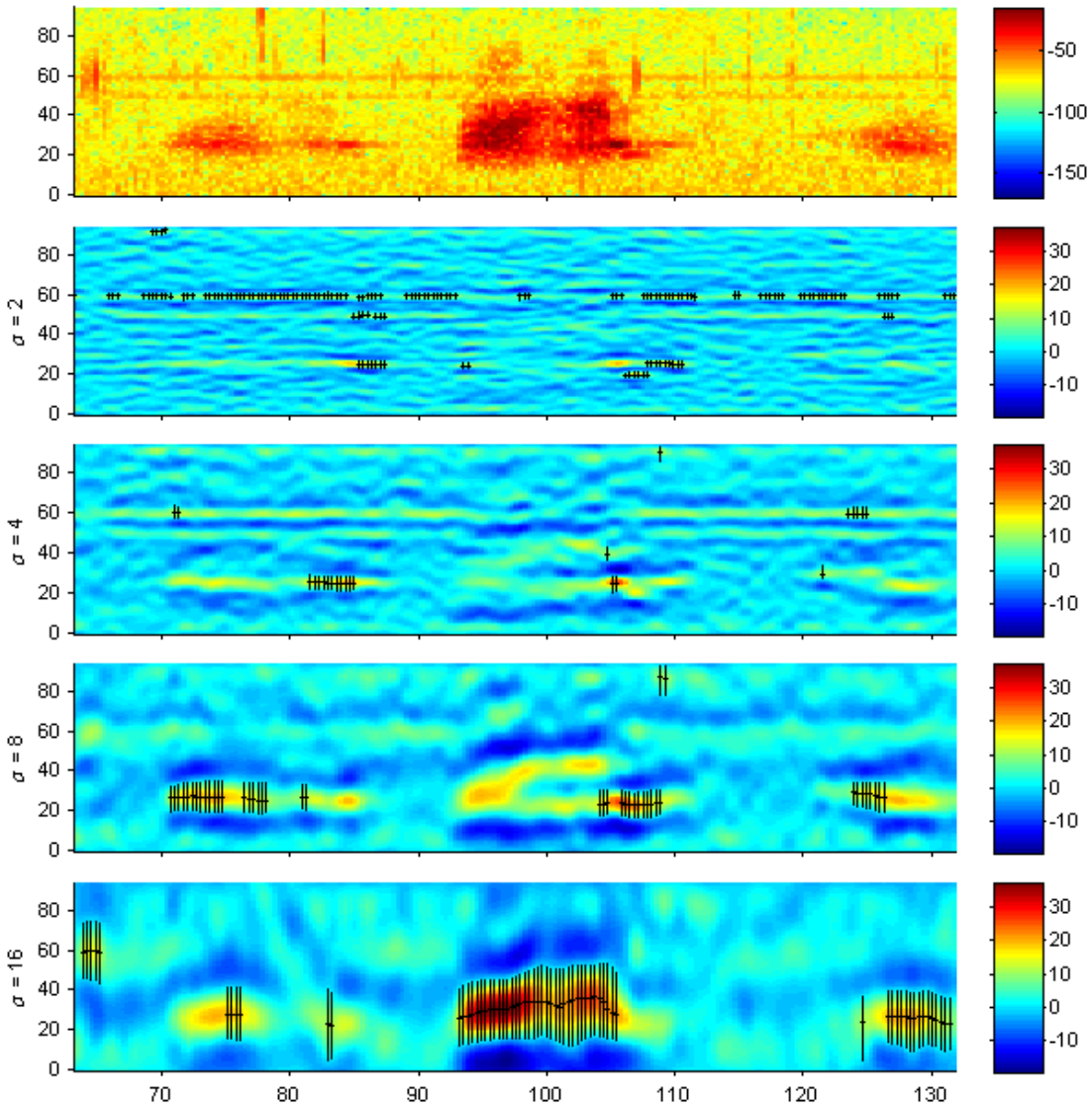

Figure 4.5. Multi-scale detection of intensity plateaus demonstrated using a segment of a spectrogram produced from an underwater recording containing Bryde's whale calls and other noises. The top row shows the spectrogram and the following rows show the operator responses at four different scales. In all of the plots, frequency (in Hz) and time (in seconds) are respectively shown along the vertical and horizontal axes. Detection was performed with `SNR_threshold` value set at 10 dB. The bounds of the detected 1D plateaus are shown with overlaid vertical lines and the corresponding response apices are indicated with a marker along the lines.

4.2.3. Tracing temporal evolution of blobs

One-dimensional intensity plateaus detected in each frame are "joined" with those in neighbouring frames to trace the 2D blobs representing the underlying broadband signal. The "joining" is performed iteratively – processing the spectrogram frame after frame along increasing time and extending traced blobs using detected 1D plateaus available in succeeding frames. At any frame, a spectrogram blob being traced is referred to as an active blob if it had an extension at the previous frame. A candidate 1D plateau is assigned to an active blob based on its similarity to the blob's frontier which is the 1D plateau that had extended the blob at previous frame. The similarities considered are –

- value of the response apex (SNR estimate),
- position of the response apex along the frequency axis (intensity centroid), and
- width of the plateau.

At any frame there may be multiple active blobs and multiple candidate extensions available. For each of the pairwise combinations of active blobs and candidate 1D plateaus, assignment costs are determined based on the 3-dimensional Mahalanobis distance (Mahalanobis, 1936) quantifying the aforementioned similarities. Possible dubious pairings are heuristically discarded when the edges of the 1D plateaus exhibit no overlap or when the centroid of the candidate does not occur within the edges of the blob frontier. Of the pairings that remain, a minimalistic subset is chosen such that a one-to-one mapping of blobs to 1D plateaus exists in the subset and that the resulting sum of the assignment costs of the pairings in the subset is a minimum. This assignment problem is solved using the Munkres algorithm (Munkres, 1957), which finds global-minimum-cost assignments with a computational complexity of $O(n^3)$.

Tracing of an active blob ceases at a frame when no suitable extensions are available in the succeeding frame. Blobs whose tracing have ceased are reported if their duration (temporal bounds) is longer than `minimum_duration` and are otherwise discarded. Detected 1D plateaus that were not used for extending an active blob are

used for starting a new blob trace. An example of the outcomes of the tracing process is shown in Figure 4.6.

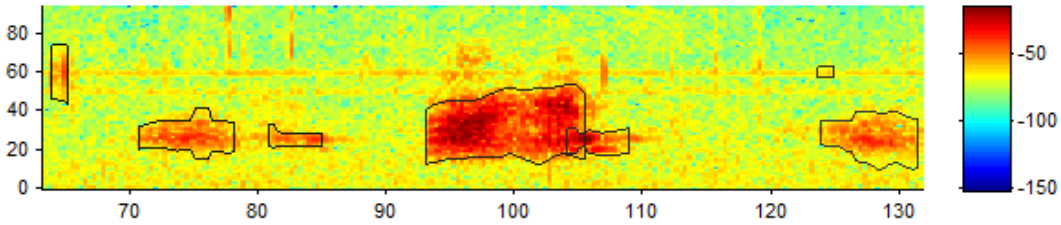

Figure 4.6. Outcomes of the described blob tracing process shown with overlaid black curves on the spectrogram considered in the example of Figure 4.5. The value of the parameter `minimum_duration` was set to 500 ms. The vertical and horizontal axes show frequency (in Hz) and time (in seconds), respectively.

4.3. Analysis

Systematic testing of the algorithm proposed in this chapter requires several instances of verifiable hand-annotated inputs. At this moment, I have not been able to procure the same. As such, a thorough performance analysis of the algorithm is yet to be conducted. At the current stage, the detection performance of the algorithm is demonstrated only using a few examples. Figure 4.7 shows an example of the detector outcome for a recording corresponding to an underwater quake. Figure 4.8 shows a few more examples using Bryde's whale calls. It can be seen from the example shown in Figure 4.5 and from the examples shown in this section the number of false positives can be kept low by using well-formed spectrograms and choosing well-defined values for the algorithm settings. The utility of the algorithm in detecting long-lasting acoustic events from LTSAs is demonstrated in Figure 4.9. Instances of fish choruses, lasting about 2.5 hours, were successfully detected. Detection of the presence of humpback whales (from the detection at mid-frequencies on day 5), possibly others cetaceans (at lower frequencies) and shipping activities are also made possible.

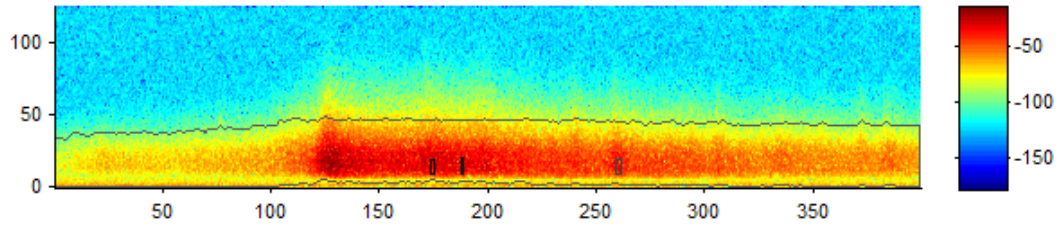

Figure 4.7. Demonstration of blob detection in spectrograms using an example of a high-intensity low-frequency long-duration sound caused by an underwater earthquake. The vertical and horizontal axes show frequency (in Hz) and time (in seconds), respectively.

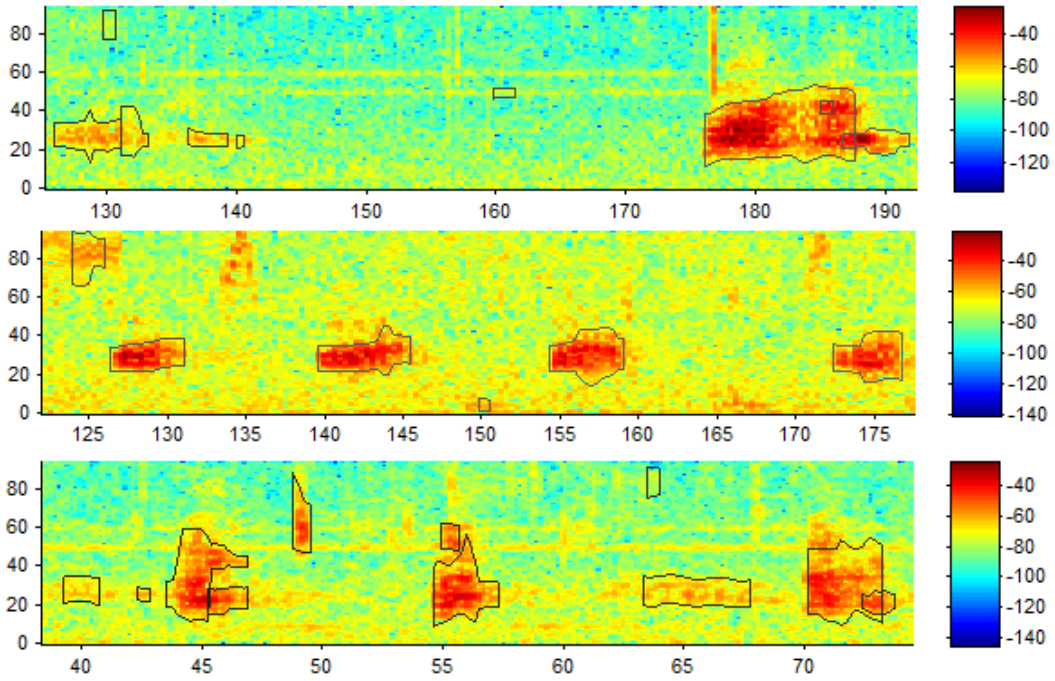

Figure 4.8. Demonstration of blob detection in spectrograms using examples of Bryde's whale calls amidst other noises. The vertical and horizontal axes show frequency (in Hz) and time (in seconds), respectively.

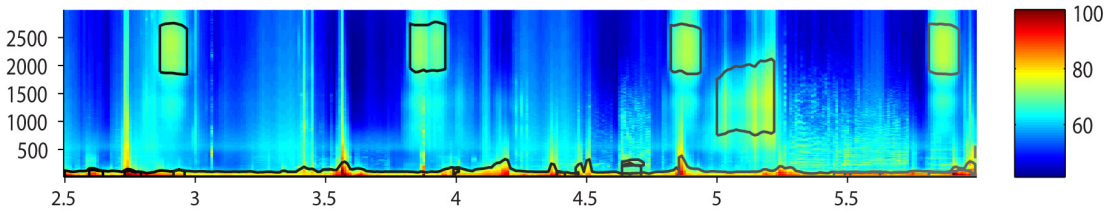

Figure 4.9. Demonstration of blob detection in LTSAs using a multi-day segment as an example. The vertical axis shows frequency (in Hz), the horizontal axis shows time (in fractional days, starting at midnight) and the color levels indicate average spectral power density (in dB re 1μPa^2/Hz). Each frame or time slice in the LTSA represents 900 s. The value of the parameter `minimum_duration` was accordingly set to 4500 s. The LTSA shows, among other events, fish choruses occuring following sunset every day, highly vocal presence of humpback whales early on day 5 and several passing ships (at values of 2.7, 3.55, 4.2 and 4.8 on the horizontal axis).

4.4. Discussion

A generic approach for detecting broadband acoustic activity is presented. Demonstration of the algorithm's detection performance is presented using indicative examples of biological, anthropogenic and geophysical sounds. The approach is shown to yield reasonable results with both spectrograms and LTSAs. A systematic performance analysis of the algorithm using a substantial set of test data covering varieties of noisy conditions is pending.

In the proposed algorithm, no long-duration spectral means-subtraction or normalisation techniques were employed. A direct benefit of this is that the system's inertia in adapting to changing ambient noise levels is considerably lowered. Not having to wait for large amounts of data to be gathered before processing them makes the algorithm more suitable for application in an on-site scenario. The time taken by successive per-frame 1D convolution operations remains constant over the entire spectrogram and the convolutions usually execute faster than real-time on modest computers. The tracing sub-process described in Section 4.2.3 is the only

aspect of the algorithm whose computational complexity is tied to the number of putative blobs in each spectrogram frame. Given that its computational complexity is polynomial and that the number of underlying mathematical computations is small, the tracing process executes faster than real-time as well. A performance metric commonly considered, particularly in the case of on-site monitoring applications, is a system's response latency. Seeing that all operations in the algorithm can be performed at faster-than-real-time speeds, the response latency of the algorithm is defined by the group delay resulting from convolution with the horizontal blur filter, i.e. its half width or $\lceil 3\sigma \rceil \times \Delta t$ seconds.

Systematic testing of the algorithm would most likely warrant several improvements to be considered. Some such considerations are provided here. Currently, tracing of blobs from successive frames ceases immediately at the frame where no suitable extensions can be found. This may result in fragmented detections to be reported. Additional functionality to bridge such fragmented detections can be implemented by considering a small limiting factor on the temporal separation. In the example shown in Figure 4.7, the per-frame peak energy varies considerably (by about 50 dB) over the full duration of the detected blob. Permitting such large variations could be detrimental in some situations, especially for spectrograms of recordings that have high levels of ambient noise. The temporal extent of a detected blob corresponding to a time-limited event could get falsely stretched indefinitely by the background levels. In the example of Figure 4.7, the detection could possibly have been reported to start at around 100 s instead. In the tracing component of the algorithm, per-frame peak (or median) energy levels (within an active blob's frequency bounds) can be considered and limitations can be imposed on the amount of change allowed across successive frames.

Chapter 5.
Future Work - Automatic Soundscape Characterisation

Developing a unified solution for the automatic classification of all underwater sounds may seem a near-impractical undertaking. The solution becomes more tractable with the suggested breakdown of the automatic recognition problem into a two-phase detection-and-classification approach as was shown in Figure 1.1. Based on this idea, a framework is presented in this chapter for realising an automatic soundscape characterisation system using the detectors presented in the preceding chapters.

5.1. Considerations for realising an automatic characterisation system

5.1.1. Application-specific operating conditions

An implementation of an automatic soundscape characterisation system must consider the operating conditions specific to a particular application. The operating conditions of an application include the variety of sounds occurring in the environment being monitored and the limitations imposed by the PAM hardware employed. Soundscapes pertaining to different environments have differing characteristics. For example, consider an estuary marked for biota conservation where recreational and commercial vessel activity may be banned. Monitoring of soundscape in such an environment would not require, within the automatic system, components meant for recognising vessel noise. The operational efficiency of the automatic characterisation of a soundscape may be improved based on such *a priori* knowledge about the soundscape. Similar examples can be stated for near-shore and deep-ocean environments. Also, different PAM hardware configurations result in different operating conditions. For example, a recording made at 24 kHz sampling rate would not contain a majority of the odontocete echolocation clicks and hence,

the automatic system may not need to execute the components pertinent to identifying those sounds.

5.1.2. Context of a sound

Automatic classification of sounds becomes more feasible with the consideration of their *context*, particularly in cases where available information about signal's spectro-temporal characteristics do not suffice in decision-making. The notion of the use of a sound's context towards automatic recognition has been garnering interest in the marine PAM community. The fourth international workshop on Detection, Classification, Localization and Density Estimation (DCLDE) of marine mammals held at Pavia, Italy in 2009 was focused on the use of contextual information in automatic recognition. Researchers from JASCO Applied Sciences (http:// http://www.jasco.com/) presented their study titled "Improving the performance of marine mammal call classifiers using contextual information" at the sixth DCLDE workshop at St. Andrews, UK in 2013. There was a wider expression of acknowledgement of the need for using contextual information during the discussion session headed by Dr. Marie Roch of Scripps Institution of Oceanography (http://sio.ucsd.edu/) at the succeeding DCLDE workshop held in San Diego, USA in 2015. There is yet no widely agreed-upon understanding of what constitutes a sound's context. It could include information such as visual sighting records of the species producing the received sounds, weather data, seismic activity records, information on the sound propagation environment, etc. Where such non-acoustic contextual data are unavailable, classification may rely on acoustic context that is inherently available in the input audio. The acoustic context of a sound accounts for the acoustic activity within its immediate temporal and spectral neighbourhoods. The examples below should help with understanding the definition of context as it applies to classification.

Without the explicit use of the term "context" in the literature, a few prior studies have already benefited from the use of contexts. Sirovic *et al*. (2013) show that the detection of multiple call types within a certain temporal neighbourhood provides a more certain estimate of a species' presence. Oswald *et al*. (2003) speculate

achieving improved performance in their automatic classifier with the inclusion of species' spatial distribution probabilities (non-acoustic context) in its classification models. Whistles of many marine mammal species are known to contain one or more harmonics. Along with several parameters of the fundamental frequency contour, the Real-time Odontocete Call Classification Algorithm (ROCCA) described in Oswald *et al*., 2007 uses information about the presence or absence of call harmonics to aid in the classification of delphinid whistles. Note that although harmonics are, in fact, components of an odontocete whistle, their consideration in ROCCA as supplementary information indicates that they may be viewed as being a part of a signal's spectro-temporal acoustic context. Following the detection of individual echolocation clicks, some studies have used, as temporal acoustic context, information from click trains (more specifically, the interval between successive clicks) (Harland, 2008; Gerard *et al*., 2009) and information about delayed surface reflections (Zimmer and Pavan, 2008) towards classification of the detected clicks. Application-specific contextual information considered in classification tasks may include other information such as time of day, prevailing season, weather data, *a priori* information on seasonality of species' presence and migratory patterns, proximity of recording equipment to shipping lanes, Automatic Identification System (AIS) data on vessel activity, etc.

5.2. System design

A modular implementation framework is presented here for realising an automatic soundscape characterisation system. The proposed framework is to be used only as guidelines and a particular implementation of the suggested framework needs to consider the requirements of the specific application (see Section 5.1.1) for good operational efficiency. Successful first-step detection of signals is vital for subsequent meaningful characterisation. Most signals that are distinguishable from background noise are expected to be detected by one of the three detectors presented in the preceding chapters. A composite module comprising of the three detectors could hence form the first phase of any implementation of an automatic characterisation system. With signal detection handled in a context-free manner, the

considerations for operational efficiency indicated in Section 5.1.1 must be dealt with in the classifier phase. A schematic of the proposed framework is shown in Figure 5.1. Implementation of different modules making up the classifier will be application-specific and such potential modules are not detailed in this study.

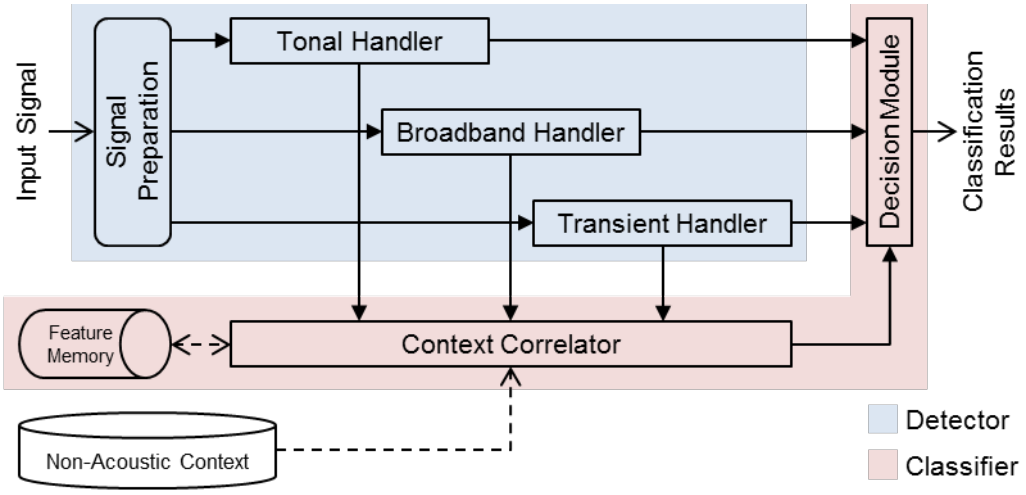

Figure 5.1. Proposed framework of an automatic soundscape characterisation system.

The three proposed detectors are part of the corresponding modular "handlers" shown in the framework. Following signal detection by the contained detectors, the handlers identify and pick out detected signals' salient distinctive characteristics, which are called "features" in this chapter. For example, the Tonal Handler delivers a series of contours. Certain features can be measured off such contours, e.g., minimum frequency, maximum frequency, start frequency, end frequency, locations of local extrema, locations of inflection points, presence or absence of contour discontinuities, presence and frequency spacing of overtones, etc. These features can be combined into a feature vector that is used by the classifier. The Transient Handler could extract features such as peak frequency/frequency band, presence of FM, FM rate, transient duration, peak-to-peak amplitude, etc. The Broadband Handler could extract features such as the spectral and temporal extents of detected blobs, peak and average energy within blob extents, frequency at the peak energy, etc. The Signal Preparation module is included in the schematic to indicate the potential need for pre-processing input signal appropriately for the different

detectors. Certain signal pre-processing operations may be performed inherently within the handlers. This is discussed further in Section 5.3.

As indicated in Section 5.1.2, the classification of certain sounds is made easier when the various extracted features of detected signals are augmented with contextual information. This is enabled by the Context Correlator. Non-acoustic contextual information, when available, is generally known beforehand and can be made available to the Context Correlator in stored form. Acoustic context of a detected signal comprises of other signals detected within its spectro-temporal neighbourhood. The bounds of considered neighbourhoods vary depending on the type of detected signal in consideration. For example, an individual echolocation click detected by the Transient Handler can be associated to an ongoing click-train by considering other detected clicks occurring within a short time window; individual TF contours reported by the Tonal Handler could be associated as harmonic components of a tonal signal if they occurred within a certain frequency bandwidth of each other and are contained within similar temporal extents. As input audio is continuously processed, detected signals' features are retained temporarily in the Feature Memory to enable establishing of acoustic contexts for other signals. For each detected signal, the Context Correlator identifies the necessary context and supplies the corresponding features and any pertinent non-acoustic contextual information to the Decision Module. Over time, some features retained in the Feature Memory become impertinent for any future contexts depending on the bounds of temporal neighbourhoods considered for the different signal varieties. The Context Correlator is also responsible for the discarding of such stored features that may no longer be useful.

The Decision Module brings about the final classification of the detected signals. It considers the extracted features of the detected signals and any contextual information, whenever available, in its decision-making process. The identification of features by the different handlers could sometimes suffice for making direct inferences about the type of source without needing contextual information. Several choices are available for the implementation of the Decision Module. For example, it may be implemented as a rule-based expert system utilising rules derived from

human interpretations of the detected signal features, or it may be implemented similar to the multivariate discriminant function analysis (DFA) or classification and regression tree (CART) analysis as shown in Oswald *et al*. (2003). The number of rules, constraints or parameters, whichever applicable, considered for an implementation may be restricted based on the operational requirements of a particular application.

5.3. Notes on detector configurations and signal preparation

In order to avoid a sound unit from being multiply classified, detector parameters need to be chosen such that a single sound is not detected by multiple handlers. Some considerations are provided here. Short impulsive signals such as echolocation clicks, which are detected by the Transient Handler, appear as broadband signals in a spectrogram. Detection of such signals by the Broadband Handler can be avoided by setting the minimum duration for acceptable signals to cover multiple frames in input spectrograms. The provision of multiple spatial smoothing scales in the Tonal Handler and 1D smoothing scales in the Broadband Handler provide flexibility in detecting signals of various spectral bandwidths. The largest scale in the Tonal Handler and the smallest scale in the Broadband Handler must be chosen such that spectrogram frames containing continuous signals do not trigger detections in both handlers.

Both the Tonal Handler and the Broadband Handler handle spectrograms as inputs. An implementation of the framework may choose to feed same spectrograms to both handlers or different spectrograms may be computed (using different sets of parameters) from the input audio for each handler. The frequency band of frequency modulated signals is governed in general by the quality factor (a measure of sharpness of spectral peaks, defined as the ratio of the central frequency to the width of frequency band at -3 dB level) of underwater sources of sound, which varies within a limited range for most sources of biological and physical origin. Consequently, the rate of frequency modulation in the absolute units of Hz/s is typically higher for high-frequency sources than that of low-frequency sounds. This

means that *time* × *frequency* contours of high-frequency tonal sounds are traced better in spectrograms with a shorter time window of spectral analysis. So, when input signals have a wide recording bandwidth, it is beneficial to compute multiple spectrograms with different width of the FFT window for different frequency ranges. Such a frequency-range restricted portion of a spectrogram will be referred to with the term *segment*. When an implementation is chosen to process multiple segments, the respective handlers would clone their operations for each segment. With the available range of frequencies for any recording and the operational requirement in consideration, one may either wish to extract the tonal and broadband signals within only one frequency range of interest or wish to extract the contours of all tonal signals contained in the recording. In the latter scenario, it is up to the user to carefully select segment boundaries (with or without overlap). The choices of the boundaries are influenced by three considerations:

- The need to maximize the efficacy of a set of values chosen for the spectrogram parameters imposes limitations on the bandwidth of the segment.
- On the contrary, excessive segmentation leads to undesirable increase in overall processing time.
- For effective contour tracing, the possibility of TF contours and broadband signals cutting across segment boundaries has to be minimized.

5.4. Conclusion

A generic framework is proposed for realising an automatic soundscape characterisation system as a two-phase detection-and-classification system using the independent detectors presented in this study. The role of an application's operating conditions in specific implementations of the framework is illustrated. The notion of *context* of a signal is introduced and its importance in developing classifiers is described. Finally, guidelines are provided for preparation of inputs to the detectors and on their configurations within the overall system.

Chapter 6.
Summary

The task of automatic detection of "signals of interest" in underwater audio was split into three components based on the signals' spectro-temporal characteristics. Such splitting of the detection process enables pertinent characteristics of detected signals to be readily extracted, e.g. temporal extents and energy levels of impulsive signals and amplitude and frequency modulation rates of tonal signals. This capability is conducive for the development of the signal "handlers" proposed in Chapter 5 as parts of an automatic characterisation system.

Detailed performance analyses of the transient and tonal detectors were provided in the respective chapters. The performances were also compared against available approaches in the literature. Detailed analysis of the broadband signal detector is currently pending and only a brief analysis of its performance is provided. The limitations of the detectors and considerations for future development were also discussed in the respective chapters. In this chapter, general characteristics of the detectors are highlighted in view of the expectations outlined in Section 1.2. Note that, due to lack of evidence, not all characteristics discussed below are relevant to the broadband signal detector.

The transient detector was tested using both synthetic and real audio; and the other two detectors were tested using real audio alone. For testing with real underwater audio, the transient and tonal detectors were tested using annotated data from the MobySound database. The recordings in the data used for testing were gathered at different locations by different organisations. The recording equipment employed in procuring the various recordings also differed. The consistency exhibited in detection performance across the different data sets demonstrates that the "flexibility" and "adaptability" requirements were successfully met. "Robustness" of the transient signal detector was demonstrated using both real and synthetic data where synthetic data were regenerated to produce different SNR conditions; and the robustness of the

tonal detector was evident from the provided testing results, and examples of the same were demonstrated using indicative segments from the test set.

No long-duration signal measurements (e.g. mean energy, background level estimates, etc.) were considered for pre-conditioning detector inputs. So, the detectors need not gather chunks of data before beginning processing. The response latencies of the three detectors are finite and constant. This makes them well-suited for on-site real-time applications, both as independent detectors and as part of the categorisation system proposed in Chapter 5.

Summarising the above discussion, automatic detectors of various types of underwater acoustic signals were developed and the desired performance characteristics (robustness, flexibility, adaptability and operational efficiency) were achieved. Together with the simplicity and ease of configuration, these standalone detectors can be readily redeployed for different purposes and hence, they provide an attractive choice for uptake in targeted recognition applications as well. General guidelines are provided in Chapter 5 for realising a comprehensive automatic soundscape characterisation system using the detectors presented in this study.

List of Abbreviations

AIS	Automatic Identification System
AM	Amplitude Modulation/Modulated
CART	Classification And Regression Tree (analysis)
CFCW	Constant Frequency Carrier Wave
CTBT	Comprehensive nuclear-Test-Ban Treaty
DCLDE	Detection, Classification, Localization and Density Estimation
DFA	Discriminant Function Analysis
ERMA	Energy Ratio Mapping Algorithm
FDR	Filter Difference Ratio
FM	Frequency Modulation/Modulated
GBDT	Gradient Boosted Decision Trees
ICI	Inter-Click Interval
IMS	International Monitoring System
IPI	Inter-Pulse Interval
LCCW	Linearly Chirped Carrier Wave
LM	Levenberg-Marquardt (algorithm)
LTI	Linear Time Invariant
LTSA	Long-Term Spectral Average
MAF	Moving Average Filter
PAM	Passive Acoustic Monitoring
PR	Precision-Recall (trade-off)
ROCCA	Real-time Odontocete Call Classification Algorithm
SNR	Signal-to-Noise Ratio
TKEO	Teager-Kaiser Energy Operator

References

Adam, O. (**2008**). Segmentation of killer whale vocalizations using the Hilbert-Huang transform. *EURASIP Journal on Advances in Signal Processing*, *2008*, **162**.

Adam, O., Lopatka, M., Laplanche, C., & Motsch, J. F. (**2005**). Sperm Whale Signal Analysis: Comparison Using the AutoRegressive Model and the Daubechies 15 Wavelets Transform. *WEC (2)*, *2005*, 188-195.

Andrew, R.K., B.M. Howe, and J.A. Mercer. (**2011**). Long-time trends in ship traffic noise for four sites off the North American West Coast. *Journal of the Acoustical Society of America* **129**(2):642-651.

Andrew, R. K., Howe, B. M., Mercer, J. A., & Dzieciuch, M. A. (**2002**). Ocean ambient sound: comparing the 1960s with the 1990s for a receiver off the California coast. *Acoustics Research Letters Online*, ***3***(2), 65-70.

Au, W. W. L. (**1993**). *The Sonar of Dolphins* (Springer-Verlag, New York), 277 p.

Au, W. W., & Würsig, B. (**2004**). Echolocation signals of dusky dolphins (Lagenorhynchus obscurus) in Kaikoura, New Zealand. *The Journal of the Acoustical Society of America*, ***115***(5), 2307-2313.

Au, W. W., Branstetter, B., Moore, P. W., & Finneran, J. J. (**2012**). Dolphin biosonar signals measured at extreme off-axis angles: Insights to sound propagation in the head. *The Journal of the Acoustical Society of America*, ***132***(2), 1199-1206.

Bahoura, M., & Simard, Y. (**2008**). Blue whale calls characterization using chirplet transform. *The Journal of the Acoustical Society of America*, ***123***(5), 3779-3779.

Bailey, H., Senior, B., Simmons, D., Rusin, J., Picken, G., and Thompson, P. M. (**2010**). Assessing underwater noise levels during pile-driving at an offshore windfarm and its potential effects on marine mammals. *Marine Pollution Bulletin*, **60**(6), 888-897.

Bao, F., Li, C., Wang, X., Wang, Q., & Du, S. (**2010**). Ship classification using nonlinear features of radiated sound: An approach based on empirical mode decomposition. *The Journal of the Acoustical Society of America*, ***128***(1), 206-214.

Baumgartner, M. F., & Mussoline, S. E. (**2011**). A generalized baleen whale call detection and classification system. *The Journal of the Acoustical Society of America*, ***129***(5), 2889-2902.

Baumgartner, M. F., Van Parijs, S. M., Wenzel, F. W., Tremblay, C. J., Esch, H. C., & Warde, A. M. (**2008**). Low frequency vocalizations attributed to sei whales (Balaenoptera borealis). *The Journal of the Acoustical Society of America*, ***124***(2), 1339-1349.

Bourassa, S. L. (**1984**). Measurement of oceanic wind speed using acoustic ambient sea noise.

Brown, J. C., & Smaragdis, P. (**2009**). Hidden Markov and Gaussian mixture models for automatic call classification. *The Journal of the Acoustical Society of America*, ***125***(6), EL221-EL224.

Buck, J. R., & Tyack, P. L. (**1993**). A quantitative measure of similarity for tursiops truncatus signature whistles. *The Journal of the Acoustical Society of America*, ***94***(5), 2497-2506.

Carevic, D. (**2013**) Modelling and Tracking of Dynamic Spectra Using a Non-Parametric Bayesian Method. In *Acoustics 2013*, Victor Harbor, Australia.

Caudal, F., and Glotin, H. (**2008**). Stochastic matched filter outperforms Teager-Kaiser-Mallat for tracking a plurality of sperm whales. In *New Trends for Environmental Monitoring Using Passive Systems, 2008* (Hyeres, French Riviera), pp. 1-9.

Chapman, N.R., and A. Price. (**2011**). Low frequency deep ocean ambient noise trend in the Northeast Pacific Ocean. *Journal of the Acoustical Society of America* ***129***(5):EL161-EL165.

Chen, C. H., Lee, J. D., & Lin, M. C. (**2000**). Classification of underwater signals using neural networks. *Tamkang Journal of Science and Engineering*, ***3***(1), 31-48.

Clark, C. W., Charif, R., Mitchell, S., and Colby, J. (**1996**). Distribution and behavior of the bowhead whale, Balaena mysticetus, based on analysis of acoustic data collected during the 1993 spring migration off Point Barrow, Alaska. *Report-International Whaling Commission*, **46**, 541-554.

Clark, C. W., Ellison, W. T. (**2004**). Potential use of low-frequency sound by baleen whales for probing the environment: Evidence from models and empirical measurements. In: Thomas, JA, Moss, CF, Vater, M eds. (2004) Echolocation in bats and dolphins. University of Chicago Press, Chicago, IL, pp. 564-581.

Clark, C. W., Ellison, W. T., Southall, B. L., Hatch, L., Van Parijs, S. M., Frankel, A., & Ponirakis, D. (**2009**). Acoustic masking in marine ecosystems: intuitions, analysis, and implication. *Marine Ecology Progress Series*, ***395***, 201-222.

Dashen, R., Flatté, S. M., Munk, W. H., Watson, K. M., & Zachariasen, F. (Eds.). (**2010**). *Sound transmission through a fluctuating ocean*. Cambridge University Press.

Datta, S., & Sturtivant, C. (**2002**). Dolphin whistle classification for determining group identities. *Signal Processing*, ***82***(2), 251-258.

DeRuiter, S. L., Bahr, A., Blanchet, M. A., Hansen, S. F., Kristensen, J. H., Madsen, P. T., Tyack, P. L., and Wahlberg, M. (**2009**). Acoustic behaviour of echolocating porpoises during prey capture. *Journal of Experimental Biology* **212**, 3100-3107.

Dobbins, P. (**2009**). Time and frequency shifted click detection. In *Underwater Acoustic Measurements 2009* (Nafplion, Greece), pp. 1-8.

Erbe, C. (**2000**). Detection of whale calls in noise: Performance comparison between a beluga whale, human listeners and a neural network. *Journal of the Acoustical Society of America*, ***108***(1), 297-303.

Erbe, C. (**2012**). Effects of underwater noise on marine mammals. In *The Effects of Noise on Aquatic Life* (pp. 17-22). Springer New York.

Erbe, C. (**2013**). Underwater passive acoustic monitoring & noise impacts on marine fauna—a workshop report. *Acoustics Australia*, ***41***(1), 211.

Erbe, C., & King, A. R. (**2008**). Automatic detection of marine mammals using information entropy. *The Journal of the Acoustical Society of America*, ***124***(5), 2833-2840.

Esfahanian, M., Zhuang, H., & Erdol, N. (**2014**). A new approach for classification of dolphin whistles. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 6038-6042). IEEE.

Farina, A., & Pieretti, N. (**2012**). The soundscape ecology: A new frontier of landscape research and its application to islands and coastal systems. *Journal of Marine and Island Cultures*, ***1***(1), 21-26.

Farmer, D. M., and Vagle, S. (**1988**). Observations of high frequency ambient sound generated by wind. In *Sea Surface Sound* (pp. 403-415). Springer Netherlands.

Fox, C. G., Matsumoto, H., & Lau, T. K. A. (**2001**). Monitoring Pacific Ocean seismicity from an autonomous hydrophone array. *Journal of Geophysical Research: Solid Earth (1978–2012)*, ***106***(B3), 4183-4206.

Fox, C. G., Radford, W. E., Dziak, R. P., Lau, T. K., Matsumoto, H., and Schreiner, A. E. (**1995**). Acoustic detection of a seafloor spreading episode on the Juan de Fuca Ridge using military hydrophone arrays. *Geophysical Research Letters*, **22**(2), 131-134.

Gabor, D. (**1946**). Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering* **93**, 429-441.

Gerard, O., Carthel, C., Coraluppi, S. (**2009**). Classification of Odontocete Buzz Clicks using a Multi-Hypothesis Tracker. In *OCEANS 2009-EUROPE* (pp. 1-7). IEEE.

Gervaise, C., Barazzutti, A., Busson, S., Simard, Y., and Roy, N. (**2010**). Automatic detection of bioacoustics impulses based on kurtosis under weak signal to noise ratio. *Applied Acoustics* **71**, 1020-1026.

Gill, P. E., Murray, W., and Wright, M. H. (**1981**). The Levenberg-Marquardt method. In *Practical optimization* (London : Academic Press, London), pp. 136-137.

Gillespie, D. (**2004**). Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram. *Canadian Acoustics*, ***32***(2), 39-47.

Gillespie, D., and Chappell, O. (**2002**). An automatic system for detecting and classifying the vocalisations of harbour porpoises. *Bioacoustics*, **13**(1), 37-61.

Gillespie, D., & Leaper, R. (**1996**). Detection of sperm whale (Physeter macrocephalus) clicks and discrimination of individual vocalisations. *Eur. Res. Cetaceans*, ***10***, 87-91.

Gledhill, K. R. (**1985**). An earthquake detector employing frequency domain techniques. *Bulletin of the Seismological Society of America*, ***75***(6), 1827-1835.

Goold, J. C., and Jefferson, T. A. (**2002**). Acoustic signals from free-ranging finless porpoises (Neophocaena phocaenoides) in the waters around Hong Kong. *The Raffles Bulletin of Zoology*, 131-139.

Gray, R. (**1984**). Vector quantization. *IEEE Assp Magazine*, **1**(2), 4-29.

Greco, M., and Gini, F. (**2006**). Analysis and modeling of echolocation signals emitted by Mediterranean bottlenose dolphins. *Eurasip Journal on Applied Signal Processing*, 1-10.

Halkias, X. C., & Ellis, D. P. (**2006**). Call detection and extraction using Bayesian inference. *Applied Acoustics*, ***67***(11), 1164-1174.

Hanson, J., Le Bras, R., Brumbaugh, D., Guern, J., Dysart, P., & Gault, A. (**2001**). Operational processing of hydroacoustics at the Prototype International Data Center. In *Monitoring the Comprehensive Nuclear-Test-Ban Treaty: Hydroacoustics* (pp. 425-456). Birkhäuser Basel.

Harland, E. (**2008**). Processing the workshop datasets using the TRUD algorithm. *Canadian Acoustics, **36***, 27-33.

Harris, F.J. (**1978**). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, ***66***(1), 51-83.

Hazlett, B. A., & Winn, H. E. (**1962**). Sound production and associated behavior of Bermuda crustaceans (Panulirus, Gonodactylus, Alpheus, and Synalpheus).*Crustaceana*, ***4***(1), 25-38.

Heimlich, S., Klinck, H., and Mellinger, D. K. (**2011**). *The Moby Sound Database for Research in the Automatic Recognition of Marine Mammal Calls*, http://www.mobysound.org/ (Last viewed on March 31, 2015).

Heindsmann, T. E., Smith, R. H., and Arneson, A. D. (**1955**). Effect of rain upon underwater noise levels. *The Journal of the Acoustical Society of America*, **27**(2), 378-379.

Helble, T. A., Ierley, G. R., Gerald, L. D., 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**(4), 2682-2699.

Hildebrand, J. A. (**2009**). Anthropogenic and natural sources of ambient noise in the ocean. *Marine Ecology Progress Series*, ***395***(5).

Holland, R. A., Waters, D. A., and Rayner, J. M. V. (**2004**). Echolocation signal structure in the Megachiropteran bat Rousettus aegyptiacus Geoffroy 1810. *Journal of Experimental Biology* **207**, 4361-4369.

Houser, D. S., Helweg, D. A., & Moore, P. W. (**1999**). Classification of dolphin echolocation clicks by energy and frequency distributions. *The Journal of the Acoustical Society of America*, ***106***(3), 1579-1585.

Huang, X., Acero, A., and Hon, H.-W. (**2001**). *Spoken Language Processing* (Prentice Hall PTR, Upper Saddle River, NJ).

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. and Liu, H. H. (**1998**). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **454**, 903-995.

Ioana, C., Quinquis, A., & Stephan, Y. (**2006**). Feature extraction from underwater signals using time-frequency warping operators. *Oceanic Engineering, IEEE Journal of*, ***31***(3), 628-645.

Jarvis, S., DiMarzio, N., Morrissey, R., and Moretti, D. (**2008**). A novel multi-class support vector machine classifier for automated classification of beaked whales and other small odontocetes. *Canadian Acoustics* **36**, 34-40.

Johnson, M., Madsen, P. T., Zimmer, W. M. X., de Soto, N. A., and Tyack, P. L. (**2006**). Foraging Blainville's beaked whales (Mesoplodon densirostris) produce distinct click types matched to different phases of echolocation. *Journal of Experimental Biology* **209**, 5038-5050.

Kaiser, J. F. (**1990a**). On a simple algorithm to calculate the `energy' of a signal. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on* (Albuquerque, NM), pp. 381-384.

Kaiser, J. F. (**1990b**). On Teager's energy algorithm and its generalization to continuous signals. In *4th IEEE Digital Signal Processing Workshop* (Mohonk, NY), pp. 17-18.

Kalman, R. (**1960**). A new approach to linear filtering and prediction problem. *Journal of Basic Engineering Transactions*, ***82***(1), 34-45.

Kamminga, C., and Beitsma, G. R. (**1990**). Investigations on cetacean sonar IX: Remarks on dominant sonar frequencies from Tursiops truncates. *Aquatic Mammals* **16**, 14-20.

Kamminga, C., Cohen Stuart, A., and Silber, G. K. (**1996**). Investigations on cetacean sonar XI: Intrinsic comparison of the wave shapes of some members of the Phocoenidae family. *Aquatic Mammals* **22**, 45-55.

Kamminga, C., and Stuart, A. C. (**1995**). Wave shape estimation of delphinid sonar signals, a parametric model approach. *Acoustics Letters* **19**, 70-76.

Kamminga, C., van Hove, M. T., Engelsma, F. J., and Terry, R. P. (**1993**). Investigations on cetacean sonar X: A comparative analysis of underwater echolocation clicks of Inia spp. and Sotalia spp. *Aquatic Mammals* **19**, 31-43.

Kandia, V., and Stylianou, Y. (**2006**). Detection of sperm whale clicks based on the Teager–Kaiser energy operator. *Applied Acoustics* **67**, 1144-1163.

Kandia, V., and Stylianou, Y. (**2008**). A phase based detector of whale clicks. In *New Trends for Environmental Monitoring Using Passive Systems, 2008* (Hyeres, French Riviera), pp. 1-6.

Kastelein, R. A., Gransier, R., Hoek, L., & Rambags, M. (**2013**). Hearing frequency thresholds of a harbor porpoise (Phocoena phocoena) temporarily affected by a continuous 1.5 kHz tone. *The Journal of the Acoustical Society of America*, ***134***(3), 2286-2292.

Kerman, B. R. (Ed.). (**1988**). *Sea surface sound: natural mechanisms of surface generated noise in the ocean* (Vol. 238). Springer Science & Business Media.

Kershenbaum, A., & Roch, M. A. (**2013**). An image processing based paradigm for the extraction of tonal sounds in cetacean communications. *The Journal of the Acoustical Society of America*, ***134***(6), 4435-4445.

Klimley, A. P., Voegeli, F., Beavers, S. C., and Le Boeuf, B. J. (**1998**). Automated listening stations for tagged marine fishes. *Marine Technology Society. Marine Technology Society Journal*, **32**(1), 94.

Klinck, H., and Mellinger, D. K. (**2011**). The energy ratio mapping algorithm: A tool to improve the energy-based detection of odontocete echolocation clicks. *The Journal of the Acoustical Society of America* **129**, 1807-1812.

Krause, B. L., (**1987**). Bioacoustics, habitat ambience in ecological balance. *Whole Earth Rev* **57**:14 – 18.

Laiolo, P. (**2010**). The emerging significance of bioacoustics in animal species conservation. *Biological Conservation*, ***143***(7), 1635-1645.

Leis JM, Carson-Ewart BM, Hay AC, Cato DH (**2003**). Coral reef sounds enable nocturnal navigation by some reef-fish larvae in some places and at some times. *J Fish Biol* **63**: 724–737.

Lewis, T., Gillespie, D., Lacey, C., Matthews, J., Danbolt, M., Leaper, R., McLanaghan, R. and Moscrop, A. (**2007**). Sperm whale abundance estimates from acoustic surveys of the Ionian Sea and Straits of Sicily in 2003. *Journal of the Marine Biological Association of the United Kingdom*, **87**(01), 353-357.

Li, B. (**2010**). Acoustic observation of ice rifting and breaking events on the Antarctic ice shelf using remote hydroacoustic listening stations.

Li, B., and Gavrilov, A. (**2008**). Localization of Antarctic ice breaking events by frequency dispersion of the signals received at a single hydroacoustic station in the Indian Ocean. In *Acoustics '08*, Paris 29 June – 4 July, 2008. ISBN 978-2-9521105-4-9.

Lindeberg, T. (**1993**). Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, ***11***(3), 283-318.

Lindeberg, T. (**1998a**). Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, ***30***(2), 117-156.

Lindeberg, T. (**1998b**). Feature detection with automatic scale selection. *International journal of computer vision*, ***30***(2), 79-116.

Lourens, J. G. (**1990**). Passive sonar detection of ships with spectrograms. *Proceedings of the South African Symposium on Communications and Signal Processing*, 147-151.

Madhusudhana, S., Gavrilov, A., & Erbe, C. (**2015**). Automatic detection of echolocation clicks based on a Gabor model of their waveform. *The Journal of the Acoustical Society of America*, ***137***(6), 3077-3086.

Madhusudhana, S. K., Oleson, E. M., Soldevilla, M. S., Roch, M., & Hildebrand, J. (**2008**). Frequency based algorithm for robust contour extraction of blue whale B and D calls. In *OCEANS 2008-MTS/IEEE Kobe Techno-Ocean* (pp. 1-8). IEEE.

Madhusudhana, S. K., Roch, M. A., Oleson, E. M., Soldevilla, M. S., & Hildebrand, J. A. (**2009**). Blue whale B and D call classification using a frequency domain based robust contour extractor. In *OCEANS 2009-EUROPE*(pp. 1-7). IEEE.

Madsen, P. T., Johnson, M., de Soto, N. A., Zimmer, W. M. X., & Tyack, P. (**2005**). Biosonar performance of foraging beaked whales (Mesoplodon densirostris). *Journal of Experimental Biology*, ***208***(2), 181-194.

Madsen, P. T., Kerr, I., and Payne, R. (**2004**). Echolocation clicks of two free-ranging, oceanic delphinids with different food preferences: false killer whales Pseudorca crassidens and Risso's dolphins Grampus griseus. *Journal of Experimental Biology* **207**, 1811-1823.

Mahalanobis, P. C. (**1936**). On the generalized distance in statistics. *Proceedings of the national institute of sciences (Calcutta)*, **2**, 49-55.

Mallawaarachchi, A., Ong, S. H., Chitre, M., & Taylor, E. (**2008**). Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles. *The Journal of the Acoustical Society of America*, ***124***(2), 1159-1170.

Manasseh, R., Babanin, A. V., Forbes, C., Rickards, K., Bobevski, I., and Ooi, A. (**2006**). Passive acoustic determination of wave-breaking events and their severity across the spectrum. *Journal of Atmospheric and Oceanic Technology*, **23**(4), 599-618.

Mann, S., and Haykin, S. (**1991**). The chirplet transform: A generalization of Gabor's logon transform. In *Vision Interface '91* (Calgary, Canada), pp. 205-212.

Mann, D. A., and Lobel, P. S. (**1995a**). Passive acoustic detection of fish sound production associated with courtship and spawning. *Bulletin of marine science-Miami*, **57**, 705-705.

Mann, D. A., and Lobel, P. S. (**1995b**). Passive acoustic detection of sounds produced by the damselfish, Dascyllus albisella (Pomacentridae). *Bioacoustics*, **6**(3), 199-213.

Marčelja, S. (**1980**). Mathematical description of the responses of simple cortical cells*. *Journal of the Optical Society of America*, **70**(11), 1297-1300.

Marques, T. A., Thomas, L., Ward, J., DiMarzio, N., and Tyack, P. L. (**2009**). Estimating cetacean population density using fixed passive acoustic sensors: an example with Blainville's beaked whales. *The Journal of the Acoustical Society of America*, **125**(4), 1982-1994.

Marr, D., and Hildreth., E. (**1980**). Theory of edge detection. *Proc. Roy. Soc. London*, ***B 207***, 187-217.

McCarthy, E., Moretti, D., Thomas, L., DiMarzio, N., Morrissey, R., Jarvis, S., Ward, J., Izzi, A., and Dilley, A. (**2011**). Changes in spatial and temporal

distribution and vocal behavior of Blainville's beaked whales (Mesoplodon densirostris) during multiship exercises with mid-frequency sonar. *Marine Mammal Science*, **27**(3), E206-E226.

McCauley, R.D., Fewtrell, J., Popper, A.N. (**2003**). High intensity anthropogenic sound damages fish ears. *Journal of the Acoustical Society of America*, **113** (1), pp. 638-642.

McConnell, S. (**1983**). Remote sensing of the air-sea interface using microware acoustics. In *OCEANS '83*, (pp. 85-92). IEEE.

Medwin, H., Nystuen, J. A., Jacobus, P. W., Ostwald, L. H., and Snyder, D. E. (**1992**). The anatomy of underwater rain noise. *The Journal of the Acoustical Society of America*, **92**(3), 1613-1623.

Mellinger, D. K., & Clark, C. W. (**1997**). Methods for automatic detection of mysticete sounds. *Marine & Freshwater Behaviour & Phy*, ***29***(1-4), 163-181.

Mellinger, D. K., & Clark, C. W. (**2000**). Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America*, ***107***(6), 3518-3529.

Mellinger, D. K., Martin, S. W., Morrissey, R. P., Thomas, L., & Yosco, J. J. (**2011**). A method for detecting whistles, moans, and other frequency contour sounds. *The Journal of the Acoustical Society of America*, ***129***(6), 4055-4061.

Miksis-Olds JL, Madden LE (**2014**) Environmental Predictors of Ice Seal Presence in the Bering Sea. *PLoS ONE* **9**(9): e106998.

Møhl, B., Wahlberg, M., Madsen, P. T., Heerfordt, A., and Lund, A. (**2003**). The monopulsed nature of sperm whale clicks. *The Journal of the Acoustical Society of America* **114**, 1143-1154.

Moretti, D., DiMarzio, N., Morrissey, R., Ward, J., and Jarvis, S. (**2006**). Estimating the density of Blainville's beaked whale (Mesoplodon densirostris) in the

Tongue of the Ocean (TOTO) using passive acoustics. In *OCEANS 2006* (pp. 1-5). IEEE.

Morrissey, R. P., Ward, J., DiMarzio, N., Jarvis, S., and Moretti, D. J. (**2006**). Passive acoustic detection and localization of sperm whales (Physeter macrocephalus) in the tongue of the ocean. *Applied Acoustics* **67**, 1091-1105.

Mouy, X., Leary, D., Martin, B., and Laurinolli, M. (**2008**). A comparison of methods for the automatic classification of marine mammal vocalizations in the Arctic. In *New Trends for Environmental Monitoring Using Passive Systems, 2008*, pp. 1–6.

Munger, L. M., Mellinger, D. K., Wiggins, S. M., Moore, S. E., & Hildebrand, J. A. (**2005**). Performance of spectrogram cross-correlation in detecting right whale calls in long-term recordings from the Bering Sea. *Canadian Acoustics*, ***33***(2), 25-34.

Munkres, J. (**1957**). Algorithms for the assignment and transportation problems. *Journal of the society for Industrial and Applied Mathematics*, 5(1), 32-38.

Nystuen, J. A., McPhaden, M. J., and Freitag, H. P. (**2000**). Surface measurements of precipitation from an ocean mooring: The underwater acoustic log from the south china sea*. *Journal of Applied Meteorology*, **39**(12), 2182-2197.

Ogden, G. L., Zurk, L. M., Jones, M. E., & Peterson, M. E. (**2011**). Extraction of small boat harmonic signatures from passive sonar. *The Journal of the Acoustical Society of America*, ***129***(6), 3768-3776.

Oppenheim A. V., & Schafer R. W. (**1975**). *Digital Signal Processing* (Prentice-Hall, Inc., Englewood Cliffs).

Oswald, J. N., Barlow, J., Norris, T., F. (**2003**). Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean. *Marine Mammal Science, **19**(1),* 20-37.

Oswald, J. N., Rankin, S., Barlow, J., Lammers, M. O. (**2007**). A tool for real-time acoustic species identification of delphinid whistles. *The Journal of the Acoustical Society of America, **122(1)***, 587-595.

Ou, H., Au, W. W., Zurk, L. M., & Lammers, M. O. (**2013**). Automated extraction and classification of time-frequency contours in humpback vocalizations. *The Journal of the Acoustical Society of America*, ***133***(1), 301-310.

Popper, A.N., Salmon, M., Horch, K.W., (**2001**). Acoustic detection and communication by decapod crustaceans. *J. Comp. Physiol.* **187**, 83–89.

Prosperetti, A., and Oguz, H. N. (**1993**). The impact of drops on liquid surfaces and the underwater noise of rain. *Annual Review of Fluid Mechanics*, **25**(1), 577-602.

Rabiner, L. R. (**1989**). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257-286.

Radford, C.A., Stanley, J.A., Simpson, S.D., Jeffs, A.G., **2011**. Juvenile coral reef fish use sound to locate habitats. *Coral Reefs* **30** (2), 295–305.

Rankin, S., Baumann-Pickering, S., Yack, T., and Barlow, J. (**2011**). Description of sounds recorded from Longman's beaked whale, Indopacetus pacificus. *The Journal of the Acoustical Society of America* **130**, EL339-EL344.

Richardson, W. J. , and Thomson, D. H. (**1995**). Marine mammal sounds. In *Marine Mammals and Noise*, edited by W. J.Richardson, C. R. Greene, Jr., C. I. Malme, and D. H. Thomson (Academic Press, San Diego), pp. 159–204.

Roch, M. A., Klinck, H., Baumann-Pickering, S., Mellinger, D. K., Qui, S., Soldevilla, M. S., and Hildebrand, J. A. (**2011a**). Classification of echolocation clicks from odontocetes in the Southern California Bight. *The Journal of the Acoustical Society of America* **129**, 467-475.

Roch, M. A., Brandes, T. S., Patel, B., Barkley, Y., Baumann-Pickering, S., and Soldevilla, M. S. (**2011b**). Automated extraction of odontocete whistle contours. *The Journal of the Acoustical Society of America* **130**, 2212-2223.

Roch, M. A., Soldevilla, M. S., Burtenshaw, J. C., Henderson, E. E., & Hildebrand, J. A. (**2007**). Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California. *The Journal of the Acoustical Society of America*, ***121***(3), 1737-1748.

Roch, M. A., Soldevilla, M. S., Hoenigman, R., Wiggins, S. M., and Hildebrand, J. A. (**2008**). Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes. *Canadian Acoustics* **36**, 41-47.

Sakoe, H., and Chiba, S. (**1978**). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, **26**(1), 43-49.

Scambos, T. A., Hulbe, C., Fahnestock, M., & Bohlander, J. (**2000**). The link between climate warming and break-up of ice shelves in the Antarctic Peninsula. *Journal of Glaciology, **46***(154), 516-530.

Schafer, R.M., (**1977**). *The soundscape: our sonic environment and the tuning of the world*. Destiny Books. Rochester, NY. US.

Schmitz, B. (**2002**). Sound production in Crustacea with special reference to the Alpheidae. In *The crustacean nervous system* (pp. 536-547). Springer Berlin Heidelberg.

Scrimger, J. A. (**1985**). Underwater noise caused by precipitation. *The Journal of the Acoustical Society of America*, **78**(S1), S2-S2.

Scrimger, J. A., Evans, D. J., McBean, G. A., Farmer, D. M., and Kerman, B. R. (**1987**). Underwater noise due to rain, hail, and snow. *The Journal of the Acoustical Society of America*, **81**(1), 79-86.

Shaw, P. T., Watts, D. R., and Rossby, H. T. (**1978**). On the estimation of oceanic wind speed and stress from ambient noise measurements. *Deep Sea Research*, **25**(12), 1225-1233.

Simpson S.D., M.G. Meekan, A. Jeffs, J.C. Montgomery, R.D. McCauley. (**2008**). Settlement-stage coral reef fish prefer the higher-frequency invertebrate-generated audible component of reef noise, *Animal Behaviour*, Volume **75**, Issue 6, June 2008, Pages 1861-1868, ISSN 0003-3472.

Sirovic, A., Williams L. N., Kerosky, S. M., Wiggins, S. M., Hildebrand, J. A. (**2013**). Temporal separation of two fin whale call types across the eastern North Pacific. *Mar Biol, **160***, 47-57.

Soldevilla, M. S., Henderson, E. E., Campbell, G. S., Wiggins, S. M., Hildebrand, J. A., and Roch, M. A. (**2008**). Classification of Risso's and Pacific white-sided dolphins using spectral properties of echolocation clicks. *The Journal of the Acoustical Society of America* **124**, 609-624.

Solé Marta, Marc Lenoir, Mercè Durfort, Manel López-Bejar, Antoni Lombarte, Mike van der Schaar, Michel André. (**2013**). Does exposure to noise from human activities compromise sensory information from cephalopod statocysts?, *Deep Sea Research Part II: Topical Studies in Oceanography*, Volume **95**, 15 October 2013, Pages 160-181, ISSN 0967-0645.

Sorensen, E., Ou, H. H., Zurk, L. M., & Siderius, M. (**2010**). Passive acoustic sensing for detection of small vessels. In *OCEANS 2010 MTS/IEEE Seattle*, 1-8.

Southall, B.L., A.E. Bowles, W.T. Ellison, J.J. Finneran, R.L. Gentry, C.R. Greene, D. Kastak, D.R. Ketten, J.H. Miller, P.E. Nachtigall, W.J. Richardson, J.A. Thomas, and P.L. Tyack. (**2007**). Marine mammal sound exposure criteria: Initial scientific recommendations. *Aquatic Mammals* **33**:411-522.

Stafford, K. M., Fox, C. G., & Clark, D. S. (**1998**). Long-range acoustic detection and localization of blue whale calls in the northeast Pacific Ocean. *The Journal of the Acoustical Society of America*, ***104***(6), 3616-3625.

Stafford, K. M., Fox, C. G., & Mate, B. R. (**1994**). Acoustic detection and location of blue whales (Balaenoptera musculus) from SOSUS data by matched filtering. *The Journal of the Acoustical Society of America*, ***96***(5), 3250-3251.

Svend, G., and Herlufsen, H. (**1987**). Windows to FFT analysis (Part I). *BRÜEL & KJÆR, Technical Review 3*.

Sukhovich, A., Irisson, J. O., Perrot, J., & Nolet, G. (**2014**). Automatic recognition of T and teleseismic P waves by statistical analysis of their spectra: An application to continuous records of moored hydrophones. *Journal of Geophysical Research: Solid Earth*, ***119***(8), 6469-6485.

Sukhovich, A., Irisson, J. O., Simons, F. J., Ogé, A., Hello, Y., Deschamps, A., & Nolet, G. (**2011**). Automatic discrimination of underwater acoustic signals generated by teleseismic P-waves: A probabilistic approach. *Geophysical Research Letters*, ***38***(18).

Szeliski, R. (**2010**). *Computer vision: algorithms and applications*. Springer.

Talandier, J., Hyvernaud, O., Reymond, D., & Okal, E. A. (**2006**). Hydroacoustic signals generated by parked and drifting icebergs in the Southern Indian and Pacific Oceans. *Geophysical Journal International*, ***165***(3), 817-834.

Tavolga, William N., (**1971**). 6 Sound Production and Detection, In: W.S. Hoar and D.J. Randall, Editor(s), *Fish Physiology*, Academic Press, 1971, Volume 5, Pages 135-205, ISSN 1546-5098, ISBN 9780123504050.

Thode, A. M., Kim, K. H., Blackwell, S. B., Greene Jr, C. R., Nations, C. S., McDonald, T. L., & Macrander, A. M. (**2012**). Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys. *The Journal of the Acoustical Society of America*, ***131***(5), 3726-3747.

Thompson, P. O., and Friedl, W. A. (**1982**). A long term study of low frequency sounds from several species of whales off Oahu, Hawaii. *Cetology* 45:1–19.

Thorpe, C. W., and Dawson, S. M. (**1991**). Automatic measurement of descriptive features of Hector's dolphin vocalizations. *The Journal of the Acoustical Society of America* **89**, 435-443.

Tolimieri N, Jeffs A, Montgomery JC (**2000**) Ambient sound as a cue for navigation by the pelagic larvae of reef fishes. Mar Ecol Prog Ser **207**:219–224.

Tyack, P. L., & Clark, C. W. (**2000**). Communication and acoustic behavior of dolphins and whales. In *Hearing by whales and dolphins* (pp. 156-224). Springer New York.

Tyack, P.L., Zimmer, W.M., Moretti, D., Southall, B.L., Claridge, D.E., Durban, J.W., Clark, C.W., D'Amico, A., DiMarzio, N., Jarvis, S. and McCarthy, E., (**2011**). Beaked whales respond to simulated and actual navy sonar. *PloS one*, **6**(3), p.e17009.

Urazghildiiev, I. R., Clark, C. W., Krein, T. P., & Parks, S. E. (**2009**). Detection and recognition of North Atlantic right whale contact calls in the presence of ambient noise. *Oceanic Engineering, IEEE Journal of*, ***34***(3), 358-368.

Urick, R.J., (**1983**). Principles of Underwater Sound, third ed. McGraw-Hill Book Company, New York.

Valinski, W., & Rigley, L. (**1981**). Function of sound production by the skunk loach Botia horae (Pisces, Cobitidae). *Zeitschrift für Tierpsychologie*, ***55***(2), 161-172.

van der Schaar, M., Delory, E., van der Weide, J., Kamminga, C., Goold, J. C., Jaquet, N., and Andre, M. (**2007**). A comparison of model and non-model based time-frequency transforms for sperm whale click classification. *Journal of the Marine Biological Association of the United Kingdom* **87**, 27-34.

von Seggern, D. (**1993**). Standard curves and surfaces. CRC Press, Boca Raton, p. 324.

Vermeij, M. J., Marhaver, K. L., Huijbers, C. M., Nagelkerken, I., & Simpson, S. D. (**2010**). Coral larvae move toward reef sounds. *PloS one*, ***5***(5), e10660.

Webb, S. C. (**1998**). Broadband seismology and noise under the ocean. *Reviews of Geophysics*, ***36***(1), 105-142.

Weisburn, B. A., Mitchell, S. G., Clark, C. W., and Parks, T. W. (**1993**). Isolating biological acoustic transient signals. In *ICASSP-93*, IEEE, Vol. **1**, pp. 269–272.

Wenz, G.M., (**1962**). Acoustic ambient noise in the ocean: spectra and sources. *J. Acoust. Soc. Am.* **34**, 1936–1956.

Woodman, G. H., Wilson, S. C., Li, V. Y., & Renneberg, R. (**2004**). A direction-sensitive underwater blast detector and its application for managing blast fishing. *Marine pollution bulletin*, ***49***(11), 964-973.

Wright, A. J., Soto, N. A., Baldwin, A. L., Bateson, M., Beale, C. M., Clark, C., *et al*. (**2007**). Do Marine Mammals Experience Stress Related to Anthropogenic Noise?. *International Journal of Comparative Psychology*, ***20***(2).

Wyatt, R. (**2008**). Review of Existing Data on Underwater Sounds Produced by the Oil and Gas Industry-Issue 1. *Report by Seiche Measurements Ltd., Great Torrington, to Joint Industry Programme on Sound and Marine Life, Seiche Measurements Limited Ref–S186.*

Yang, S., Li, Z., & Wang, X. (**2002**). Ship recognition via its radiated sound: The fractal based approaches. *The Journal of the Acoustical Society of America*, ***112***(1), 172-177.

Zakarauskas, P. (**1993**). Detection and localization of nondeterministic transients in time series and application to ice-cracking sound. *Digital Signal Processing*, **3**(1), 36-45.

Zimmer, W. M. X., Johnson, M. P., Madsen, P. T., and Tyack, P. L. (**2005**). Echolocation clicks of free-ranging Cuvier's beaked whales (Ziphius cavirostris). *The Journal of the Acoustical Society of America* **117**, 3919-3927.

Zimmer, W. M. X., and Pavan, G. (**2008**). Context driven detection/classification of Cuvier's beaked whale (Ziphius cavirostris). In *New Trends for Environmental Monitoring Using Passive Systems, 2008* (Hyeres, French Riviera), pp. 1-6.

Appendix A

A.1. Tonal detector: Training data and analyses

The data used for training included 879 hand-traced TF contours from four different recordings. Spectrograms computed for the different recordings differed in their time and frequency resolutions. The resolutions were chosen to be appropriate for analysing the tonal signals contained in the respective recordings. The training results thus obtained were more generic and not biased to a particular set of spectrogram parameter values. The ridge detection procedure described in Section 3.2.2 was applied to each spectrogram. The tracing of the TF contours was performed by manually connecting spectral peaks corresponding to detected ridge points across successive frames. The correspondence of chosen spectral peaks to detected ridge points is necessary because of the way process and measurement errors have been defined.

At each time step i (spectrogram frame) in a traced contour, the differences between f_i, ρ_i and τ_i values and their respective estimates obtained using f_{i-1}, ρ_{i-1}, τ_{i-1} and ψ_{i-1} and Eqs. (3.16), (3.17) and (3.18) describe the process error at the time step, as per the definition of process error in Section 3.2.3. The process errors in f, ρ and τ computed over successive time steps across all hand-traced contours are characterised with histogram plots in Figure A.1. Since different spectrograms are considered, the training results shown for error in f were normalised by the frequency resolution of the respective spectrogram. This enables f error variance to be expressed in terms of the number of frequency bins. Considering the bell-shaped profile evident in the histograms and the symmetric distribution of values around the respective means, I assume that the data in the histograms can be represented by Gaussian distributions. The process error covariance Q was determined using the computed f, ρ and τ errors. The bottom right plot of Figure A.1 characterises the differences in ψ between successive points in a traced contour. This value is used

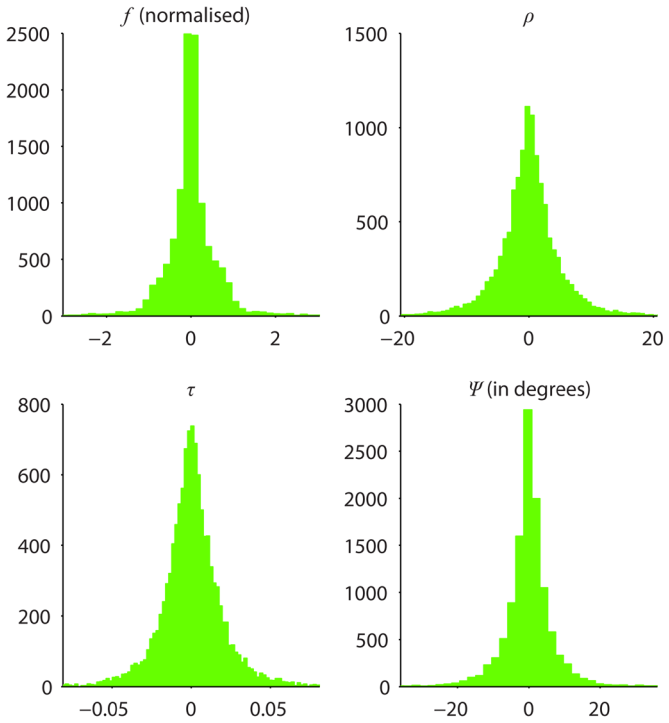

Figure A.1. Histograms of the differences in the predicted and measured values for f (top left), ρ (top right) and τ (bottom left). A histogram of the differences in ψ across successive contour points is shown in the bottom right panel.

along with f, ρ and τ errors in the determination of Mahalanobis distances in the tracing process.

Instantaneous FM rate at each ridge point was determined as the ratio of frequency difference with a connected ridge point in the succeeding frame to the time resolution of the spectrogram. The magnitudes of the instantaneous FM rates were analysed as a function of the start frequency (the frequency of the first ridge point). The recorded FM rates are shown in the scatter plots of Figure A.2. Observably, higher rates of FM were possible at higher start frequencies than at lower start frequencies. The conceptual analogy for bioacoustic tonal signals here is that an animal's sound production system more easily or readily changes the FM rate for high-frequency whistles than for low-frequency whistles. I aimed to characterise the limiting positive and negative FM rates at different start frequencies. Following empirical analysis, the positive and negative FM rate capping functions are defined as

$$\begin{aligned}\overline{\mathrm{F}}^{+}(f_s) &= f_s[\log 10(f_s)+7] \\ \overline{\mathrm{F}}^{-}(f_s) &= -f_s[\log 10(f_s)+3]\end{aligned}, \tag{A.1.1}$$

where f_s is the start frequency, $\overline{\mathrm{F}}^{+}$ and $\overline{\mathrm{F}}^{-}$ are the corresponding positive and negative FM rate limits, respectively. Note that better modelling of the limits than those considered in Eq. (A.1.1) may be available. However, it is not vital for this study and the coarse limits considered do suffice.

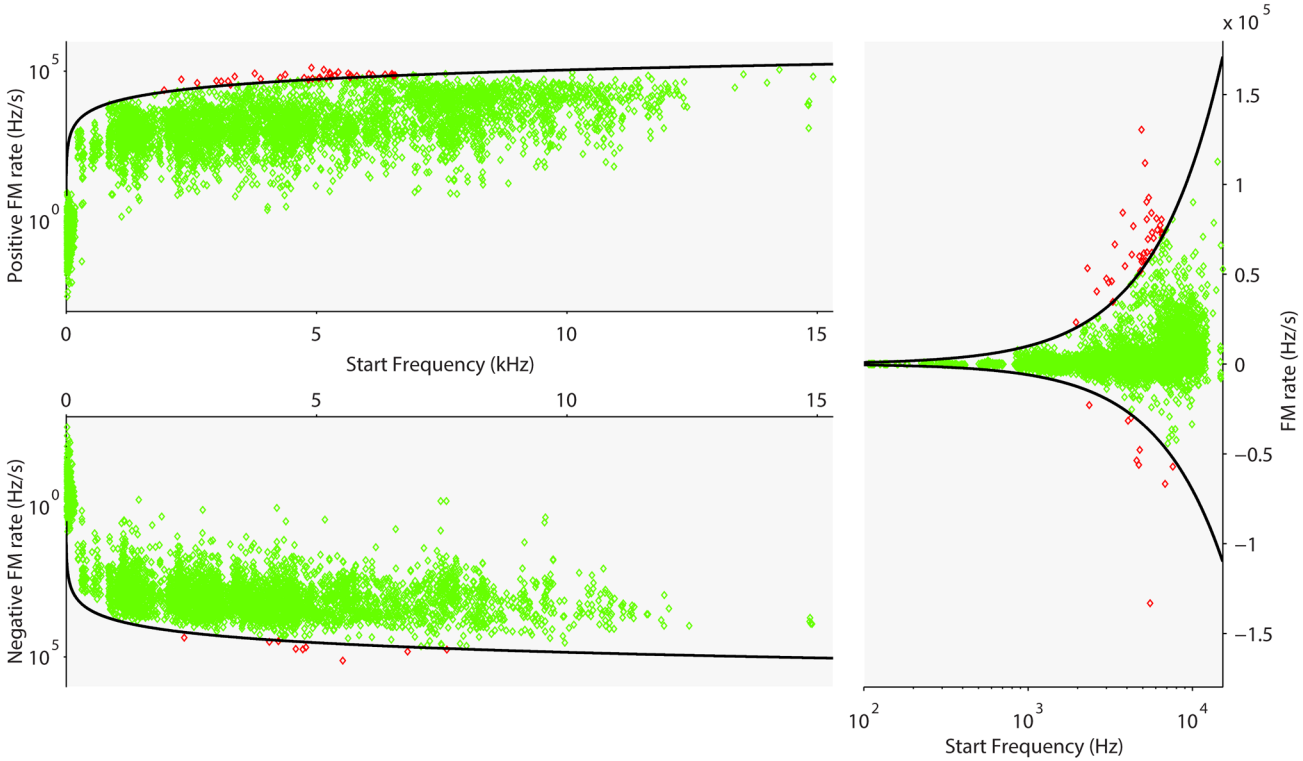

Figure A.2. Scatter plots of instantaneous rates of frequency modulation as a function of the start frequency. The plots in the left column show positive and negative FM rates separately while the plot in the right column shows both positive and negative rates collectively. The equivalent plots show data with different axes scaling for better emphasis at different start frequency ranges. The overlaid black curves indicate the considered capping functions. With respect to the capping functions, the green (light) and red (dark) data points indicate the contained and outlier data values, respectively.

A.2. Tonal detector: Defining frequency range limitations for candidate extensions

Candidate ridge points considered at frame $i + 1$ for the extension of a TF contour being traced are restricted to only those ridge points that occur within a certain frequency range around the contour's frontier. Two independent lookup frequency ranges are determined, one based on the contour's $\tilde{\psi}_m$ and the other based on its frequency value at the last frame $f_{i|m}$. The final frequency range restriction is defined by combining the two lookup ranges.

The lookup range based on $\tilde{\psi}_m$ is obtained by considering an additional range of angles around it. However, consideration of a predefined range of additional angles around $\tilde{\psi}_m$ is not a viable option. Given that the f values of ridge points are multiples of the spectrogram's frequency resolution, choosing a small range may not be beneficial with small $\tilde{\psi}_m$ as it would allow for few or no additional frequency bins to be considered. Choosing a large range, on the other hand, may be detrimental with large $\tilde{\psi}_m$ as it would result in a large number of additional bins to be considered. I define an angle lenience quantity of

$$\phi = \phi_0 - \alpha|\tilde{\psi}_m| \tag{A.2.1}$$

which allows for the specification of an adaptive range of angles around $\tilde{\psi}_m$. Setting $\phi_o = 60°$ and $\alpha = 2/3$ yields a lenience angle of $\phi = 60°$ for $\tilde{\psi}_m = 0°$ and ϕ approaches 0° as $\tilde{\psi}_m$ approaches 90°. A frequency lookup range can now be defined as $\left[\left(\text{truncate}\left(\tan\left(\tilde{\psi}_m - \phi\right)\right) - 1\right)\cdot \Delta f,\ \left(\text{truncate}\left(\tan\left(\tilde{\psi}_m + \phi\right)\right) + 1\right)\cdot \Delta f\right]$. The truncate(·) function rounds the specified quantity towards zero. For $\tilde{\psi}_m \approx 0°$, added lenience angle of 60° translates to a lookup range of two frequency bins on either side of the bin corresponding to $f_{i|m}$. For larger $|\tilde{\psi}_m|$, although ϕ decreases, the defined range function still allows for a fair number of frequency bins to be considered in looking for candidate ridge points. The effect of the added lenience angles and the number of additional frequency bins resulting from it are shown in Figure A.3.

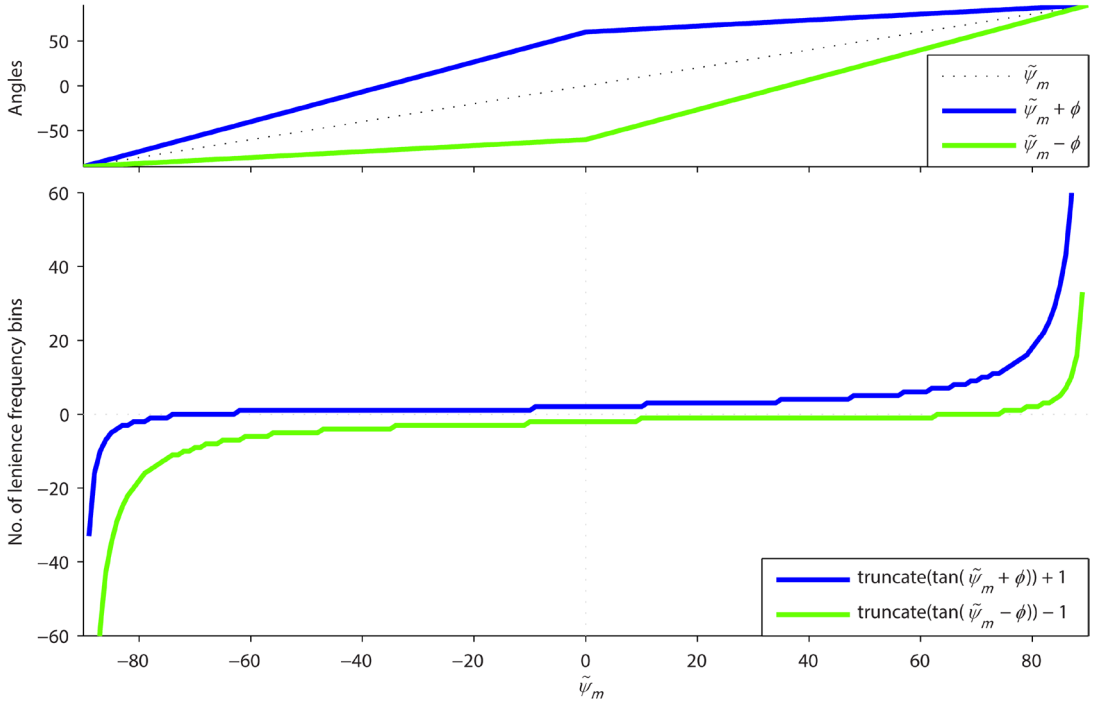

Figure A.3. Lenience angles (top row) and the corresponding additional frequency bins (bottom row) arising from Eq. (A.2.1) for different values of $\tilde{\psi}_m$, with $\phi_o = 60°$ and $\alpha = 2/3$. All angles are specified in degrees.

Based on $f_{i|m}$, a different frequency range can be specified using the frequency slope capping functions $\overline{\mathrm{F}}^{+}$ and $\overline{\mathrm{F}}^{-}$ defined in Appendix A.1. The final range of frequencies within which to look for candidate ridge points is based on the more restrictive combination arising out of the two specified ranges –

$$\begin{aligned} \mathrm{f\_max}(f_{i|m}, \tilde{\psi}_m) &= f_{i|m} + \min\left(\overline{\mathrm{F}}^{+}(f_{i|m}) \cdot \Delta t, \quad \left[\mathrm{truncate}\left(\tan\left(\tilde{\psi}_m + \phi\right)\right) + 1\right] \cdot \Delta f\right) \\ \mathrm{f\_min}(f_{i|m}, \tilde{\psi}_m) &= f_{i|m} + \max\left(\overline{\mathrm{F}}^{-}(f_{i|m}) \cdot \Delta t, \quad \left[\mathrm{truncate}\left(\tan\left(\tilde{\psi}_m - \phi\right)\right) - 1\right] \cdot \Delta f\right) \end{aligned}. \qquad \text{(A.2.2)}$$