Detection and Classification of Whale Acoustic Signals

by

Yin Xian

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Loren Nolte, Supervisor

_____
Douglas Nowacek (Co-supervisor)

_____
Robert Calderbank (Co-supervisor)

_____
Xiaobai Sun

_____
Ingrid Daubechies

_____
Galen Reeves

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering in the Graduate School of Duke University
2016

Abstract

Detection and Classification of Whale Acoustic Signals

by

Yin Xian

Department of Electrical and Computer Engineering
Duke University

Date: _____
Approved:

_____
Loren Nolte, Supervisor

_____
Douglas Nowacek (Co-supervisor)

_____
Robert Calderbank (Co-supervisor)

_____
Xiaobai Sun

_____
Ingrid Daubechies

_____
Galen Reeves

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University
2016

Abstract

This dissertation focuses on two vital challenges in relation to whale acoustic signals: detection and classification.

In detection, we evaluated the influence of the uncertain ocean environment on the spectrogram-based detector, and derived the likelihood ratio of the proposed Short Time Fourier Transform detector. Experimental results showed that the proposed detector outperforms detectors based on the spectrogram. The proposed detector is more sensitive to environmental changes because it includes phase information.

In classification, our focus is on finding a robust and sparse representation of whale vocalizations. Because whale vocalizations can be modeled as polynomial phase signals, we can represent the whale calls by their polynomial phase coefficients. In this dissertation, we used the Weyl transform to capture chirp rate information, and used a two dimensional feature set to represent whale vocalizations globally. Experimental results showed that our Weyl feature set outperforms chirplet coefficients and MFCC (Mel Frequency Cepstral Coefficients) when applied to our collected data.

Since whale vocalizations can be represented by polynomial phase coefficients, it is plausible that the signals lie on a manifold parameterized by these coefficients. We also studied the intrinsic structure of high dimensional whale data by exploiting its geometry. Experimental results showed that nonlinear mappings such as Laplacian Eigenmap and ISOMAP outperform linear mappings such as PCA and MDS, suggesting that the whale acoustic data is nonlinear.

We also explored deep learning algorithms on whale acoustic data. We built each layer as convolutions with either a PCA filter bank (PCANet) or a DCT filter bank (DCTNet). With the DCT filter bank, each layer has different a time-frequency scale representation, and from this, one can extract different physical information. Experimental results showed that our PCANet and DCTNet achieve high classification rate on the whale vocalization data set. The word error rate of the DCTNet feature is similar to the MFSC in speech recognition tasks, suggesting that the convolutional network is able to reveal acoustic content of speech signals.

To God. To my family, my brothers and sisters at Chinese Christian Mission Church, Blacknall Memorial Presbyterian Church and Guangzhou Shifu Christian Church.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

H_0	Null Hypothesis.
H_1	Alternative Hypothesis.
P_D	Probability of Detection.
P_F	Probability of False Alarm.

Abbreviations

AM	Amplitude-Modulated.
AUC	Area Under the Curve.
DCT	Discrete Cosine Transform.
DFT	Discrete Fourier Transform.
DOF	Degree Of Freedom.
EMD	Empirical Mode Decomposition.
FM	Frequency-Modulated.
LFSC	Linear Frequency Spectral Coefficients.
LLE	Local Linear Embedding
LRT	Likelihood Ratio Test.
MDS	Multidimensional Scaling
MFCC	Mel Frequency Cepstral Coefficients.
MFSC	Mel-Frequency Spectral Coefficients.

PAM	Passive Acoustic Monitoring.
PCA	Principal Component Analysis.
PDF	Probability Density Function.
PMF	Probability Mass Function
PPS	Polynomial-Phase Signal.
ROC	Receiver Operating Characteristic.
SKE	Signal Known Exactly.
SNR	Signal to Noise Ratio.
STFT	Short Time Fourier Transform.
WER	Word Error Rate.
WVD	Wigner Ville Distribution.

Acknowledgements

I would like to express my deepest gratitude toward and dedicate my dissertation to Jesus Christ. His amazing love, guidance and discipline led me to understand the meaning of my PhD studies and my life. He humbles me, and opened my heart. I would have quit my PhD studies two years ago if he hadn't come and found me. He showed me a path that I had never known, and saved me from my bad habits, my funding difficulties, and multiple research challenges. He bestows his abundant blessings on me that I don't deserve, and lets me know who I am meant to be.

This dissertation could never have been finished without the constant prayers, support, love and encouragements of my parents, my brothers and sisters at Chinese Christian Mission Church, Blacknall Presbyterian Church, and Guangzhou Shifu Christian Church. They are my family. They give me full understanding and support every time I fail, and celebrate my every progress. It is through them that I experience the amazing love of God.

I would also especially like to thank Dr. Andrew Thompson for his great help, instruction, prayers and encouragements. He is my role model for teaching and my brother in heaven. He helped me and did not ask for anything in return. I am really thankful for his time, patience, and generosity. May the Lord remember his kindness and bless him.

I am indebted to Duke University and the Department of Electrical and Computer Engineering. They could have kicked me out any time during my first four

years of study given my poor academic performance, but they didn't. They provided me help instead, and financial support that I didn't and don't deserve. My thanks go especially to Dr. Steve Cummer, Dr. Krishnendu Chakrabarty, Dr. Stacy Tantum, Ms. Amy Kostrewa and Ms. Samantha Morton. I also owe an apology to Dr. Qing Liu for not performing well when I was in his lab. I thank him for his instruction on partial differential equations and electromagnetics, which have benefited my understanding of the physics behind signal processing. I would like to thank Duke University for the provision of an excellent learning environment. I learned a lot from the symposiums, workshops and seminars.

I thank Dr. Loren Nolte, my advisor, for his advice and help when I was met with difficulties. I thank him for teaching me detection theory and presentation skills. His concern about research details helped me uncover details I never would have seen on my own. I thank him for introducing me to underwater acoustics and marine mammal research.

I thank Dr. Douglas Nowacek for his generous help funding and recommending me, which made this dissertation possible. I thank God for introducing me to Dr. Robert Calderbank. He is brilliant and has a kind heart. I thank him for providing me the opportunity to study subjects that advanced my research. I am thankful for his effort to maintain the IID, which is a great place for study and research.

I also thank God for introducing me to Dr. Xiaobai Sun. She has great insight on research direction, and is sensitive to the research frontier. She is a very nice, responsible teacher, and would listen to my research presentations for hours without having her lunch. She is very knowledgeable and keen on research. I benefit every time I talk to her. I am forever indebted to her.

I am honored to have Dr. Ingrid Daubechies to be my committee member, and I am thankful for her time, instruction and patience. I would also like to thank Dr. Galen Reeves for being a committee member of my qualifying, preliminary and

final exam. I am thankful for his instruction and sincere advice on doing research.

I thank Dr. Jianfeng Lu for his generous help and instruction on time-frequency analysis. I am grateful to have received his help, care and encouragement. I would also like to thank Dr. Mauro Maggioni for his great instruction, and his encouragement to study signal processing.

I am also thankful for the great advice and instruction of Dr. Jeffery Krolik, Dr. Guillermo Sapiro, Dr. Henry Pfister, Dr. Rebecca Willett, Dr. Surya Tokdar, Dr. Mary Knox, Dr. Tom Witelski, Dr. Mike West, Dr. Bill Allard, Dr. Lillian Pierce, and Dr. Ioanna Manolopoulou.

My thanks also go especially to Dr. Yuan Zhang and Dr. Wenjing Liao for their great friendship, discussion, instruction, collaboration and encouragement. I am also thankful for the help and collaboration from Dr. Qiang Qiu, Dr. Xiuyuan Cheng and Dr. Liang Lu.

I am lucky to have a lot of friends and great people help me. I would like to especially thank Dr. Yingbo Li, Dr. Lingling Tang, Dr. Esteban Vera, Shaobo Han, Dr. Changyou Chen, Dr. James Murphy, Dr. Qiangliang Su, Jiaji Huang, Xin Jiang, Yangbo Xie, Dr. Xuejun Liao, Dr. Xin Yuan, Dr. Pan Wu, Dr. Guang Yang, Yuncheng Pu, Yichuan Zhao, Nan Yi, Chunyuan Li, Wanyi Fu, Kai Fan, Dr. Granger Hickman, Dr. Andrew Harms, Dr. Li Li, John Soli, Juan Ramirez Jr., Dr. Mengqiang Yuan, Qiwei Zhan, Dr. Jiazhou Liu, Dr. Tingran Gao, Jieren Xu, Chi Xu, Qian Gong, Leon Cai, Kathy Peterson, Ellen Currin and Edith Allen.

I am thankful for all the challenges that the Lord has placed on me. They constantly shape my character, and let me see the blessings through the difficulties. Lord, may you continue to be the light of my path and let me live out your glory. May you continue to write my story and testimony.

1

Introduction

This dissertation focuses on detection and classification of whale vocalizations in the time-frequency plane. We will first introduce the background and motivation to study whales in Section 1.1, and basic knowledge of whale vocalizations in Section 1.2. We will then briefly go through the time-frequency representations on signals in Section 1.3. The review of current works on whale acoustic signals detection and classification will be in Section 1.4 and Section 1.5 respectively. We present the outline and contributions of the dissertation in Section 1.6. The contributions of the dissertation, in short, are using linear and nonlinear methods to analyze the time-frequency characteristics and acoustic structures of whale vocalizations in order to improve detection and classification.

1.1 Background and Motivation

There are 129 species of marine mammals [1], all of which rely on the aquatic environment for feeding [2]. Marine mammals can be divided into four recognized groups; cetaceans, pinnipeds, sirenians, and fissipeds [3]. Whales, dolphins, and porpoises belong to the group of cetaceans. There are two suborders of them: mysticetes

and odontocetes. Mysticetes, or baleen whales, do not have teeth, and use baleen to filter small prey from sea water. Odontocetes, or toothed whales, have teeth and feed largely on fish and squid [4].

Studies of whales can help understand the ocean and better protect the environment. Human activities have heavily affected the ocean environment. Many whales have been killed because of vessel collision, hunting, and fishing. Sound pollution-such as military sonar, commercial vessel traffic noise, and seismic exploration from oil industry can have a significant impact on whales. Sadly, some cetaceans have even become extinct, and it is thought that sound pollution could be a contributing factor [5].

Whales have made contributions to national security and economic development. Their sense of hearing is more well-developed than that of humans. Research has shown that whales can translate sounds into mental images, and distinguish brass, aluminum and stainless steel in the mud. They have been trained to hunt for mines so that the acoustic transponders can be dropped nearby [6]. They have also been used by the US Navy to patrol and lookout for the enemy who try to infiltrate the submarine base. During the Vietnam War, dolphins patrolled Cam Ranh Bay to prevent attack from the Vietnam army [6].

1.2 Whales Vocalizations

1.2.1 Sound Production

The sound production of whales is different from that of humans. Humans produce sound by the vibration of vocal cords in the larynx. The sound is also affected by the shape changes of tongue, lips and oral cavity (mouth) [7]. The sound productions of toothed whales and baleen whales are different. Toothed whales produce sound through the phonic lips, which function like the human nasal cavity. The

vibration of phonic lips produces clicks, whistles and burst pulses [8].

Baleen whales use the larynx for sound production. By contracting muscles in the chest and the laryngeal sac in the throat, the whales can change the frequency and amplitude of sound [9, 10].

Whales can produce sounds by slapping the water with their body. For example, humpback whale and bottlenose dolphins can slap their tail to produce a broadband sound (30-12kHz) [11].

1.2.2 Types and Purposes

Toothed whales are able to produce echolocation clicks at frequency ranged from 0.2 to 150kHz for navigation and food hunting. The higher frequencies can reveal detail information of the target, while the lower frequencies are likely used for distance communication. Echoes from clicks can convey the size, shape, speed, texture and direction of the target [12].

In addition to clicks, most toothed whales can produce whistle type sound, which are used for communication. Bottlenose dolphins, for example, make a wide variety of whistles to distinguish and recognize each other. Each dolphin also makes its own unique sound called a "signature whistle". Scientists believe that these signature whistles may be used to identify itself to pod mates [13].

Baleen whales produce primarily low frequency sounds - mostly below 5000 Hz. The low frequency sounds, which range from 20 to 200 Hz, are moans, grunts, thumps and knocks. The high frequency sounds, which are above 1000 Hz, are whistles, cries and songs. Humpback whales are able to generate a series of complex repeated units of sounds, which together is called a song. The song can last for up to 20 minutes. It is produced by the male during the mating season to attract females [14].

1.2.3 Signal Model

We now turn to the crucial question of how to model whale vocalizations mathematically. The whale sounds can be considered to carry information, therefore they can be modeled as amplitude modulated (AM) and frequency modulated (FM) signals [4, 15]. They can be simply represented as [4, 16]:

$$s(t) = A(t)\cos(2\pi \sum_{m=0}^{N} a_m t^m) \tag{1.1}$$

where $A(t)$ is the amplitude, $\sum_{m=0}^{N} a_m t^m$ is a polynomial phase, and $\{a_m\}_{m=0}^{N}$ are the polynomial coefficients. Signals that has the form like eq. (1.1) are known as polynomial phase signals. For quadratic phase signals ($m = 2$), a_0 is the start frequency, a_1 is the slope, and a_2 is the curvature [16].

For complex whale vocalizations, such as humpback whale song, we can use piecewise polynomial phase model like the music signals [17]:

$$s(t) = \sum_i s_i(t) = \sum_i \sum_q A_{i,q}(t)\cos(2\pi \sum_{m=0}^{N} a_{i,q,m} t_{i,q}^m) \tag{1.2}$$

where $A_q(t)\cos(2\pi \sum_{m=0}^{N} a_{q,m} t_q^m)$ is the q-th sinusoid or harmonic at time t, and $s_i = \sum_q A_{i,q}(t)\cos(2\pi \sum_{m=0}^{N} a_{i,q,m} t_{i,q}^m)$ is the signal at time frame i. Many bioacoustics detection and classification studies are based on formula eq. (1.1) and eq. (1.2).

1.3 Time-Frequency Representations

Many whale vocalizations are non-stationary signals, in order to study frequency properties at a particular time, or to study time properties at a particular frequency, we need the time-frequency techniques to analyze such signals [18]. Methods such

as the short time Fourier Transform (STFT) [19], the wavelet transform [20, 21, 22], the ambiguity function [23, 24] and the Wigner-Ville distribution (WVD) [25, 26], the chirplet transform [27] and the Empirical Mode Decomposition (EMD) [28] have been used in this field [4, 29, 30, 31]. We present briefly the STFT, the wavelet transform, the ambiguity function and the WVD in the following paragraphs.

The STFT has been used extensively for analyzing speech, music and other non-stationary signals. The STFT is to take the Fourier transform of a window segment of a signal, as the window slides in time. From the STFT, we can generate a spectrogram to examine the energy distribution of the signal in the time frequency plane. Spectrograms are among the simplest and natural tools to analyze AM-FM types signals [18, 32]. It can sparsely represent signals' energy ribbons localized along time-frequency trajectories [33]. Examples of spectrograms of whale vocalizations from Mobysound [34] are shown in Figure 1.1 and Figure 1.2.

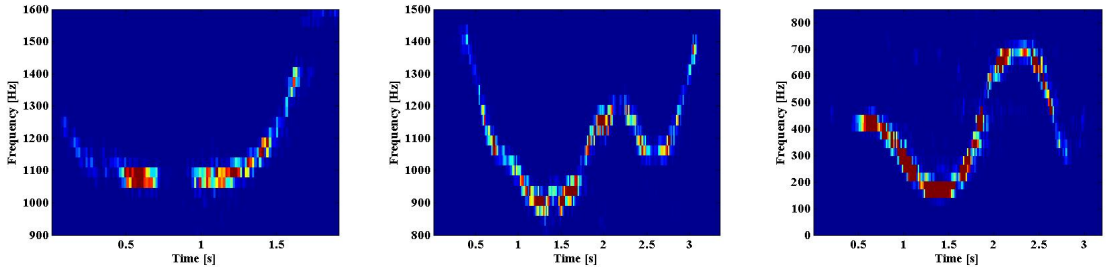


FIGURE 1.1: Spectrogram of bowhead whale signals

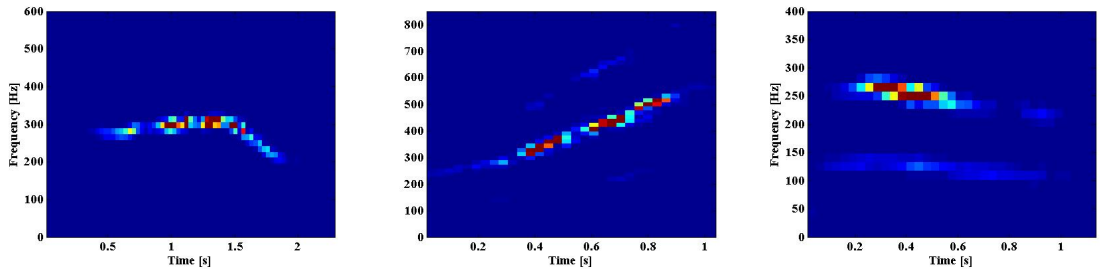


FIGURE 1.2: Spectrogram of humpback whale signals

The wavelet transform is another time-frequency representation, or more appro-

priately, time-scale representation. The wavelet transform can be expressed as an inner product of signals with a family of translates and dilates of a mother wavelet, or a wavelet function. The difference between the wavelet transform and the STFT is that the STFT uses a single analysis window, while the wavelet uses short windows at high frequencies and long windows at low frequencies. Therefore, the wavelet transform has a coarser frequency resolution at high frequencies, and a finer frequency resolution at low frequencies; while the STFT has the same frequency resolution at all frequencies. We can generate a scalogram to examine the energy distribution of signals by using the wavelet transform.

The WVD and the ambiguity function are among the Cohen's class of bilinear time-frequency energy distribution [18]. The WVD and the ambiguity function have a better energy localization in the time-frequency plane compared with spectrograms, but they have confusing artifacts for multicomponent signals [18, 32]. To attenuate the artifacts, we can apply a kernel function to the ambiguity function and the WVD. Distributions such as the pseudo Wigner Ville distribution, the Born-Jordan distribution and the Choi-Williams distributions are techniques which incorporate kernel functions into the ambiguity functions and the WVD [18, 35].

We perform detection and classification on whale vocalizations based on these time frequency representations.

1.4 Detection

By detection, we refer to monitoring, tracking and identifying whales. Detecting the occurrence of whales can give us a better understanding of their living pattern, the ocean environment and underwater sound propagation. Passive Acoustic Monitoring (PAM) is essential to detect marine mammals. PAM is based on listening to whale vocalizations without interfering with their behaviors. With the use of PAM, marine

biologists have been able to listen to the sounds of marine mammals, and track and locate them [36, 37, 38].

In order to successfully identify whale vocalizations, we have to understand the ocean acoustics propagation and some important relevant signal processing techniques. The following constitutes a brief review on these topic.

1.4.1 Ocean Acoustics Propagation

Information transmission in the ocean occurs primarily via acoustic waves. Sound speed is the major physical characteristic affecting acoustic propagation through the ocean. Sound speed in the ocean is influenced by temperature, salinity and pressure of the sea water [39, 40, 41]. Because temperature, salinity and pressure vary temporally and spatial, the sound speed also varies. The speed of sound changes with depth, yielding what is known as a sound speed profile [40]. The spatial variability gives rise to refraction of the sound. Refraction and reflection from the sea surface and bottom contribute to multipath propagation [40]. Multipath ocean propagation leads to dispersion and time spread of transmitted signals [42], which is a particular challenge for whale acoustic detection.

Propagation of sound through water is described by the wave equations, with appropriate boundary conditions. A number of models, such as the ray theory model and the normal mode model have been developed to simplify propagation calculations [39, 43]. The ray theory model assumes the potential function, or the propagated signal, to be the product of a phase function and a amplitude function. In simulation, we can use the Bellhop ray tracing model [44] to obtain the amplitude and phase, and from which we obtain the propagated signal. The normal mode model assumes the potential function to be the product of a depth function and a range function. In simulation, we use KRAKEN [45] to obtain the potential function.

In this dissertation, we evaluate the multipath propagation environment effect on

detection in the time-frequency plane. To make the problem trackable, we use the Bellhop ray tracing model to obtain transmitted signals. Experimental results show that when the signal to noise ratio (SNR) is low, detection based simply on cross correlation coefficients between the target signal's spectrograms and the predefined signal template is not desirable, because of multipath dispersion effect. To improve the detection performance, we need to have a probability distribution of transmitted signals' spectrograms. Section 1.6 and Chapter 3 have more details on improving multipath signal detection.

1.4.2 Detection Techniques in the Time Frequency Plane

Time-frequency techniques are widely applied in whale vocalizations detection. Lopatka *et al*, Adam, and Adam *et al* used the STFT, the wavelet transform, the chirplet transform and the EMD transform to analyze whale vocalizations and perform detection [31, 46, 47].

Many detection methods have been proposed to detect marine mammals based on spectrograms, Mellinger and Clark proposed the spectrogram correlation [48, 49], Gillespie proposed frequency contour edge detection method [50], and Mellinger *et al*, Buck and Tyack, Madhusuhana *et al*, Mallawaarachchi *et al* used extracted contours from whale calls' spectrograms for detection [51, 52, 53, 54]. The idea behind these methods is to form a template of the target whale vocalization in the time-frequency plane, and cross correlate received signals with the predefined template.

However, as stated above, multipath propagation environment will give rise to dispersion to transmitted signals. Detectors based on cross correlation coefficients of signals and a fixed template, without considering multipath dispersion effect, are likely having false alarms or miss detections. When the SNR is low, or the source of whale vocalizations is far from receivers, performance of detectors based simply on spectrograms is far from desirable.

In addition, spectrograms neglect phase information. Phase information is particularly useful for signal reconstruction and source localization [55]. Studies have found that the phase spectrum is more sensitive than the magnitude spectrum for signal recognition when the window function of the STFT is appropriately chosen [56, 57]. By incorporating phase information and magnitude information, we can improve detection performance.

In this dissertation, we quantify the loss of detection performance of the spectrogram based detectors, as well as detectors' sensitivity on uncertain ocean environment in different SNR.

1.5 Classification

After signals are detected, the next step is often to perform a classification process on the signals. By classification, we refer to the problem of identifying the category to which a new observation belongs; that is mapping the input data to a category. Signal representation, or feature extraction, is an important aspect in classification. Effective signal representation can reduce data dimension, and reduce computational complexity, and consequently avoid "curse of dimensionality" [58]. We can use parametric models and non-parametric models for signal representation. We first talk about parametric models, then non-parametric models.

Given a signal $s(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$, a parametric model represents $s(\mathbf{x})$ as:

$$s(\mathbf{x}) = f(\mathbf{x}|\theta_i, i = 1, \cdots, K)$$

where $\{\theta\}_{i=1}^K$ are the parameters to be fitted, and where $f(\cdot)$ is the mapping from the K-dimensional parameter space to the signal manifold. Broadly speaking, the aim is to find the best possible parameter fit for $\{\theta\}_{i=1}^K$, that is:

$$\min_{\{\theta_i\}} ||s(\mathbf{x}) - f(\mathbf{x}|\theta_i)||. \tag{1.3}$$

We can also map the signal $s(\mathbf{x})$ to another space $T(s(\mathbf{x}))$, where $T(\cdot)$ is the transformation chosen to highlight certain patterns or structure in the data. The transformation $T(\cdot)$ can be linear or nonlinear. We then compute

$$\min_{\{\theta_i'\}} ||T(s(\mathbf{x})) - g(\mathbf{y}|\theta_i')||. \tag{1.4}$$

where g is a function, and $\mathbf{y}$ are the points defined in the space of $T(s(\mathbf{x}))$. Many time-frequency representation techniques such as the STFT, the wavelet transform, the WVD and the chirplet transform are examples of the transformation.

Non-parametric models do not seek to fit for $\{\theta\}_{i=1}^{K}$, instead, they exploit signal geometric and topological information for representation. Manifold mapping is one of the examples. It computes a low dimensional embedding of high dimension data from an underlying manifold for decision making [59]. Manifold mappings can be linear and nonlinear. The principal component analysis (PCA) [60, 61] and multidimensional scaling (MDS) [62] are linear mapping methods. For nonlinear mappings, the low dimensional embedding lets the nearby points in the high dimensional space remain nearby [59]. Therefore, the nonlinear mappings can preserve geometric properties of nearest neighbors. Popular nonlinear manifold mappings includes Local linear Embedding (LLE) [63], Laplacian Eigenmap [64], ISOMAP [65] and diffusion map [66].

The following gives a brief review on signal representation techniques in whale vocalizations classification.

1.5.1 Classification using the Time Frequency Analysis Techniques

In whale acoustic classification, many time-frequency representation techniques have been applied to perform feature extraction. O'Neil *et al* used chirplet transform for whale acoustic signals representation [67], then Bahoura and Simard, Bahoura *et al* used the STFT, the wavelet transform and the chirplet transform for whale vo-

calizations classifications [29, 30].

In acoustic signal processing, the MFCC (Mel Frequency Cesptral Coefficients) is one of the most popular features for speech recognition and music classification [68]. The process of MFCC is to project and bin the spectrogram of a signal according to a log frequency (Mel) scale. MFCC has been applied as a feature to whale vocalizations classification. Pace *et al* and Roch *et al* used MFCC as features then applied Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM) for classification [69, 70, 71]

The MFCC involves first order frequency information alone, and therefore gives no direct information about the higher order coefficients of polynomial phase signals. Therefore, it should be possible to design a better feature set for whale vocalizations.

Because whale sounds can be modeled as polynomial phase signals, we can use the estimated polynomial coefficients in the time-frequency plane for representation. We can also exploit the geometry of the data and perform nonlinear mapping to the manifold parameterized by polynomial coefficients $\{a_m\}_{m=0}^{N}$. By doing this, we can use fewer parameters for data representation. Details on whale data representation can be found in Section 1.6, and Chapters 4 - 6.

Deep learning, a current popular topic in machine learning community, has been applied in whale vocalizations classification recently [72, 73]. The effectiveness of multi-layer convolutional networks for learning features has been established in audio signal processing, such as speech recognition and music classification [74, 75, 76, 77]. The idea of a convolutional network is to convolve signals with filters, and use the obtained features for classification. In order to explore the time-frequency content of audio signals, Andén and Mallat have proposed the scattering transform [78] for audio signal. Although it demonstrated some success upon a music genre dataset and some speech datasets, more testing or more improvement are needed for this algorithm.

1.5.2 Manifold Mapping

The manifold mapping approach involves exploiting the geometrical information of the signal for representation. It is based on the assumption that the high dimensional natural data actually resides on a low dimensional manifold, so fewer degrees of freedom are required to understand such data. The manifold structure of speech sounds stresses the nonlinear relation between articulatory and acoustic space [79]. We can apply alternative distance metrics such as geodesic distance and diffusion distance along these manifold to represent the acoustic signals. Algorithms such as LLE, ISOMAP, Laplacian Eigenmap and diffusion maps have been shown to be successful in the context of speech signal processing [64, 79, 80, 81]. By applying nonlinear manifold mapping methods to whale vocalization data and examining the eigenvalue distributions, we can estimate the intrinsic dimension of the data, which relates to the number of polynomial phase coefficients of eq. (1.1) and eq. (1.2). The nonlinearity of the data structure can be justified by comparing classification accuracy of the nonlinear manifold mapping methods with the linear mapping methods, such as PCA and MDS, on the whale data. Manifold mapping methods can be viewed as an alternative to parametric methods.

1.6 Outline and Contribution

The remaining chapters are organized as follows:

Chapter 2 presents detailed background knowledge on time-frequency methods, such as the STFT, the ambiguity function and the WVD. The interpretations and characteristics of these methods are essential for the contributions of this dissertation. We also present basic formula and concepts on hypothesis testing and ocean propagation models.

Chapter 3 is about underwater passive acoustic detection. It gives the probability

distribution, likelihood ratio and acoustic detection bound based on spectrograms and the STFT under the multipath propagation ocean environment. We found that by retaining the phase information in detection, the detection performance will be improved, while being sensitive to environmental changes, compared with detectors based on spectrograms. Detailed derivations and proof of probability distribution of detection based on the STFT can be found in Appendix A.

Chapters 4 - 6 are about whale vocalizations classification. In Chapter 4, we apply the Weyl transform to whale vocalizations, and obtained from it global features which capture chirp rate information of signals. Experimental results show that our proposed feature sets outperform both the MFCC and the chirplet transform in the Duke marine lab vocalization dataset.

We study the intrinsic structure of whale vocalizations in Chapter 5. We apply manifold mapping techniques of PCA, MDS, ISOMAP and Laplacian Eigenmap on whale data. The nonlinear methods can well represent the data in a lower dimension, and well capture the physical information of whale calls. This observation strongly suggests that the structure of whale vocalization is nonlinear rather than linear. This chapter gives a geometric interpretation to the Weyl feature set in Chapter 4.

We explore deep learning algorithms on whale vocalizations in Chapter 6. We associate convolutional network with filterbanks in the STFT, and apply multilayer PCA filterbanks (PCANet) and propose multilayer DCT filterbanks (DCTNet) for acoustic signals feature extraction. For the DCTNet, each layer has different time-frequency scale representation, and can reveal different physical content of signals. The PCANet, on the other hand, can filtered out noise in each layer. Experimental results show that using both PCANet and DCTNet features can achieve high classification rate in the whale vocalization data, and the DCTNet feature achieves state-of-art performance. The word error rate of the DCTNet feature is similar to the MFSC in speech recognition tasks, suggesting that the convolutional network is

able to reveal acoustic content of speech signals.

Chapter 7 gives conclusion of this dissertation and future research direction.

To summarize, the contributions of this dissertation are as follows:

- We quantify the loss of detection performance of spectrograms based detectors, and evaluate their sensitivity on uncertain ocean environment in different SNR. We also theoretically and experimentally evaluate the detection performance and the environmental sensitivity in the time-frequency plane of detector with phase information (the STFT detector), which can achieve optimal detection performance when noise is additive white gaussian, and the environmental parameters and signals are known exactly (Chapter 3).

- We apply Weyl transform for acoustic feature extraction, and represent the whale vocalizations with two-dimensional feature set, which has the chirp rate and base frequency information (Chapter 4). The fact that we can use two dimensional feature set to represent whale vocalizations is related to its intrinsic structure and intrinsic dimension. We apply linear mappings such as PCA and MDS, and nonlinear mappings such as Laplacian Eigenmap and ISOMAP to the data, and examine their eigenvalue distributions, the clustering effect, and classification accuracy. The results suggests the nonlinearity of whale vocalizations (Chapter 5).

- We use PCANet and propose DCTNet for acoustic feature extraction. We associate filterbanks for convolutions in deep learning with filterbanks in the STFT, and show that DCTNet are essentially multi-level time-frequency representations, revealing different physical acoustic content. (Chapter 6).

2

Background

2.1 Time-Frequency Representations

In this sections, we will go through the definitions of the Short Time Fourier Transform, the spectrogram, the wavelet transform, the chirplet transform, the ambiguity function and the Wigner-Ville distribution. The Short Time Fourier Transform, the wavelet transform and the chirplet transform are linear. The spectrogram, the ambiguity function and the Wigner-Ville distribution are among the Cohen's class bilinear or quadratic representations.

2.1.1 Short Time Fourier Transform

1. Definition and Interpretations

The Short Time Fourier Transform (STFT) of a function s with respect to g is defined as [82]:

$$V_g s(x,\omega) = \int_{\mathbb{R}^d} s(t)\overline{g(t-x)}\exp(-jt\omega)dt, \quad \text{for} \quad x,\omega \in \mathbb{R} \tag{2.1}$$

The discrete time STFT is defined as [83]:

$$S(n,\omega) = \sum_{k=-\infty}^{\infty} s(k)g(n-k)\exp(-j\omega k). \tag{2.2}$$

where $s(n)$ in this case is an acoustic sequence. It represents the Fourier transform of a window segment of the acoustic waveform as the window $g(n)$ slides in time.

Eq. (2.2) can be written as:

$$S(n.\omega) = [s(n)\exp(-j\omega n)] * g(n) \tag{2.3}$$

where $*$ denotes convolution. In this case, the STFT can be interpreted as modulating the signal $s(n)$ with $\exp(-j\omega n)$, and convolving it with the window $g(n)$. The modulated signal passes through a low pass filter. The process is shown in Figure 2.1 [83].

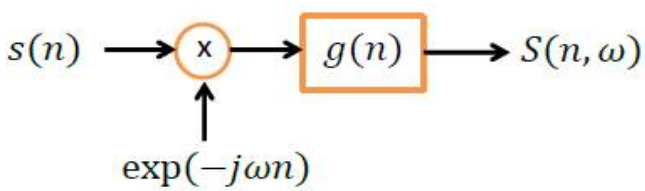


FIGURE 2.1: System evaluating Short Time Fourier Transform

Eq. (2.2) can also be written as:

$$S(n,\omega) = \exp(-j\omega n)\{s(n) * (g(n)\exp(-j\omega n))\} \tag{2.4}$$

In this case, the STFT can be interpreted as the signal passes through a bandpass filter. The process is shown in Figure 2.2 [83].

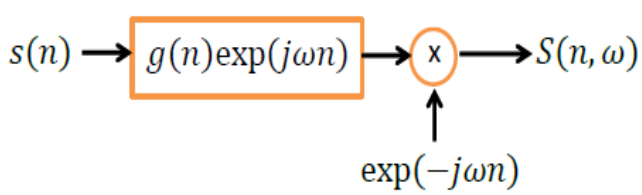


FIGURE 2.2: Alternative system evaluating Short Time Fourier Transform

When we sample the angular frequency ω, that is $\omega = 2\pi r/N, \;\; r = 0, 1, \cdots, N-1$, we can interpret the STFT as the signal pass through the filterbank, as shown in Figure 2.3 [83], and eq. (2.4) becomes:

$$S(n,r) = \exp(-j\frac{2\pi r}{N}n)\{s(n) * (g(n)\exp(-j\frac{2\pi r}{N}n))\} \tag{2.5}$$

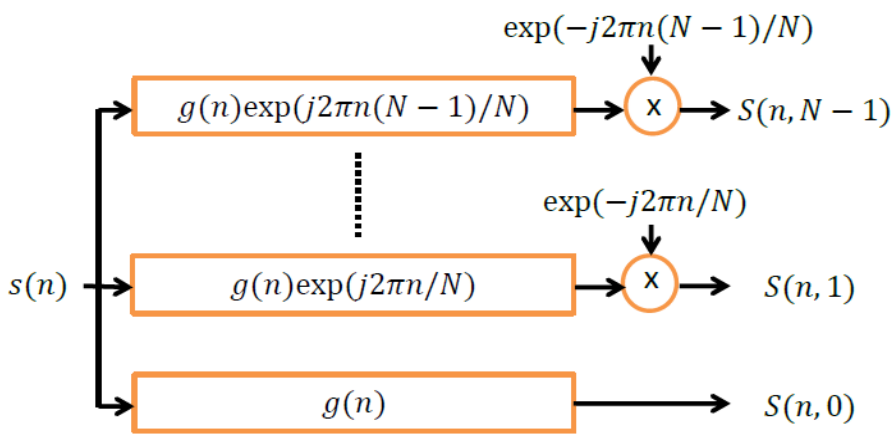


FIGURE 2.3: Filterbank interpretation of Short Time Fourier Transform

2. Matrix Form

Suppose the length of the signal is N, the discrete Fourier transform (DFT) can be written as:

$$\tilde{S}(r) = \sum_{r=0}^{N-1} s(n)\exp(-j2\pi rn/N) \tag{2.6}$$

In matrix form, it can be written as:

$$\begin{bmatrix} \tilde{S}(0) \\ \tilde{S}(1) \\ \vdots \\ \tilde{S}(N-1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix} \begin{bmatrix} s(0) \\ s(1) \\ \vdots \\ s(N-1) \end{bmatrix} \tag{2.7}$$

where ω is as defined before: $\omega = \exp(-j2\pi rn/N)$. We can write eq. (2.7) simply:

$$\tilde{\mathbf{S}} = \mathbf{Fs} \tag{2.8}$$

where, $\mathbf{F}$ is the Fourier matrix, and

$$\tilde{\mathbf{S}} = [\tilde{S}(0), \tilde{S}(1), \cdots, \tilde{S}(N-1)]^T$$

$$\mathbf{s} = [s(0), s(1), \cdots, s(N-1)]^T$$

For the STFT, supposed that the window size is M, and we take M point DFT, Eq. (2.2) can be written as [84]:

$$S(n,r) = \sum_{m=0}^{M-1} s(m+n)g(m)\exp(-j2\pi rm/M) \tag{2.9}$$

Let $x(m) = s(m+n)g(m)$, Eq. (2.9) becomes [84]:

$$S(n,r) = \sum_{m=0}^{M-1} x(m)\exp(-j2\pi rm/M) \tag{2.10}$$

In matrix form, it can be written as:

$$\begin{aligned}\mathbf{S} &= \begin{bmatrix} S(0,0) & S(1,0) & \cdots & S(N-M,0) \\ S(0,1) & S(1,1) & \cdots & S(N-M,1) \\ \vdots & \vdots & \ddots & \vdots \\ S(0,M-1) & S(1,M-1) & \cdots & S(N-M,M-1) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{M-1} & \cdots & \omega^{(M-1)(M-1)} \end{bmatrix} \begin{bmatrix} x(0) & x(1) & \cdots & x(N-M) \\ x(1) & x(2) & \cdots & x(N-M+1) \\ \vdots & \vdots & \ddots & \vdots \\ x(M-1) & x(M) & \cdots & x(N-1) \end{bmatrix}\end{aligned} \tag{2.11}$$

The STFT is invertible, that is, the original signal can be recovered from the inverse STFT. We can apply the overlap-add method to do the inversion.

3. Spectrogram

We can generate the spectrogram to examine the power spectrum of the short time fourier transform. Using formula eq. 2.2, we have [82]:

$$\mathrm{SPEC}_g s(x,\omega) = |V_g s(x,\omega)|^2 \tag{2.12}$$

Examples of spectrograms can be found in Figure 1.1 and Figure 1.2 in chapter one.

2.1.2 Wavelet Transform

The continuous wavelet transform (CWT) projects a signal $s(t)$ on a family of zero-mean functions (the wavelets) [35, 85]:

$$WT(x,a) = \frac{1}{\sqrt{a}} \int_{\mathbb{R}^d} s(t)\psi(\frac{t-x}{a})dt \tag{2.13}$$

where $\psi(t)$ is called the "mother" wavelet, and a is a scale factor.

The difference between the wavelet transform and the STFT is that the bandwidth and duration of the wavelet change when the scale factor a is changed, while the STFT uses the same window function. The CWT uses short windows at high frequencies and long windows at low frequencies.

From the wavelet transform one can obtain a energy density in the time-scale plane, which is the scalogram [85]:

$$P_{WT}(x,a) = \frac{1}{2\pi C a^2} |WT(x,a)|^2 \tag{2.14}$$

where C is chosen so that the energy of the scalogram is identical to the signal energy.

2.1.3 Chirplet Transform

The chirplet transform can be viewed as a generalization of the wavelet transform. While modulated windows of the STFT and wavelets of the wavelet transform can be regarded as "portions of waves", chirplets can be regarded as "portions of chirps" [86]. Wavelets provides a tiling of the time-frequency plane with tiles that lined up with the time-frequency axis. Chirplets have a more general tiling of the time-frequency plane because the tiles can be rotate and shear. For the Gaussian chirplet, it is defined as [86]:

$$g_{t_c,f_c,log(\Delta_t),c}(t) = \frac{1}{\sqrt{\sqrt{\pi}\Delta_t}} \exp(-\frac{1}{2}(\frac{t}{\Delta_t})^2) \exp(j2\pi(c(t-t_c)^2 + f_c(t-t_c)))$$

where t_c is the time-center, f_c is the frequency center, Δ_t is the duration, and c is the chirp rate.

2.1.4 Ambiguity Function

The ambiguity function is an analytical tool for waveform design and analysis. The ambiguity a function of s ($s \in L^2(\mathbb{R}^d)$) is defined as [82]:

$$\begin{aligned} As(x,\omega) &= \int_{\mathbb{R}^d} s(t+\frac{x}{2})\overline{s(t-\frac{x}{2})}\exp(-jt\omega)dt \\ &= \exp(jx\omega/2)\cdot V_s s(x,\omega). \end{aligned} \tag{2.15}$$

We can see that it is symmetry:

$$(As)^*(x,\omega) = \overline{As(-x,-\omega)} = As(x,\omega) \tag{2.16}$$

so it is always real valued.

Some books in engineering define it as [87]:

$$As(x,\omega) = \int_{\mathbb{R}^d} s(t)\overline{s(t-x)}\exp(-jt\omega)dt = V_s s(x,\omega). \tag{2.17}$$

The cross ambiguity function of s and $g \in L^2(\mathbb{R}^2)$ is:

$$\begin{aligned} A(s,g)(x,\omega) &= \int_{\mathbb{R}^d} s(t+\frac{x}{2})\overline{g(t-\frac{x}{2})}\exp(-jt\omega)dt \\ &= \exp(jx\omega/2)\cdot V_g s(x,\omega). \end{aligned} \tag{2.18}$$

Most properties of STFT carry over to the ambiguity function. For more information of ambiguity function and its application, we can refer to Richards' *Fundamental of Radar signal processing* [87].

2.1.5 Wigner-Ville Distribution

The Wigner-Ville Distribution (WVD) of a function $s \in L^2(\mathbb{R}^d)$ is defined as [82]:

$$Ws(x,\omega) = \frac{1}{2\pi}\int_{\mathbb{R}^d} s(x+\frac{t}{2})\overline{s(x-\frac{t}{2})}\exp(-j\omega t)dt \tag{2.19}$$

The cross WVD of function s and $g \in L^2(\mathbb{R}^d)$ is defined as:

$$W(s,g)(x,\omega) = \frac{1}{2\pi}\int_{\mathbb{R}^d} s(x+\frac{t}{2})\overline{g(x-\frac{t}{2})}\exp(-j\omega t)dt \tag{2.20}$$

The relation of the WVD and the ambiguity function can be established though the characteristic function of the WVD [18]:

$$
\begin{aligned}
M(\theta,\tau) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp(j\theta x + j\tau\omega) W s(x,\omega) dx d\omega \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp(j\theta x + j\tau\omega) s(x+\frac{t}{2})\overline{s(x-\frac{t}{2})}\exp(-j\omega t) dt dx d\omega \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \exp(j\theta x)\delta(\tau - t) s(x+\frac{t}{2})\overline{s(x-\frac{t}{2})} dt dx \\
&= \int_{\mathbb{R}^d} s(x+\frac{\tau}{2})\overline{s(x-\frac{\tau}{2})}\exp(j\theta x) dx \\
&= As(\theta,\tau)
\end{aligned}
$$

1. Pseudo Wigner Ville Distribution

When we want to emphasize the properties near the time of interest compared to the far away time, we can apply a function $h(t)$ to the WVD. t is called the lag variable. The WVD is the Fourier transform with respect to t of the quantity $s(x+\frac{t}{2})\overline{s(x-\frac{t}{2})}$. The pseudo Wigner Ville Distribution, which emphasize the signal around time x, is defined as:

$$
W_p s(x,\omega) = \int_{\mathbb{R}^d} h(t) s(x+\frac{t}{2})\overline{s(x-\frac{t}{2})}\exp(-j\omega t) dt \tag{2.21}
$$

2. Modified Wigner Ville Distribution

In order to have positive distribution (the WVD can be negative), we can smooth the WVD by convolving with a smoothing function $L(x,\omega)$, that is:

$$
W_m s(x,\omega) = \int_{\mathbb{R}^d} L(x-x',\omega-\omega') W s(x',\omega') dx' d\omega' \tag{2.22}
$$

2.1.6 Cohen's Class

The Cohen's class is the family of time-frequency energy distributions covariant by translations in time and frequency [18, 35]. All time frequency representations can be obtain from the following formula [18]:

$$C(\tau,\omega) = \frac{1}{4\pi}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(\theta,\tau)s(x+\frac{t}{2})\overline{s(x-\frac{t}{2})}\exp(j\theta x - j\tau\omega - j\omega t)dtdxd\theta \tag{2.23}$$

where $\phi(\theta,\tau)$ is a two dimensional function called the kernel. Common bilinear signal representations such as the spectrogram, the WVD, the Choi-Williams distribution [88] and the Born-Jordan distribution [89] can be expressed through eq. (2.23) [18].

2.2 Hypothesis Testing

The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis H_0 and the alternative hypothesis H_1. Let Θ_0 be some subset of the parameter space and Θ_0^c be its complement. If θ denotes a population parameter, the general format of the null and alternative hypotheses is: $H_0 : \theta \in \Theta_0$ verses $H_1 : \theta \in \Theta_0^c$. A hypothesis testing is a rule that determine the sample values that accept H_0 as true, and the samples values that reject H_0 and accept H_1. The subset that reject H_0 is called the rejection region or critical region [90]. Typically, a hypothesis test is specified in terms of test statistic, which is a function of sample.

The likelihood ratio method is related to the maximum likelihood estimators. Let $X_1, \cdots, X_n$ be a random sample from a population, and let the probability of accepting H_0 be $p_{\mathbf{x}|H_0}(\mathbf{X}|H_0)$, and the probability of accepting H_1 be $p_{\mathbf{x}|H_1}(\mathbf{X}|H_1)$.

The likelihood ratio $\Lambda(\mathbf{X})$ is defined as [91]:

$$\Lambda(\mathbf{X}) = \frac{p_{\mathbf{x}|H_1}(\mathbf{X}|H_1)}{p_{\mathbf{x}|H_0}(\mathbf{X}|H_0)} \tag{2.24}$$

The likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\mathbf{X} : \Lambda(\mathbf{X}) \leqslant c\}$, where $0 \leqslant c \leqslant 1$.

For $H_0 : \ \theta \in \Theta_0$ and $H_1 : \ \theta \in \Theta_0^c$, the probability of false alarm P_F (we say the target is present when it is not), and the probability of detection P_D (we say the target is present when it is) are defined as:

$$P_F = \int_{\Theta_0^c} p_{\mathbf{x}|H_0}(\mathbf{X}|H_0) d\mathbf{X} \tag{2.25}$$

$$P_D = \int_{\Theta_0^c} p_{\mathbf{x}|H_1}(\mathbf{X}|H_1) d\mathbf{X}. \tag{2.26}$$

The probability of a miss P_M (miss the target when it is present) is:

$$P_M = \int_{\Theta_0} p_{\mathbf{x}|H_1}(\mathbf{X}|H_1) d\mathbf{X} \tag{2.27}$$

To evaluate the likelihood ratio test, we examine the relation of P_F versus P_D, and generate receiver operating characteristic (ROC) curve. The area under the ROC curve is called the AUC.

2.3 Ocean Propagation Models

The propagation of sound through water is generally described by the three dimensional time dependent wave equation. For most application, a simplified hyperbolic second order time dependent partial differential equation is used [39]:

$$\nabla^2 \Phi = \frac{1}{c^2} \frac{\partial^2 \Phi}{\partial t^2} \tag{2.28}$$

where Φ is the potential function; t is time; c is the sound speed; and where ∇^2 is the Laplacian operator ($\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$). By assuming $\Phi = \phi \exp(-j\omega t)$, where ϕ is the time dependent potential function, and where ω is the source frequency, the wave equation reduces to the Helmholtz equation:

$$\nabla^2 \phi + k^2 \phi = 0, \tag{2.29}$$

where $k = \frac{\omega}{c}$ is the wavenumber.

A number of models, such as the ray theory model and the normal mode model have been developed to simplify propagation calculations [39]. The idea of the ray theory model is to assume ϕ be the product of a pressure amplitude function A and a phase function P, that is $\phi = A\exp(jP)$. Then eq. (2.29) becomes [39, 43]:

$$\frac{1}{A}\nabla^2 A - (\nabla P)^2 + k^2 = 0 \tag{2.30}$$

$$2(\nabla A \cdot \nabla P) + A\nabla^2 P = 0 \tag{2.31}$$

In simulation, we can obtain A and P through the Bellhop ray tracing model [44], from which we can obtain in turn the solution ϕ. The ray theory model is widely used for modeling deep ocean environments.

The normal mode model is derived from an integral representation of the wave equation, and assumes ϕ be the product of a depth function $F(z)$ and a range function $S(r)$, that is $\phi = F(z) \cdot S(r)$. Then eq. (2.29) becomes:

$$\frac{d^2F}{dz^2} + (k^2 - k_r^2)F = 0 \tag{2.32}$$

$$\frac{d^2S}{dr^2} + \frac{1}{r}\frac{dS}{dr} + k_r S = 0 \tag{2.33}$$

where k_r is separation constant. Eq. (2.32) is the depth equation, or normal mode equation, describing the standing wave portion of the solution. Eq. (2.33) is the range equation, describing the traveling wave portion of the solution. The solution of the normal mode equation is known as the Green's function. Assuming a single frequency source, the solution ϕ can be represented as:

$$\phi = \frac{j}{4\rho(z_s)} \sum_{n=1}^{\infty} u_n(z_s)u_n(z)H_0^{(1)}(k_m r) \tag{2.34}$$

where u_n is the normal mode function; z_s is the source depth; z is the receiver depth; ρ is the water density; k_{rn} is the eigenvalue of the separation constant, and where $H_0^{(1)}$ is the zero order Hankel function of the first kind. In simulation, we can use KRAKEN [45] to obtain the solution ϕ. The normal mode model is frequently used in the shallow water case because it can better model the boundary conditions.

3

On Marine Mammal Acoustic Detection Performance Bounds

Since the spectrogram does not preserve phase information contained in the original data, any algorithm based on the spectrogram is not likely to be optimal for detection. In this chapter, we present the Short Time Fourier Transform detector to detect marine mammals in the time-frequency plane. The detector uses phase information for detection. We evaluate this detector by comparing it to the existing spectrogram based detectors for different SNRs and various environments including a known ocean, uncertain ocean, and mean ocean. The results show that this detector outperforms the spectrogram based detector. Simulations are presented using the polynomial phase signal model of the North Atlantic Right Whale (NARW) and the BELLHOP ray tracing propagation model.

3.1 Introduction

Many marine mammal vocalizations are non-periodic and frequency modulated signals [15, 16]. For this reason, the time-frequency analysis techniques are widely applied in analyzing such signals. Many time-frequency representation methods, such as the Short Time Fourier Transform (STFT), the Wigner Ville distribution and the wavelet transform have been used in this field. When we use the STFT, we can generate the spectrogram to examine the energy distribution of the signal in the time-frequency plane. Spectrograms are among the simplest and natural tools to analyze the AM-FM type signals [32]. It can sparsely represent the signals' energy ribbons localized along the time-frequency trajectories [33]. Many detection methods have been proposed to detect marine mammals based on the spectrogram, such as spectrogram correlation [48, 49], frequency contour edge detection method [50] and contour extraction [51, 52, 53, 54]. In order to obtain the optimal detection performance based on the spectrogram, it is natural to examine the probability distribution of the spectrogram data, and obtain the likelihood ratio for detection. Analysis of the probability distribution of the spectrogram elements has been performed to approximate the likelihood ratio of the spectrogram by assuming that the spectrogram elements are statistically independent [92]. Huillery et al [93, 94] had performed detection based on the likelihood ratio of a single spectrogram element.

However, the phase information is neglected when we do detection based on the spectrogram. The phase information is particularly useful for signal reconstruction and source localization [55]. Studies have found that the phase spectrum is more sensitive than the magnitude spectrum for signal recognition when the window function of the STFT is appropriately chosen [56, 57]. By incorporating the phase information and the magnitude information, we can improve the detection performance. In this chapter, we propose the STFT detector. Instead of extracting the phase information

from the source signal separately for detection, we incorporate the magnitude and phase information during the STFT. Results show that the STFT detector outperforms detectors based on the spectrogram, and can achieve optimal detection result when the environmental parameters and source signal are known exactly, and the noise is additive white Gaussian. Technical details of the detector can be found in Section 3.3 and the Appendix.

The sound speed in the ocean is influenced by temperature, salinity, and pressure of the sea water, so it varies spatially [40, 95]. The speed of sound changes with depth, yielding what is known as a sound speed profile. This spatial variability gives rise to refraction of the sound. Refraction and reflection from the sea surface and bottom contribute to multipath propagation [40]. Multipath ocean propagation environment will lead to dispersion and time spread for the transmitted signal [42]. This is a challenge for underwater acoustic communication systems. Shorey, Book, Tantum, Sha, Nolte, and Wazenski and Alexandrou [96, 97, 98, 99, 100] had evaluated the uncertainty of environmental parameters to source localization, and detection in the time domain. In this chapter, we will evaluate the influence of the uncertainty of the sound speed profile on detection in the time-frequency plane. The ROC of the time domain matched filter assuming the signal is known exactly is used to benchmark the detection performance of our proposed detector.

The organization of the chapter is as follows. Section 3.1 gives an introduction and background for this chapter. Section 3.2 presents the ocean acoustics propagation model, environmental parameter settings and propagated signal model. Overviews of the characteristics of the current detection model, and statistics of our proposed STFT detector are presented in Section 3.3. Section 3.4 gives detection performance results using the STFT detector in the matched ocean, uncertain ocean and mean ocean cases. Section 3.5 concludes the chapter.

3.2 Ocean Acoustics Propagation Model

3.2.1 Environment Configuration

We use the Hudson Canyon as the ocean propagation environment. The Hudson Canyon's typical sound speed profile is shown in Figure 3.1. The Hudson Canyon has a sandy bottom. The environmental parameters and uncertainties of the canyon are shown in Table 3.1 [96].

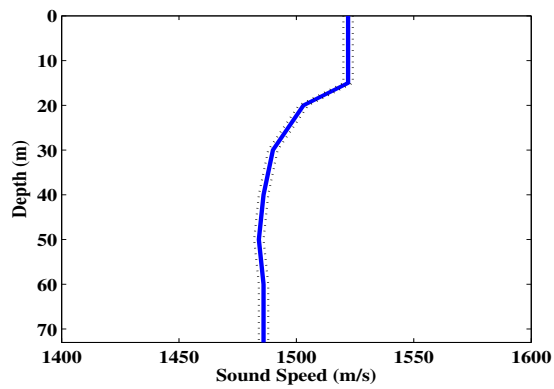


FIGURE 3.1: Hudson Canyon Typical Sound Speed Profile

Table 3.1: Hudson Canyon Environment Parameters and Uncertainties

Description	Depth	Basis sound speed	Uniform uncertainty
Water column sound speed	0+ m	1522 m/s	± 2m/s
	15 m	1522 m/s	± 2m/s
	20 m	1503 m/s	± 2m/s
	30 m	1490 m/s	± 2m/s
	40 m	1486 m/s	± 2m/s
	50 m	1484 m/s	± 2m/s
	60 m	1486 m/s	± 2m/s
	73- m	1486 m/s	± 2m/s
Bottom half-space sound speed	73+ m	1550 m/s	± 50m/s

3.2.2 Signal Model

Many marine mammal vocalizations are frequency modulated, and can be modelled as polynomial-phase signals [16]. Take the North Atlantic Right Whale (NARW)

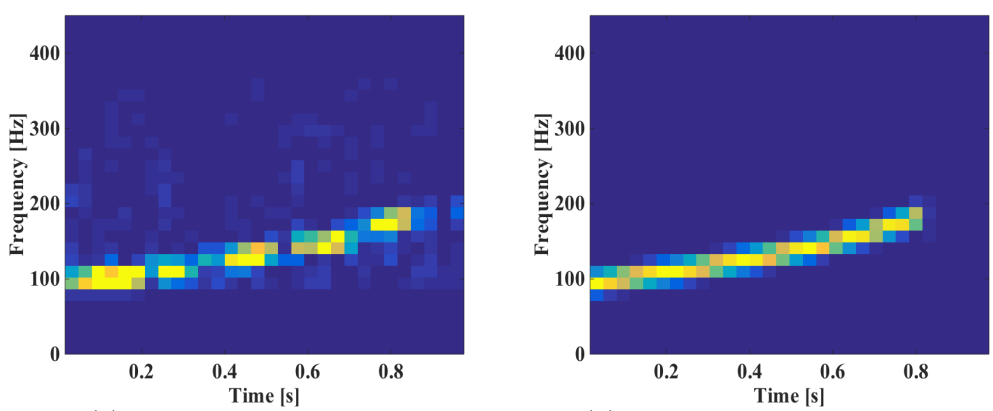
as an example. It is found that there are nine types of sound for the NARW according to its time-frequency characteristics [101], and the upsweep call is commonly found when the whales greet each other.

Assume that $s(t)$ is the source signal of the whale, $\theta(t)$ is the phase of the signal, and $A(t)$ is the amplitude[4]:

$$s(t) = A(t)\cos(\theta(t)) = A(t)\cos\Big(2\pi \sum_{m=0}^{M-1} f_m t^m\Big),\ t = 1, 2, \cdots, N, \tag{3.1}$$

where $\{f_m\}_{m=0}^{N}$ are polynomial coefficient parameters for the signal model.

By inspecting the shape of the whales' sound contour, and fitting the polynomial coefficients, we can synthesize the right whale signal (Figure 3.2). We will use the synthetic NARW signal to evaluate the effect of propagation environment on acoustic detection performance.



(a) Real NARW Contact Call (b) Synthetic NARW Contact Call

FIGURE 3.2: North Atlantic Right Whale Signal

3.2.3 Signal Propagation

Supposed that the whale is z_s meters below the sea surface. We want to determine whether there is a signal present at the receiver, which is r meters from the source

and z_r meters below the sea surface. The environment setting can be characterized as in Figure 3.3.

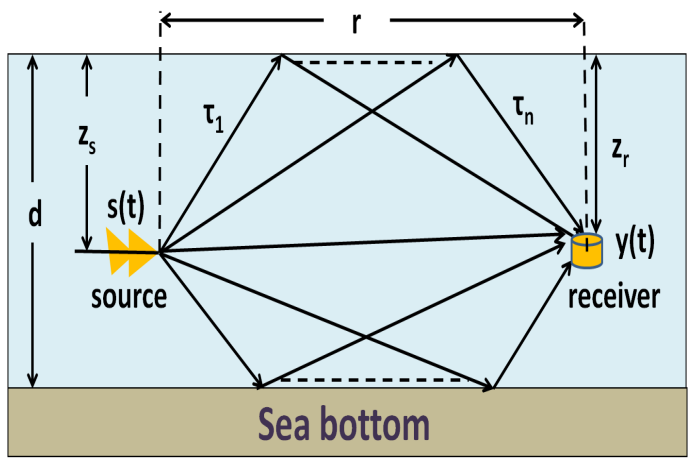


FIGURE 3.3: Ocean Acoustic Propagation Model

z_s: source depth;	z_r: receiver depth;
$s(t)$: source signal;	$y(t)$: propagated signal;
r: range of propagation;	d: depth of shallow water;
τ_1: delay of the first path;	τ_n: delay of the n-th path.

We use the Bellhop Model [44] to calculate the arrival amplitudes and delays of all possible signal paths in the ocean. Supposing there are K arrivals, the propagated signal $y(t)$ can be represented as the sum of K arrival signals. That is:

$$y(t) = \sum_{k=1}^{K} c_k s(t - \tau_k) \tag{3.2}$$

where $s(t)$ is the source signal, and c_k and τ_k are amplitude and delay of each arrived signal respectively. When the amplitudes are complex numbers, which is due, for example, to bottom reflections that introduce a phase shift, the propagated signal is [102],

$$y(t) = \sum_{k=1}^{K} \Big[Re[c_k] s(t - \tau_k) - Im[c_k] s^{+}(t - \tau_k) \Big] \tag{3.3}$$

where $s^+ = \mathcal{H}(s)$ is the Hilbert transform of $s(t)$. The Hilbert transform is a 90 degrees phase shift of $s(t)$ and accounts for the imaginary part of c_k. In this case,

$$s(t) = A(t)\cos\Big(2\pi \sum_{m=0}^{M-1} f_m t^m\Big) \tag{3.4}$$

$$s^+(t) = A(t)\sin\Big(2\pi \sum_{m=0}^{M-1} f_m t^m\Big). \tag{3.5}$$

3.3 Detection Models and Statistics

It is known that when the noise is additive white Gaussian, the time domain matched filter can achieve optimal detection when the source signal and the propagation environment are known exactly [40]. When examining the time-frequency performance of a given signal, we can apply the Short Time Fourier Transform to the signal, and generate the spectrogram to analyze it. The detection process is shown in Figure 3.4.

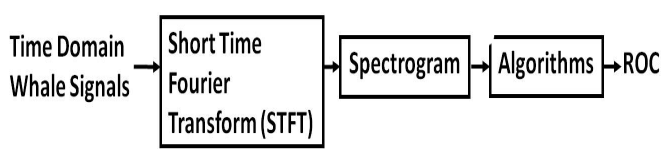


FIGURE 3.4: Detection in time-frequency domain

In this section, we will apply the likelihood ratio of the approximated probability distribution of the spectrogram [92] and present our STFT detector in the matched ocean, uncertain ocean and mean ocean cases. Simulations of detection performance comparison can be found in Section 3.4.

3.3.1 Spectrogram Distribution Detector

1. Matched ocean

We first examine the matched ocean case, that is, the source signal and sound speed profile is known exactly. Our analysis for the matched ocean case is similar to the work of Altes [92].

Let $x(t),\ t = 0, 1, \cdots, N-1$ be the data recorded at the receiver. We want to know whether the data contains the propagated NARW signal. We can form the binary hypothesis test as follows [103]

$$H_0: \quad x(t) = n(t), \tag{3.6}$$

$$H_1: \quad x(t) = n(t) + y(t), \tag{3.7}$$

where $n(t)$ is Gaussian additive noise, $n(t) \sim \mathcal{N}(0, \sigma_n^2)$, and where $y(t)$ is the propagated NARW signal. We apply the STFT to the data and generate the spectrogram. That is [32]

$$S_x[a, b] = \left| \sum_{i=0}^{M-1} x[aD + i] w[i] \exp(-2\pi j b i / M) \right|^2, \tag{3.8}$$

where $w[i]$ is the window function, M is the length of the window, D is the step of the sliding window, and a is the window shift, b is the frequency shift.

The binary hypothesis test based on the spectrogram is

$$H_0: \quad \mathbf{S_x} = \mathbf{S_n}, \tag{3.9}$$

$$H_1: \quad \mathbf{S_x} = \mathbf{S_{n+y}}, \tag{3.10}$$

where $\mathbf{S_x}$ is the spectrogram vector of the data at the receiver, $\mathbf{S_n}$ is the spectrogram vector of pure noise, and where $\mathbf{S_{n+y}}$ is the spectrogram vector of the propagated signal plus noise. Let $J = (\lfloor (N-M)/D \rfloor + 1)$, and $B = \lfloor M/2 \rfloor + 1$, where $\lfloor k \rfloor$ denotes the greatest integer less than or equal to k, the spectrogram vector at the receiver is

$\mathbf{S_x} = [S_x[0,0], \cdots, S_x[J,B]]^T$. For each spectrogram element $S_x[a,b]$ in $\mathbf{S_x}$, under the H_0 hypothesis, we have [90, 92, 93]

$$p(S_x[a,b]) = \frac{1}{M\sigma_n^2} \exp\left(-\frac{S_x[a,b]}{M\sigma_n^2}\right).$$

Under the H_1 hypothesis, we have [90, 92, 93]

$$\begin{aligned} p(S_x[a,b]) =& \frac{1}{M\sigma_n^2} \exp\left(-\frac{S_x[a,b] + m_1^2[a,b] + m_2^2[a,b]}{M\sigma_n^2}\right) \cdots \\ & \times I_0\left(\frac{2\sqrt{(m_1^2[a,b] + m_2^2[a,b])S_x[a,b]}}{M\sigma_n^2}\right), \end{aligned}$$

where

$$m_1[a,b] = \sum_{i=0}^{M-1} y[aD+i]\cos(2\pi bi/M),$$

$$m_2[a,b] = -\sum_{i=0}^{M-1} y[aD+i]\sin(2\pi bi/M),$$

and where $I_0(z)$ is the zero order modified Bessel function of the first kind. Therefore, for each spectrogram element, the likelihood ratio is [90, 92, 93]

$$\begin{aligned} \lambda[a,b] =& \frac{p(S_x[a,b]|H_1)}{p(S_x[a,b]|H_0)} \\ =& \exp\left(-\frac{m_1^2[a,b] + m_2^2[a,b]}{M\sigma_n^2}\right) I_0\left(\frac{2\sqrt{(m_1^2[a,b] + m_2^2[a,b])S_x[a,b]}}{M\sigma_n^2}\right). \end{aligned} \tag{3.11}$$

Assuming that each element in the spectrogram matrix is statistically independent, the likelihood ratio is

$$\begin{aligned} \lambda =& \frac{p(\mathbf{S_x}|H_1)}{p(\mathbf{S_x}|H_0)} = \frac{p(S_x[0,0]|H_1)\cdots p(S_x[J,B]|H_1)}{p(S_x[0,0]|H_0)\cdots p(S_x[J,B]|H_1)} \\ =& \prod_{a=0}^{J}\prod_{b=0}^{B} \frac{p(S_x[a,b]|H_1)}{p(S_x[a,b]|H_0)} = \prod_{a=0}^{J}\prod_{b=0}^{B} \lambda[a,b]. \end{aligned} \tag{3.12}$$

2. Uncertain ocean

In reality, we may not know the environment and sound speed profile of the ocean exactly. If we know the prior distribution of the sound speed profile, we can apply the Bayes rule, and incorporate the prior information in the detector. Suppose the sound speed profile is uniformly distributed over P possible cases, the likelihood ratio, based on eq. (3.12), in this case is

$$\begin{aligned}\lambda_u =& \frac{1}{P}\sum_{k=1}^{P}\Big(\prod_{a=0}^{J}\prod_{b=0}^{B}\lambda_k[a,b]\Big)\\ =& \frac{1}{P}\sum_{k=1}^{P}\Bigg(\prod_{a=0}^{J}\prod_{b=0}^{B}\Big(\exp\Big(-\frac{m_{k1}^2[a,b]+m_{k2}^2[a,b]}{M\sigma_n^2}\Big)\cdots\\ &\times I_0\Big(\frac{2\sqrt{(m_{k1}^2[a,b]+m_{k2}^2[a,b])S_{kx}[a,b]}}{M\sigma_n^2}\Big)\Big)\Bigg),\end{aligned} \tag{3.13}$$

where λ_k is the likelihood ratio of the spectrogram of the kth possible propagated signal, $S_{kx}[a,b]$ is the spectrogram element of the kth possible propagated signal, and where $m_{k1}[a,b]$ and $m_{k2}[a,b]$ are the mean of the real and imaginary parts respectively of the kth possible propagated signal.

3. Mean ocean

When the sound speed profile is uncertain, and we only have the knowledge of the mean value of the possible sound speed profiles, which is mismatched with the true sound speed profile. Based on eq. (3.12), the likelihood ratio in this case becomes [97]

$$\lambda_m = \prod_{a=0}^{J}\prod_{b=0}^{B}\left(\exp\left(-\frac{\hat{m}_1^2[a,b]+\hat{m}_2^2[a,b]}{M\sigma_n^2}\right)I_0\left(\frac{2\sqrt{(\hat{m}_1^2[a,b]+\hat{m}_2^2[a,b])S_x[a,b]}}{M\sigma_n^2}\right)\right), \tag{3.14}$$

where $S_x[a,b]$ is the spectrogram element of the received signal, and $\hat{m}_1[a,b]$ and $\hat{m}_2[a,b]$ are the mean value of the real part and imaginary part respectively of the Short Time Fourier Transform of the signal in the mean ocean environment, and where the sound speed profile is assumed to take average values.

3.3.2 STFT Detector

The detection algorithms based on spectrogram data neglect phase information, and therefore their detection performances are not likely to be optimal. In this section we derive the probability distribution and likelihood ratio for detection for the STFT detector. The STFT detector preserves the phase information of the detected signal, and achieves optimal detection performance in the matched ocean case.

1. Matched ocean

The binary hypothesis test based on the STFT in this case is [103]

$$H_0: \quad X = X_n, \tag{3.15}$$

$$H_1: \quad X = X_y + X_n, \tag{3.16}$$

where X is the vectorized STFT matrix of the received data, X_n is the vectorized STFT matrix of pure noise, and X_y is the vectorized STFT matrix of the propagated signal. $X = [X[0,0], \cdots, X[J,B]]^T$. For each STFT element $X[a,b]$ in X, we have [32]

$$X[a,b] = \sum_{i=0}^{M-1} x[aD+i]w[i]\exp(-2\pi jbi/M), \tag{3.17}$$

where $a = 0, \cdots, J$ and $b = 0, \cdots, B$. Since $X[a,b]$ is a complex value, letting

$$U[a,b] = Re\{X[a,b]\}, \tag{3.18}$$

$$V[a,b] = Im\{X[a,b]\}, \tag{3.19}$$

we have $X[a,b] = U[a,b] + jV[a,b]$.

Concatenating the elements in the STFT matrix of the received data, and separating the real and imaginary part of each element, we can obtain

$$\mathbf{X} = (U[0,0], \cdots, U[J,B], V[0,0], \cdots, V[J,B])^T. \tag{3.20}$$

Because we use the Gaussian distribution as our noise model, $\mathbf{X}$ follows a multivariate normal distribution under both H_0 and H_1 hypotheses. It can be proved

that the signal covariance matrix $\mathbf{C_x}$ is the same under both H_0 and H_1 hypotheses. The expression for $\mathbf{C_x}$ is derived in Appendix A.

When $\mathbf{C_x}$ is singular, the probability density function for $\mathbf{X}$ does not exist. According to the derivation in Appendix A, mapping $\mathbf{X}$ into a subspace formed by Q_1, where Q_1 is the eigenvector of the non-zero singular values' component, we can obtain the probability density function for $Q_1^T X$ under H_0 and H_1 hypotheses

$$p(Q_1^T\mathbf{X}|H_0) = \frac{1}{(2\pi)^{(B+1)(J+1)}\det^{1/2}(\Lambda_1)}\exp\left(-\frac{1}{2}\mathbf{X^T}Q_1\Lambda_1^{-1}Q_1^T\mathbf{X}\right), \tag{3.21}$$

$$p(Q_1^T\mathbf{X}|H_1) = \frac{1}{(2\pi)^{(B+1)(J+1)}\det^{1/2}(\Lambda_1)}\exp\left(-\frac{1}{2}(\mathbf{X}-\mu)^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T(\mathbf{X}-\mu)\right). \tag{3.22}$$

where Λ_1 is the diagonal matrix of eigenvalues corresponding to Q_1, and μ is the expected value of $\mathbf{X}$.

The likelihood ratio based on the STFT is then

$$\begin{aligned}\lambda &= \frac{p(Q_1^T\mathbf{X}|H_1)}{p(Q_1^T\mathbf{X}|H_0)} \\ &= \exp\left(-\frac{1}{2}(\mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu - 2\mathbf{X^T}Q_1\Lambda_1^{-1}Q_1^T\mu)\right).\end{aligned} \tag{3.23}$$

We can calculate the analytic solution for the ROC plot based on eq. (3.23). According to the derivation in Appendix A, the square of separation parameter, or the detection index is

$$d^2 = \mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu. \tag{3.24}$$

2. Uncertain ocean

When the sound speed profile is uncertain, and is uniformly distributed over P possible cases. Applying the Bays rule, the likelihood ratio can be calculated as

$$\lambda_u = \frac{1}{P}\sum_{k=1}^{P}\exp\left(\mathbf{X^T}Q_1\Lambda_1^{-1}Q_1^T\mu_{\mathbf{k}} - \frac{1}{2}\mu_{\mathbf{k}}^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu_{\mathbf{k}}\right), \tag{3.25}$$

where $\mu_{\mathbf{k}}$ is the expected value of the kth possible propagated signal's STFT vector.

3. Mean ocean

When the sound speed profile is uncertain, and the prior information of the sound speed profile is unknown, but we know the mean sound speed profile, we can cross-correlate the possible propagated signals with the signal corresponding to the mean sound speed profile environment. The likelihood ratio can be computed as [97]

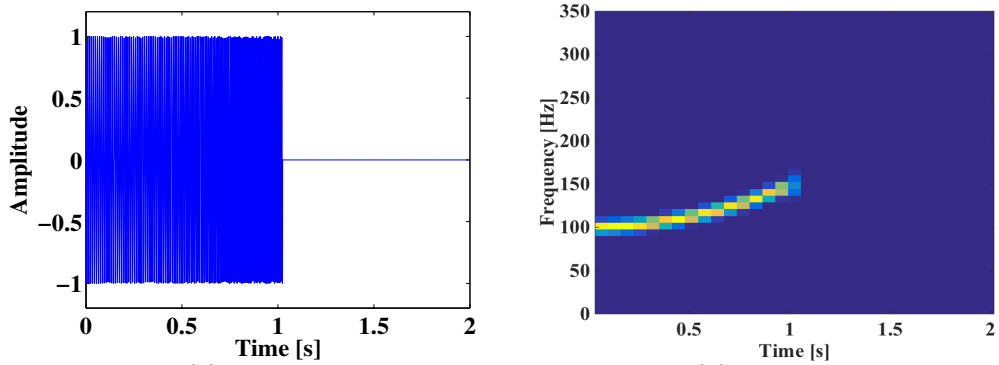
$$\lambda_m = \exp\left(-\frac{1}{2}\left(\mu_{\mathbf{m}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{m}} - 2\mathbf{X}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{m}}\right)\right), \tag{3.26}$$

where $\mu_{\mathbf{m}}$ is the expected value of the mean sound speed profile with respect to $\mathbf{X}$. We can mathematically prove that when the uncertainty of the sound speed profile is small, given that the mapping of sound speed profile to the propagated signal is linear, according to eq. (3.3), eq. (3.26) is actually the geometric mean of the likelihood ratio. A detailed proof can be found in Appendix A.

3.4 Results

Based on recordings of NARW from the years 2001 to 2003 in the Cape Cod Bay region, using bottom-mounted hydrophones with 2000Hz sampling rate, the polynomial coefficient set (f_0, f_1, f_2) was found most frequently to take the value of (100,0,48) [16]. Because the NARW upsweep call usually lasts for about 1 second, we let the duration of the signal be 1.024s. Substituting these values into eq. (3.1), and letting $A(t) = 1$, the source signal's waveform and spectrogram can be determined and are illustrated in Figure 3.5.

We placed the source signal 15 meters below the sea surface, and placed the receiver 35 meters below. The horizontal distance between the receiver and source signal is 1000 meters. Considering the uncertainty of the sound speed profile, we supposed that the sound speed profile has 50 possible cases. When the sound speed profile changes, the propagated signal will also change. Figure 6 illustrates an example that a source under two distinct sound speed profiles will produce two different propagated signals.



(a) Waveform (b) Spectrogram

FIGURE 3.5: Synthetic NARW source signal

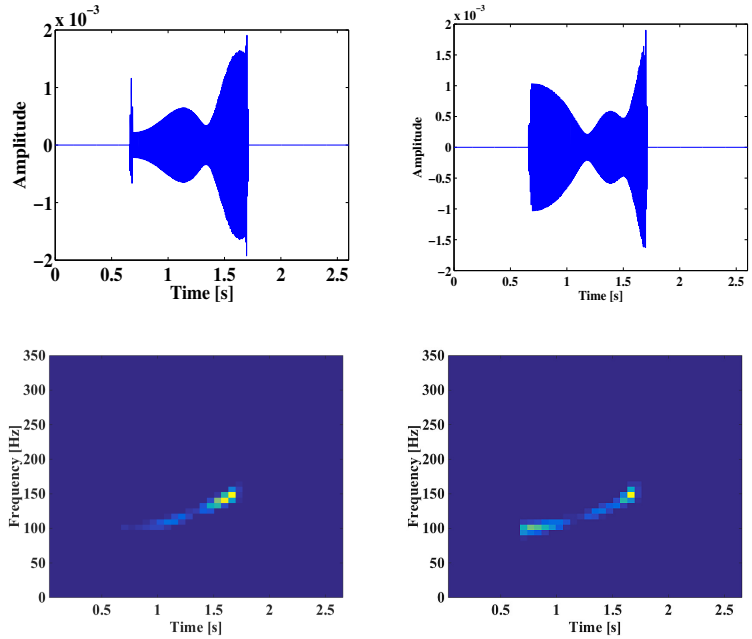


FIGURE 3.6: Example plots of time domain waveform and spectrogram of NARW propagated signal under two different sound speed profiles.

According to the spectrogram plots, we can see that there is a dispersion effect during signal propagation. Some frequency components in the source signal have been lost due to phase cancelation effects of multipath propagation environment. However, the overall shapes of 50 possible propagated signal spectrograms are similar to the source signal spectrogram; this is because the time shift will not change the

frequency content of the signal, due to the basic property of the Fourier Transform.

3.4.1 Detectors Tested

We compared the detection performance of the STFT detector, the spectrogram distribution detector, and the spectrogram correlation detector in the matched ocean, uncertain ocean and mean ocean cases. We benchmarked the results with the optimal detection performance given by the time domain matched filtering assuming that the source signal and propagation environment are known exactly.

The spectrogram correlation detector is a well known work in marine mammal acoustic detection field. Review of the spectrogram correlation detector [48] proposed by Mellinger and Clark is as follows. The spectrogram correlation constructs a kernel function for the vocalization signal, then cross correlates it with the target signal's spectrogram to calculate the recognition score, and makes detection decisions based on the recognition score. The kernel function for the signal is made up of several segments: one per FM section in the target vocalization type. The kernel value ke at a given time and frequency point (t, f) is specified by:

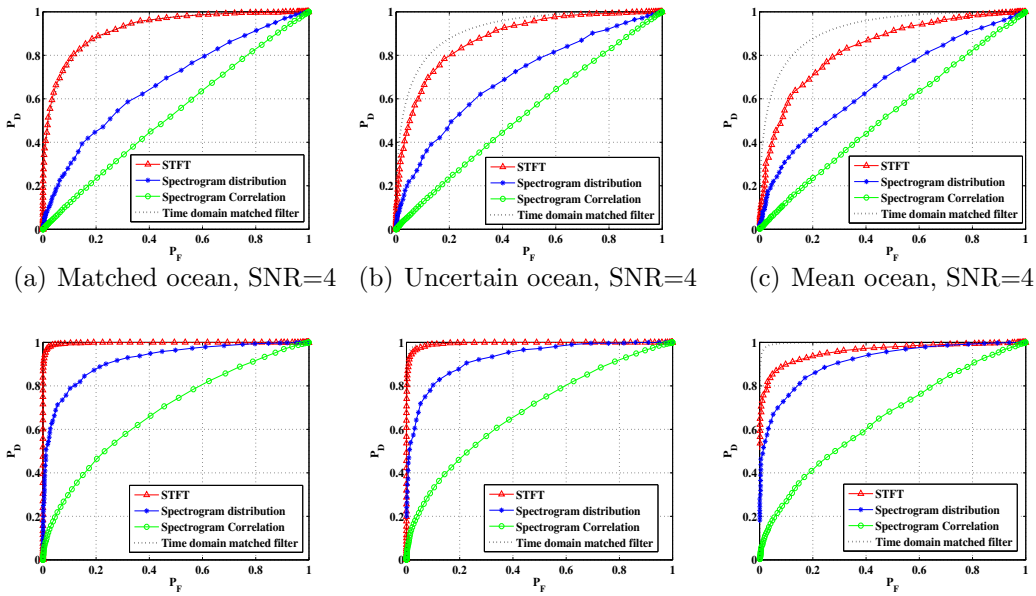
$$x = f - \left(f_0 + \frac{t}{d}(f_1 - f_0) \right)$$

$$ke(t, f) = \left(1 - \frac{x^2}{\sigma^2} \right) \exp\left(-\frac{x^2}{2\sigma^2} \right)$$

where x is the distance of the point (t, f) from the central axis of the segment at time t, f_0 is the start frequency of the segment, f_1 is the end frequency of the segment, d is the duration of the segment, and σ is the instantaneous bandwidth of the segment at time t. The recognition score is calculated by cross correlating the kernel $ke(t, f)$ with the spectrogram of the signal:

$$\alpha(t) = \sum_{t_0} \sum_{f} ke(t_0, f) S(t - t_0, f)$$

We used a rectangular window function with length 256, and overlap size 128 to generate the spectrogram of the signal. We applied these parameters to the spectrogram correlation detector and spectrogram distribution detector. We set a threshold



(a) Matched ocean, SNR=4 (b) Uncertain ocean, SNR=4 (c) Mean ocean, SNR=4

(d) Matched ocean, SNR=16 (e) Uncertain ocean, SNR=16 (f) Mean ocean, SNR=16

FIGURE 3.7: ROC plots for the detectors in various ocean environments and SNRs. (a), (b) and (c) show the detectors performances of the matched ocean, uncertain ocean and mean ocean environment case when SNR=4; (d), (e) and (f) show the matched ocean, uncertain ocean and mean ocean environment case when SNR=16.

for the recognition score and generated the Receiver Operating Characteristic (ROC) curve.

The comparison of detection performance of the STFT detector, the spectrogram distribution detector and the spectrogram correlation detector are shown in Figure 3.7. We used a Monte Carlo method to perform a numerical simulation. We analyzed the detection performance in the matched ocean, uncertain ocean and mean ocean cases.

3.4.2 Matched Ocean

We first examined the special case of detection in a matched ocean propagation environment. We can see that in this case the STFT detector achieves identical detection performance to the time domain matched filter case, which is the optimal detection performance. We used different window lengths and amounts of overlap, and the STFT detector performs identically in all situations. Figure 3.8 shows the

analytic ROC plots in different SNRs for the STFT detector, based on the separation index in eq. (3.24) and the analysis in Appendix A; along with the corresponding analytic time domain matched filtered ROC curves as a comparison. We can see the analytic derivation and the numerical simulation are a good match.

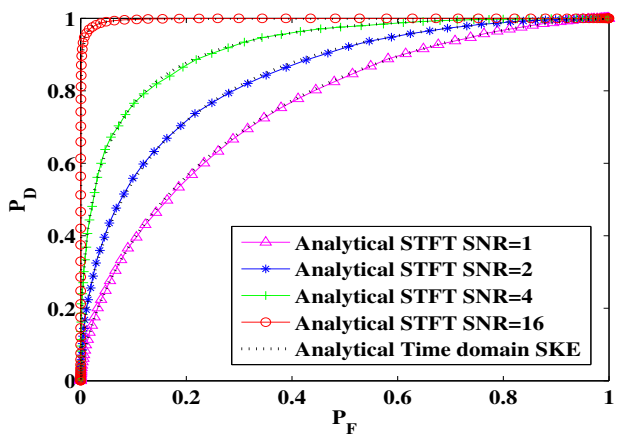


FIGURE 3.8: ROC plot based on analytical derivation when the environmental parameters are known exactly in different SNRs

Examining Figure 3.7, the STFT detector performs better than both the spectrogram distribution detector and the spectrogram correlation detector under different SNRs since it preserves the phase information. The spectrogram distribution detector performs better than the spectrogram correlation detector, and the spectrogram correlation detector does not perform well especially at low SNR. This is because the multipath effect will lead to dispersion of the signal, and some energy of the signal will be lost due to phase cancelation. When such a propagated signal is corrupted by noise, the performance of the spectrogram correlation detector suffers further, because it doesn't exploit the probability distribution of the spectrogram elements and noise.

3.4.3 Uncertain Ocean

For the uncertain ocean case, we assume there are 50 possible sets of sound speed profiles. The values of the sound speed profile are given in Table I. From the ROC

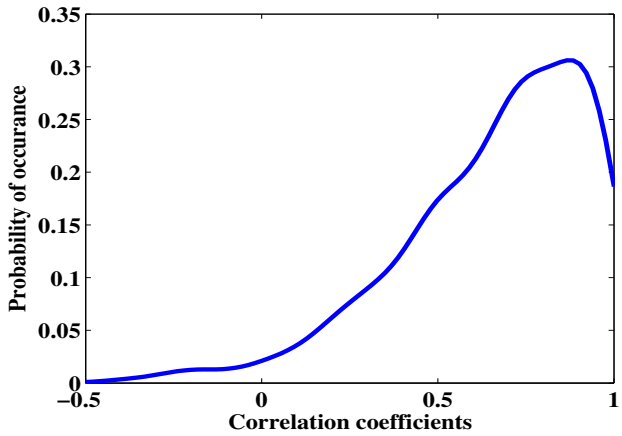


FIGURE 3.9: Kernel density function of propagated signals' correlation coefficients of different sound speed profiles

plots, we can see that the STFT detector performs close to the optimal case given by the time domain matched filter; and its detection performance is much better than the spectrogram distribution detector and spectrogram correlation detector. The kernel density function of correlation coefficients of the propagated signals of each possible sound speed profile is shown in Figure 3.9. We can see that most of the correlation coefficients are above 0.5; therefore, when we assume a uniform distribution over the possible sound speed profiles, the STFT detector performs close to the optimum.

3.4.4 Mean Ocean

For the mean ocean case, we can see that the overall detection performance deteriorates. Comparing the figures for the matched ocean case, the uncertain ocean case and the mean ocean case, we can see that the detection performance for the matched ocean case is the best, and the uncertain ocean is better than the mean ocean case. This is because the matched ocean has the most information while the mean ocean has the least. The performance of the spectrogram distribution detector does not change as much as the STFT detector, and there is little change in the performance of the spectrogram correlation detector for the three different

environment cases. This implies that, by neglecting phase information, the detector will be less sensitive to environmental uncertainty.

3.5 Conclusion

In this chapter, we have presented the STFT detector, which preserves the phase information of the signal. We have also determined the likelihood ratio of the STFT detector and the spectrogram distribution detector in the matched ocean, uncertain ocean and mean ocean cases. Experiments show that the STFT detector performs better than the spectrogram distribution detector and spectrogram correlation detector. The STFT detector is more sensitive to environmental changes, because it includes the phase information, while the spectrogram based detectors are less sensitive. By exploiting the probability distribution of noise and spectrogram element, the detector can be improved to be more robust in multipath propagation environments.

4

Classification of Whale Vocalizations using the Weyl Transform

In this chapter, we apply the Weyl transform to obtain the representation of whale vocalizations. In contrast to other popular representations, such as the MFCC and the chirplet transform coefficients, the Weyl transform coefficients capture the global information of signals. This is especially useful when the signal has low order polynomial phase. We can reconstruct the signal from the coefficients, and perform classification based on them. Experimental results show that classification using features extracted from the Weyl transform outperforms the MFCC and the chirplet transform coefficients on our collected whales data.

4.1 Introduction

There is a great deal of current research interest on better representing and classifying the vocalizations of marine mammals. However, the best feature extraction method for marine mammals classification is unknown. The variation of whale vocalizations and the uncertainty of the ocean environment can decrease the accuracy of classification and more work needs to be done in this area. A distinctive feature of many marine mammal calls is that they are frequency modulated. For this reason, it is natural to model such signals as polynomial phase signals [15, 16]. In this paper, our interest is in the task of classifying the chirp-like signals of marine mammals. It therefore becomes natural to ask that the features for classification of such signals should detect frequency modulation, also known as chirp rates.

One of the most popular features for classification of acoustic signals (including marine mammals) is the MFCC (Mel Frequency Cepstral Coefficients) [69, 104]. The MFCCs are short term spectral based features [68]. Despite being a powerful representation, the MFCC involves first order frequency information alone, and therefore gives no direct information about the chirp rates.

A recent attempt to capture chirp rate information more explicitly is using the discrete chirplet transform [86, 105], which was proposed for classification in [29, 30]. Chirplets are excellent for capturing localized chirp-like behaviour.

In this chapter, we propose a more global approach to obtain chirp rate information by using features based upon a second-order discrete time-frequency representation which we will refer to as the Weyl transform [106, 107, 108]. More technical details on the Weyl transform can be found in Section 4.4.

The features obtained from the Weyl transform is invariant to any shifts in both time and frequency. Furthermore, we show in Section 4.4 that, by pooling coefficients of the features obtained from the Weyl transform in an appropriate way, a feature

vector can be obtained which is essentially a chirp rate predictor. We will support our claims with numerical experiments in the context of the two-class NOAA test data set consisting of right whales and humpback whales. We propose two different sets of features which can be extracted from the Weyl transform, and compare them with MFCC and chirplets. We observe that both sets of features outperform the other two choices of features.

4.2 Background

Many different signal representation methods have been applied to whale signal representation, such as the chirplet transform [29, 30], the EMD transform [31], sparse coding [109], and MFCC [69]. Among them, the MFCC is one of the most popular. The chirplet transform is well known for its ability to detect a signal in a noisy environment [110]. In this section, we will briefly present these two methods and they will be the subject of our numerical experiments in Section 4.5.

4.2.1 MFCC

The MFCC is widely used in speech signal processing. The process of MFCC is to project and bin the short time Fourier transform of a signal according to a log-frequency (Mel) scale.[1] The short time Fourier transform of the signal $s(t)$ with length N is given by [111]:

$$X(k) = \sum_{t=0}^{N-1} w(t)s(t)\exp(-j2\pi kt/N), \quad k = 0, 1, \cdots, N-1$$

where $w(t)$ is the window function. We apply the Mel filter bank $H(k, m)$ to $X(k)$:

$$X'(m) = \ln\Big(\sum_{k=0}^{N-1} |X(k)|H(k, m)\Big), \quad m = 1, 2, \cdots, M$$

[1] Or sometimes a part-linear, part-logarithmic frequency scale.

where M is the number of filter banks and $M \ll N$. The Mel filter bank is a collection of triangular filters defined by the center frequency $f_c(m)$:

$$H(k,m) = \begin{cases} 0, \quad f(k) < f_c(m-1) \\ \frac{f(k)-f_c(m-1)}{f_c(m)-f_c(m-1)}, \quad f_c(m-1) \leqslant f(k) < f_c(m) \\ \frac{f(k)-f_c(m+1)}{f_c(m)-f_c(m+1)}, \quad f_c(m) \leqslant f(k) < f_c(m+1) \\ 0, \quad f(k) \geqslant f_c(m+1) \end{cases}$$

where $f(k) = kf_s/N$, f_s is the sampling frequency. The MFCCs are obtained by computing the DCT of $X'(m)$ using:

$$c(l) = \sum_{m=1}^{M} X'(m) \cos\Big(l\frac{\pi}{M}(m-\frac{1}{2})\Big), \quad l = 1, 2, \cdots, M \tag{4.1}$$

The above process is usually repeated over a sliding window, and the MFCC coefficients from each window are then concatenated.

4.2.2 Chirplet Transform

Given a signal $s(t)$, we can represent the signal as a weighted sum of chirplet functions [105, 112]:

$$s(t) = \sum_{i=1}^{M} A_i \exp(j\phi_i) k(n_i, t_i, \omega_i, c_i, d_i) \tag{4.2}$$

where $k(n_i, t_i, \omega_i, c_i, d_i)$ is the Gaussian chirplet function, and

$$k(n,t,\omega,c,d) = (\sqrt{2\pi}d)^{-\frac{1}{2}} \times \exp\Big\{-\Big(\frac{n-t}{2d}\Big)^2 + j\frac{c}{2}(n-t)^2 + j\omega(n-t)\Big\}. \tag{4.3}$$

The t, ω, c and d represents the location of time, frequency, chirp rate, duration of the Gaussian chirplet. We can represent and reconstruct the signal base on the chirplet coefficients.

4.3 Description of Signals

The marine mammal vocalizations can be represented by a family of polynomial-phase signals. The upsweep call is commonly found in right whale vocalizations,

which are typically in the 50-400 Hz frequency band and last for 1 second [16]. The humpback whale can also generate sounds like the right whale upsweep call. In this paper, we use right whale and humpback whale data, which were collected in the continental shelf off Cape Hatteras in North Carolina by NOAA and Duke Marine Lab, for experimental validations. The data was collected by using a linear array of marine autonomous recording units (MARUs) underwater, between December 2013 and February 2014. The MARUs are programmed to collect continuous acoustic recordings at a sample rate of 2 kHz. The data was collected from four different locations in Cape Hatteras. In this paper, we use the data file that was collected in the location which contains both right whales and humpback whales calls. The data is not publicly available at the moment.

The data file we use contains 24 vocalizations of right whales and 24 vocalizations of humpback whales. Example time-frequency representations using the right whale signals are shown in Figure 4.1, and the humpback whale signals in Figure 4.2.

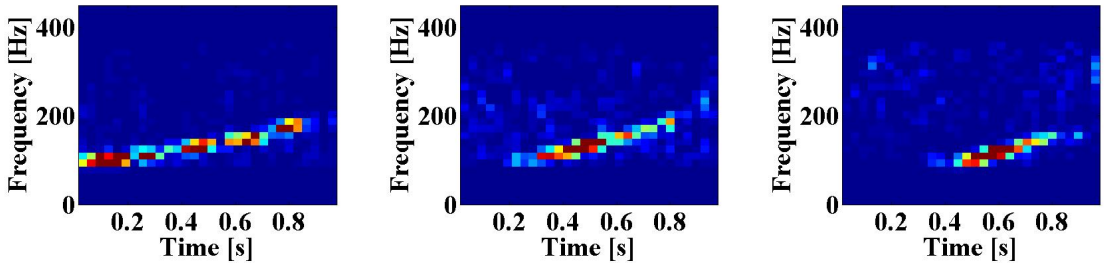


FIGURE 4.1: Examples of right whale signals

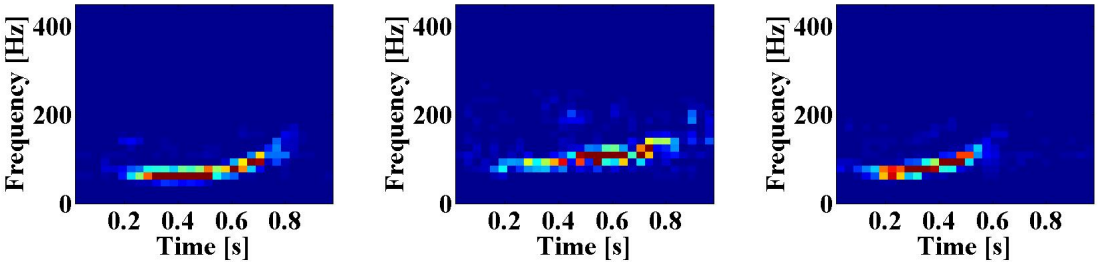


FIGURE 4.2: Examples of humpback whale signals

4.4 Using Weyl Transform to Chirp Signals

The Weyl transform in the Fourier domain is closely related to the Wigner Ville distribution [18, 113], the discrete polynomial phase transform [114, 115], and the ambiguity function, which is central in radar signal processing [116]. For a discretized signal $\mathbf{s}$ of length K, the Weyl transform coefficients has length K^2, and consists of the Fourier spectrum of diagonal bands of the covariance matrix $\mathbf{s}\mathbf{s}^{\mathbf{T}}$. It can be computed efficiently by means of K applications of the Fourier transform.

The representation of a signal obtain from the Weyl transform is as follows [107]: consider a real signal $s(l)$ over a time interval $[0, 1)$, discretized into K samples $s(t)$, where $t \in \mathbb{Z}_K = \{0, 1, \cdots, K-1\}$. Define the Weyl transform coefficients $\{\omega_{ab}\}$, where $a, b \in \mathbb{Z}_K$, as:

$$\omega_{ab} = \sum_{t=0}^{K-1} \exp\Big(-\frac{j2\pi bt}{K}\Big) s(t)s(t+a). \tag{4.4}$$

Letting $(Z_a)_t = s(t)s(t+a)$, Z_a is in fact a diagonal band of the correlation matrix $\mathbf{s}\mathbf{s}^T$, capturing periodicity. The Weyl transform coefficients consist of the Fourier transform of each correlation band $\omega_{ab} = \mathcal{F}\{Z_a\}_b$.

Now consider a linear chirp signal of the form:

$$s(l) = \cos(2\pi(ml + rl^2)), \quad m, r > 0$$

where m is the base frequency, and r is the chirp rate. Discretizing $s(l)$, we have:

$$s(t) = \cos\Big(2\pi\Big(\frac{mt}{K} + \frac{rt^2}{K^2}\Big)\Big). \tag{4.5}$$

We define two sets of features of using the Weyl transform for the signal.

Feature set 1: We use V_r as our feature set.

$$V_r = \sum_{\substack{(a,b):\\ 2ar/K=b, r\in\mathbb{Z}_n}} |\omega_{ab}|^2 \tag{4.6}$$

V_r is a chirp rate detector, because

$$\begin{aligned}\omega_{ab} =& \sum_{t=0}^{K-1} \exp\Big(-\frac{j2\pi bt}{K}\Big) \cos\Big(2\pi\Big(\frac{mt}{K} + \frac{rt^2}{K^2}\Big)\Big) \cos\Big(2\pi\Big(\frac{m(t+a)}{K} + \frac{r(t+a)^2}{K^2}\Big)\Big) \\ =& \frac{1}{2} \sum_{t=0}^{K-1} \exp\Big(-\frac{j2\pi bt}{K}\Big) \Bigg(\cos\Big(2\pi\Big(\frac{ma}{K} + \frac{(2at+a^2)r}{K^2}\Big)\Big) \\ &+ \cos\Big(2\pi\Big(\frac{m(a+2t)}{K} + \frac{r(2t^2+2at+a^2)}{K^2}\Big)\Big)\Bigg). \end{aligned} \tag{4.7}$$

The term $\cos\Big(2\pi\Big(\frac{m(a+2t)}{K} + \frac{r(2t^2+2at+a^2)}{K^2}\Big)\Big)$ is a chirp, and the sum of chirps is of lower order in V_r [106]. Therefore,

$$\begin{aligned}\omega_{ab} =& \frac{1}{4} \sum_{t=0}^{K-1} \Bigg(\exp\Big(j2\pi\Big(\frac{ma}{K} + \frac{ra^2}{K^2}\Big)\Big) \exp\Big(j2\pi\Big(\Big(\frac{2ra}{K} - b\Big)\frac{t}{K}\Big)\Big) \\ &+ \exp\Big(-j2\pi\Big(\frac{ma}{K} + \frac{ra^2}{K^2}\Big)\Big) \exp\Big(-j2\pi\Big(\Big(\frac{2ra}{K} + b\Big)\frac{t}{K}\Big)\Big) \\ &+ \text{lower order terms}\Bigg). \end{aligned} \tag{4.8}$$

We can see that ω_{ab} has two sharp peaks when $-\frac{2ra}{K} \approx b$ and $\frac{2ra}{K} \approx b$. Since the signals of interest in the current data set always have positive chirp rate, we discount the negative chirp rates, and the peak at $\frac{2ra}{K} \approx b$ indicates a chirp rate:

$$r \approx \frac{bK}{2a}. \tag{4.9}$$

We can use V_r as the feature vector, or we can instead use it to fit a quadratic polynomial to the frequency, the coefficients of which will be our second set of features.

Feature set 2: We use $(\hat{m}, \hat{r})$ as our feature set.

$$\hat{r} = \arg\max_{r \in \mathbb{Z}_n} V_r \tag{4.10}$$

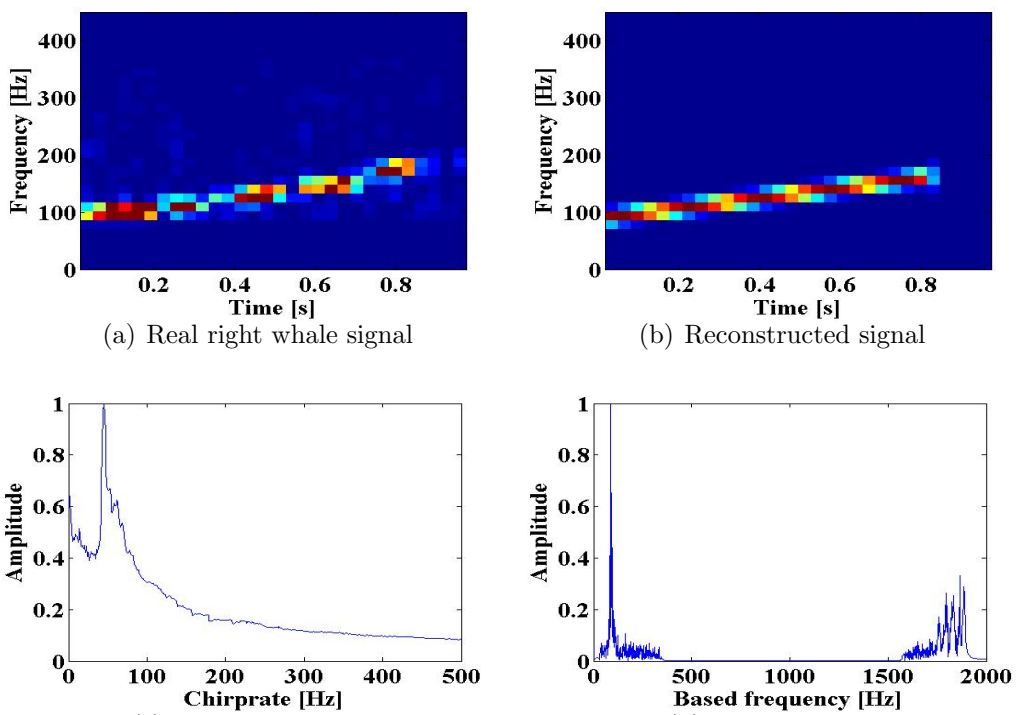
having estimated $\hat{r}$, we de-chirp:

$$\hat{s}(t) = s(t) \exp\Big(\frac{-j2\pi\hat{r}t^2}{K^2}\Big) \tag{4.11}$$

and take the Fourier transform of $\hat{s}(t)$, $\mathcal{F}(\hat{s})$, and record the location $\hat{m}$ of the largest entry as the estimate of m. We thus obtain the second feature set $(\hat{m}, \hat{r})$. The chirp is characterized by $(\hat{m}, \hat{r})$ as:

$$s(t) \approx \cos(2\pi(\hat{m}t + \hat{r}t^2)). \tag{4.12}$$

Note that this is an extremely compact feature set, where each signal has just two features. An example of signal estimation is illustrated in Figure 4.3. The right whale and humpback whale can generate upsweep calls, which can be expressed using the linear chirp model. The original right whale signal is shown in Figure 4.3(a), and Figure 4.3(c) is the plot of the features V_r and the location of the peak corresponding to the value of the estimated chirp rate $\hat{r}$. The plot of the Fourier transform of $\hat{s}(t)$ is shown in Figure 4.3(d), with the location of the peak corresponding to the estimated base frequency $\hat{m}$. Figure 4.3(b) is the estimated signal using $\hat{m}$ and $\hat{r}$.



(a) Real right whale signal (b) Reconstructed signal

(c) Chirp rate vector (d) Base frequency

FIGURE 4.3: Example of signal reconstruction

The technique to obtain feature set 2 is closely related to existing methods of

detecting polynomial phase signals, such as the discrete polynomial phase transform [114, 115] or the higher order ambiguity function [117], and also to the Weyl time-frequency strip filters described in [118].

4.5 Classification Results

We apply the Weyl transform, the chirplet transform and MFCC to obtain signal features, and apply the KNN classifier ($k = 3$) to classify the NOAA data. For the MFCCs, we form the spectrogram using the Hamming window of length 128, and the step size 64, then compute the coefficients by multiplying the filter bank function [119] with the spectrogram. We extracted 12 coefficients from each time frame, and concatenate the coefficients along the time axis. Suppose the length of the signal is 1 second in length, and the sampling frequency is 2000 Hz, then for each signal the length of MFCC features is 384. For the chirplet part, we use the Gaussian chirplet atom, and use 15 chirplet atoms to represent each signal. We use maximum likelihood estimate and the EM algorithm [67, 105] to estimate the chirplet coefficients and use the base frequency and chirp rate as features for each signal, giving a feature vector of length 30.

The ROC plot is shown in Fig 4.4. The AUCs (Area under the curves) under the ROC plots are given in Table 4.1.We use six fold cross-validation to generate the plots. We use 83% right whale and humpback whale data to do the training, and the remaining whale data to do the testing. We calculate the distance of each testing data points to the training data points, and make a decision for each testing data based on its three nearest neighbor points, and compare it with the ground truth, to obtain the value of true positive rate and false positive rate. We know that the probability of false alarm (P_F) and the probability of detection (P_D) are both from 0 to 1. In order to generate the ROC curve, we vary the value of P_F over the range $0 : 1/6 : 1$, and obtain the corresponding threshold for P_D based upon the obtained true positive and false positive rate. Since the number of vocalizations of right whales and humpback whales available for the classification results were small,

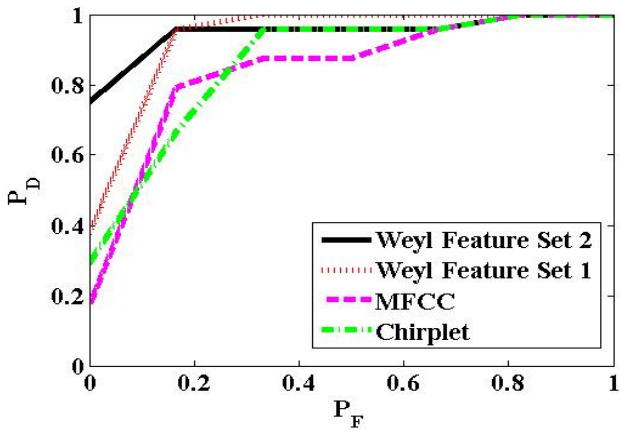


FIGURE 4.4: The ROC of classifying whales using signal representation methods

the classification results are promising but preliminary.

Table 4.1: Area Under the Curves (AUCs) of Duke Marine Data Classification

Weyl transform feature Set 2	Weyl transform feature Set 1	MFCC	Chirplet transform
0.9514	0.9410	0.8472	0.8646

With the KNN classifier, the classification accuracy of using Weyl transform feature set 1 and Weyl transform feature set 2 outperform the MFCC and chirplet coefficients. For this data set, the frequency ranges of right whale and humpback whale vocalization are both in the 50-250 Hz frequency band, and their base frequency is uniformly distributed in the range of 40 to 160 Hz, but the length of their vocalization and energy distribution are different, which means that the chirp rate information can better represent the whale calls. In addition, some of the humpback whale data have several harmonics, while all the right whale data just have one harmonic, the Weyl transform coefficients can distinguish one harmonic case and several harmonic case. The MFCC can represent the local frequency information of the signal over time, but neglect the higher order chirp rate information. Moreover, the MFCC applies the filter-bank function to the spectrogram, and it may not be able to distinguish the several harmonic and one harmonic case.

For the chirplet coefficients, the computational cost to obtain the coefficients is

high in our approach, so the limited numbers of chirplet atoms may not perfectly represent the whole signal, especially when the length of the signal is long. Like the MFCC, chirplet atoms can only represent local information of the signal, so in this data set, it does not perform as well as the features obtained from the Weyl transform.

4.6 Conclusion

In this chapter we have shown that the features obtained from the Weyl transform can well represent polynomial phase signals. We can obtain a chirp rate predictor by pooling Weyl transform coefficients appropriately. The chirp rate feature vectors and chirp coefficients have been shown to outperform the MFCC and chirplets for right whale and humpback whale data. Similar results are to be expected in classifying other marine mammal calls which have similar frequency range, but whose chirp rates can be distinguished.

5

Intrinsic Structure Study of Whale Vocalizations

By understanding the inherent low dimensional structure of whale vocalizations, we can effectively address high dimensional classification problems. A distinctive feature of many marine mammal calls lies in the fact that they are frequency modulated and can be modeled as polynomial phase signals. This implies that the intrinsic dimension of whale vocalizations can be described and estimated by the number of polynomial phase parameters. Traditional dimensional reduction methods, such as PCA and MDS, assume that the data lies on a low-dimensional plane. In this study we explore nonlinear manifold mapping methods, in particular Laplacian Eigenmap and ISOMAP, to examine the underlying manifold structure of the whale vocalization in the time-frequency plane. Experimental results show that nonlinear manifold methods outperform PCA and MDS in classification on the DCLDE whale vocalization data and Mobysound data, suggesting that the intrinsic structure of whale acoustic data is nonlinear rather than linear.

5.1 Introduction

Many geometrically motivated approaches for data analysis are based on the assumption that high dimensional natural data actually reside on a low dimensional manifold, so that much fewer degrees of freedom are required to understand such data. The manifold structure of speech sounds stresses the nonlinear relation between articulatory and acoustic space [79], and many algorithms exploiting this manifold structure have demonstrated success on speech signal processing problems [64, 79, 80, 81].

In this chapter we study the classification of whale vocalizations. Many marine mammal calls are frequency modulated, and can be modeled as polynomial phase signals [15, 16]:

$$s(t) = A(t)\cos(2\pi \sum_{m=0}^{N} a_m t^m). \tag{5.1}$$

where $A(t)$ is the amplitude, and $\{a_m\}_{m=0}^{N}$ are the polynomial coefficients. When the time-frequency contour is adequate to describe the signal, we can use the polynomial coefficients $\{a_m\}_{m=0}^{N}$ to represent the signal. Therefore, it is plausible that such signal lies on an m dimensional manifold parameterized by $\{a_m\}_{m=0}^{N}$. Moreover, since $s(t)$ is nonlinear in $\{t^m\}_{m=0}^{N}$, it suggests that the signal lies on a nonlinear manifold instead of an m dimensional plane.

We explore various dimension reduction methods on the DCLDE conference [120] and MobySound [34] whale vocalization data, and perform classification after the dimension reduction. Experimental results show that the nonlinear methods such as ISOMAP and Laplacian Eigenmap outperform linear methods such as PCA (Principal Component Analysis) and MDS (Multi-Dimensional Scaling). This is consistent with the signal model in eq. (5.1), which shows that $s(t)$ is nonlinear in $\{t^m\}_{m=0}^{N}$.

5.2 Dimension Reduction Methods

The process of linear and nonlinear models for dimension reduction is illustrated in Fig. 5.1.

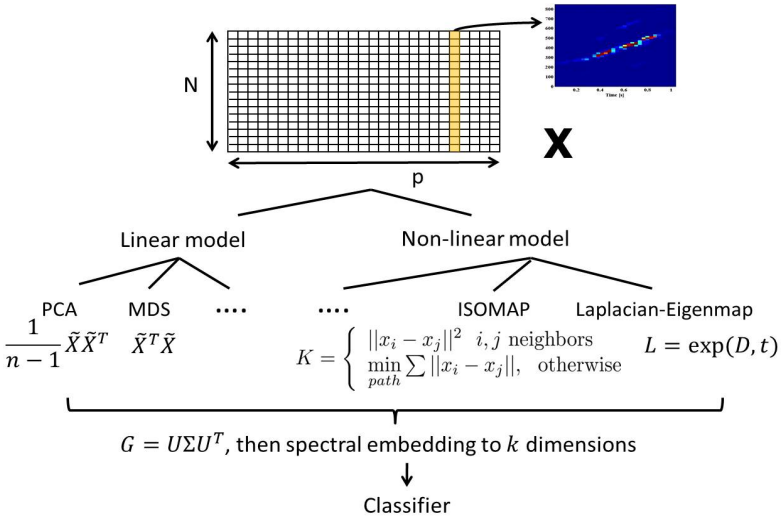


FIGURE 5.1: Methods of dimension reduction

In the figure, $X = [x_1, \cdots, x_n] \in \mathcal{R}^{p \times n}$ is the data matrix, where p is the dimension of the signal, and n is the number of signals.

PCA [60, 61] is a ubiquitous method for dimension reduction for high dimensional Euclidean data. Given the centralized data $\tilde{X} = X - \frac{1}{n}Xee^T$, we compute the sample covariance matrix $\Sigma_n = \frac{1}{n-1}\tilde{X}\tilde{X}^T$. The top k eigenvectors are called the principal components, and the data projected onto those principal components forms the feature set.

In MDS [62], given the pairwise Euclidean distance matrix D, we compute $B = -\frac{1}{2}HDH^T$, where H is the centering matrix. The data projected onto the top k eigenvectors of B forms the feature set.

MDS and PCA are related in that the top left singular vectors of $\tilde{X}$ are computed in MDS, while the top right singular vectors of $\tilde{X}$ are computed in PCA [121].

PCA and MDS rely on the assumption that the data points are near a low dimensional plane. However, when the data is sampled from a highly nonlinear curved surface, PCA and MDS will fail, while ISOMAP and Laplacian Eigenmap, the two classical spectral kernel embedding methods, succeed in capturing features of nonlinear structured data.

5.2.1 ISOMAP

ISOMAP [65] is an extension of MDS. Since the Euclidean distance in high-dimensional space usually fails to capture the intrinsic similarity between data points. ISOMAP exploits the geodesic distance along the low dimensional manifold.

ISOMAP constructs a neighborhood graph $G = (X, E, D)$ based on k nearest neighbors, or a ε-neighborhood, where X is the data, E is the edge and D is the distance matrix. When i and j are neighbors, the geodesic distance can be approximated by the Euclidian distance such that

$$D_{ij} = ||x_i - x_j||^2, \quad \text{if } i \text{ and } j \text{ are neighbors} \tag{5.2}$$

and when i and j are not neighbors, ISOMAP uses the shortest path distance such that

$$D_{ij} = \min_{\substack{\{t_1,\ldots,t_k\} \\ \text{is a path between } i,\, j}} (||x_{t_1} - x_i|| + \cdots + ||x_{t_k} - x_j||). \tag{5.3}$$

Let $H = I - ee^T/n$ be the centering matrix, and $K = -\frac{1}{2}HDH^T$, the top k eigenvectors of K give us the feature sets.

The drawback of ISOMAP is that the geodesic distance can only be accurately approximated when the data is densely sampled with high SNR. Therefore, obtaining the geodesic distance is computationally expensive. Meanwhile, when the data is noisy, according to eq. (5.3), the errors can accumulate along the path and affect data representation. The noise sensitivity of ISOMAP is illustrated in the experimental results in Section 5.3.2.

5.2.2 Laplacian Eigenmap

Laplacian Eigenmap [64] is a main tool used in spectral clustering. It constructs an undirected, weighted graph represented as a weighted matrix based on the nearest neighbors. The eigenvalues and eigenvectors of the weighted matrix are calculated to form the feature set. The goal of this method is to find a partition of the graph such that the edges between different groups have low weight, and the edges within a group have high weight. This implies that points in different clusters are dissimilar to each other, whereas points within the same cluster are similar to each other [122].

In Laplacian Eigenmap, the weighted matrix $W = (w_{ij}) \in \mathcal{R}^{n \times n}$ is defined using a heat kernel such that

$$w_{ij} = \begin{cases} \exp(-\frac{||x_i - x_j||^2}{t}), & \text{if } i \text{ and } j \text{ are neighbors} \\ 0, & \text{otherwise} \end{cases} \tag{5.4}$$

for a fixed t. Letting D be the diagonal matrix such that $D_{ii} = (\sum_j w_{ij})$, and defining the unnormalized graph Laplacian as

$$L = D - W. \tag{5.5}$$

The symmetric normalized graph Laplacian is given by:

$$\mathcal{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}. \tag{5.6}$$

We compute the eigenvalues and eigenvectors of L or $\mathcal{L}$, and drop the zero eigenvalues and the corresponding eigenvectors. The bottom k eigenvectors, which correspond to the smallest k eigenvalues, are the feature set.

The first non-zero eigenvalue is called the Fielder value, and the corresponding eigenvector is called the Fielder vector. The Fielder value is the algebraic connectivity of a graph; the further from 0, the more connected the graph. We can use the Fielder value to do spectral bi-partitioning, as illustrated in Section 5.3.2.

5.3 Experimental Results

5.3.1 Dataset

We used two sets of data. One is the DCLDE 2015 conference blue whale D call data and fin whale 40Hz call data [120], the other is the bowhead whale and humpback whale vocalizations from the Mobysound mysticetes database [34]. For the DCLDE data, there are 851 blue whale calls and 244 fin whale calls at the same sampling frequency 2000Hz. Spectrograms of the data are shown in Fig. 5.2 and Fig. 5.3.

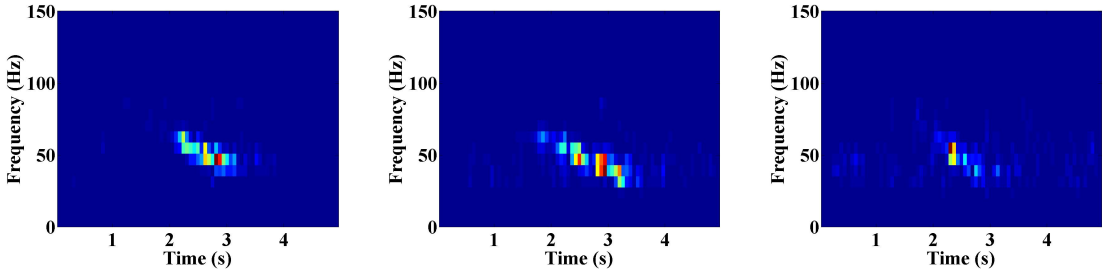


FIGURE 5.2: Blue whale signals' spectrogram

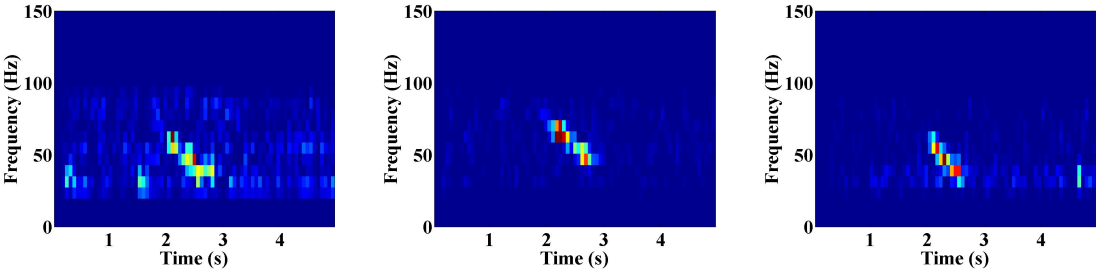


FIGURE 5.3: Fin whale signals' spectrogram

In the Mobysound data, the bowhead whale and humpback whale have the same sampling frequency 4000Hz. We used 446 bowhead whale calls and 2310 humpback whale calls. The humpback whale calls have much more variety than the bowhead whale calls. Example plots are shown in Fig. 5.4 and Fig. 5.5.

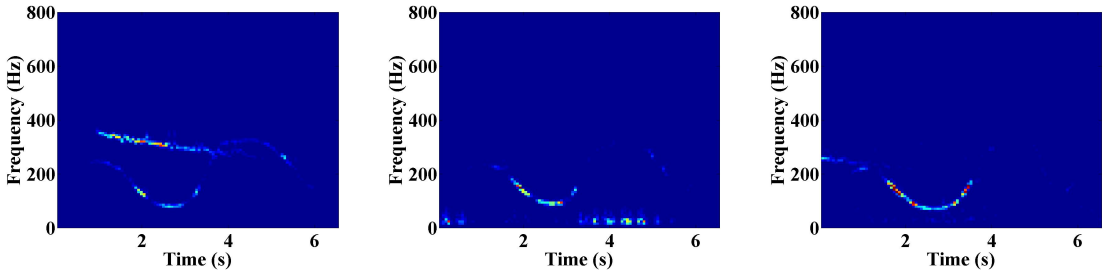


FIGURE 5.4: Example of bowhead whale signals

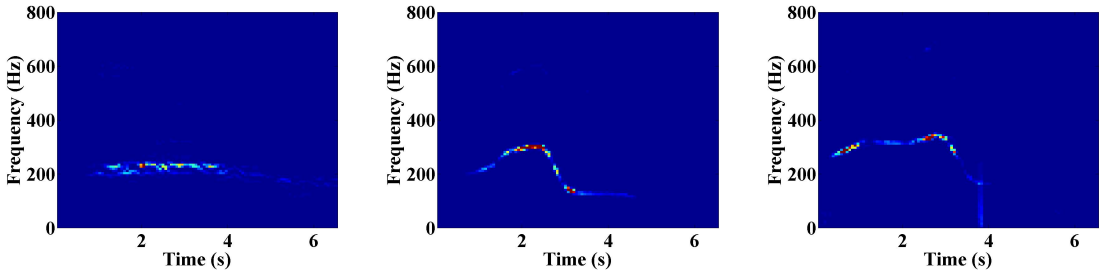
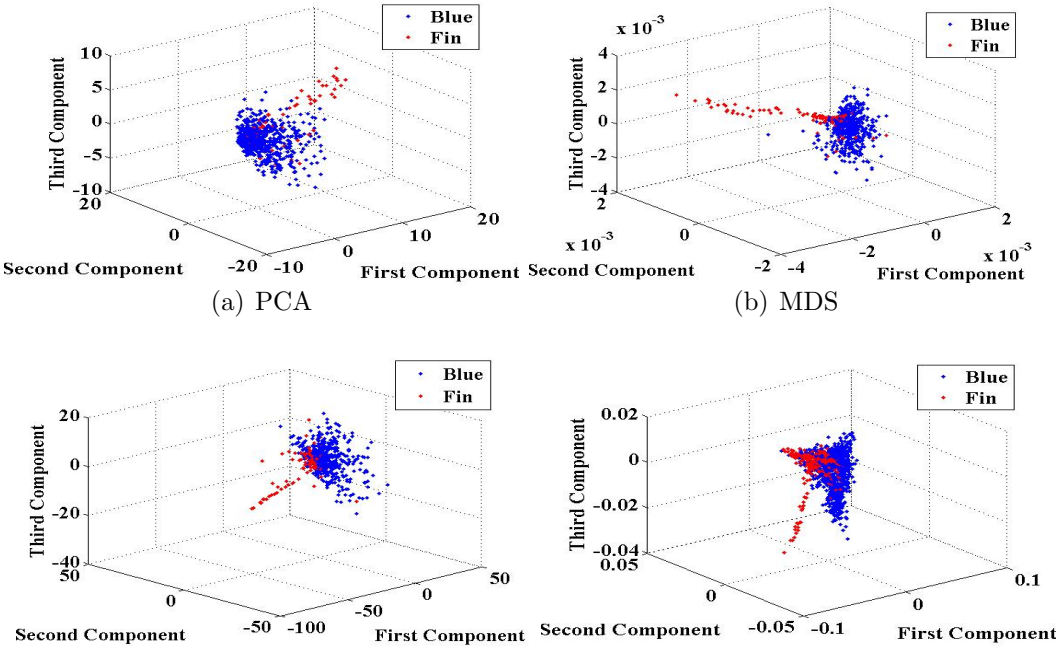


FIGURE 5.5: Example of humpback whale signals

5.3.2 Evaluation

We applied PCA, MDS, ISOMAP and Normalized Laplacian Eigenmap to the datasets. The projections of the data in the top three principle directions are shown in Fig. 5.6 and Fig. 5.7. Seven nearest neighbors are used to implement the ISOMAP and Normalized Laplacian Eigenmap. We used the spectrogram features as inputs for the DCLDE conference data and the MFCC features as inputs for the Mobysound data. Looking at the plots, we can visually separate whales from different classes by using Laplacian Eigenmap after the data projections.

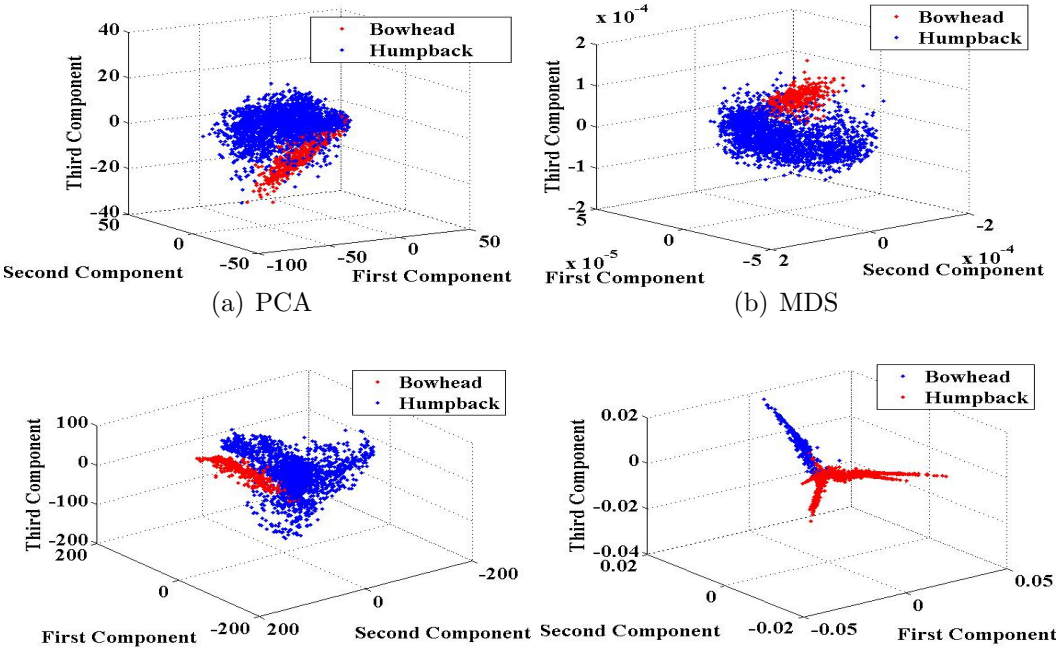
Since the whale vocalizations studied here are polynomial phase signals, the intrinsic dimension of the data can be described by the number of polynomial phase parameters. We can observe that the blue whale and fin whale vocalizations in the DCLDE conference data are both downsweep chirps. Thus we can use three parameters: phase, start frequency and slope to describe such signals. The eigenvalue distributions of the DCLDE conference data using PCA, MDS, ISOMAP and Normalized Laplacian Eigenmap are shown in Fig. 5.8.



(a) PCA (b) MDS

(c) ISOMAP (d) Normalized Laplacian Eigenmap

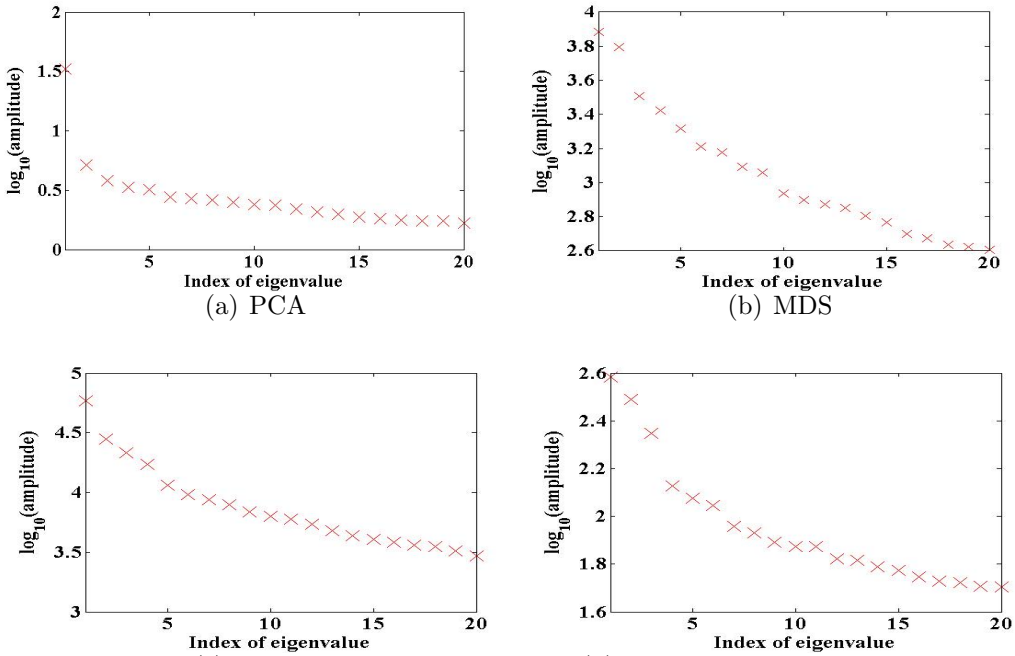
FIGURE 5.6: DCLDE data mapping



(a) PCA (b) MDS

(c) ISOMAP (d) Normalized Laplacian Eigenmap

FIGURE 5.7: Mobysound data mapping



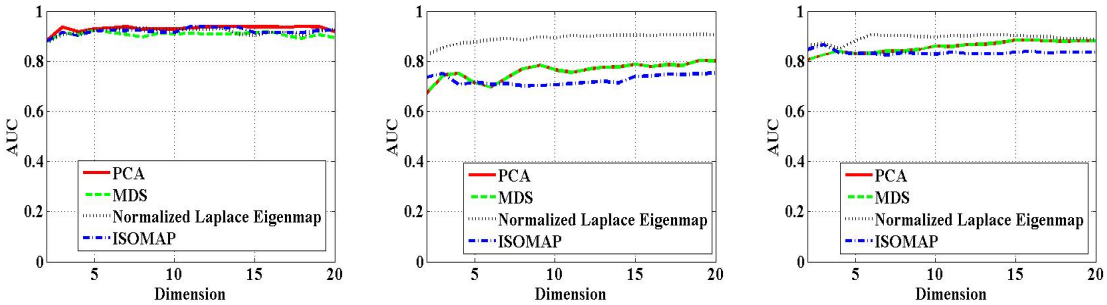
(a) PCA (b) MDS

(c) ISOMAP (d) Normalized Laplacian Eigenmap

FIGURE 5.8: Eigenvalues distribution

In Fig. 5.8, there is a gap between the third and the fourth eigenvalue in the Laplacian Eigenmap. This means that the Laplacian Eigenmap captures the intrinsic dimension of the signal information well. This is also illustrated from the plots of AUC (Area Under the Curve) versus dimensions in Fig. 5.9. We used PCA, MDS, ISOMAP, and Laplacian Eigenmap to compress the data to 2 to 20 dimensions. We used five-fold cross validation to generate the plots. In other word we used 681 blue whale calls and 196 fin whale calls for training, and 170 blue whale and 48 fin whale calls for testing. According to Fig. 5.9, the Laplacian Eigenmap outperforms all other methods. ISOMAP does not work as well as the Laplacian Eigenmap, because as stated in Section 5.2.1, the DCLDE datasets are noisy. Thus, the algorithm is affected by noise when the shortest path distance is computed, as shown in eq. (5.3).

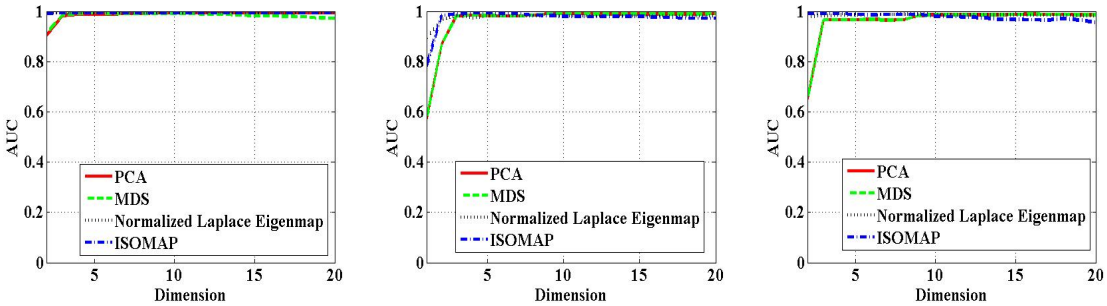
For the Mobysound dataset, the AUC plot is shown in Fig. 5.10. In this case, because the signal to noise ratio is high, both ISOMAP and Normalized Laplacian Eigenmap outperform PCA and MDS. The ISOMAP and Normalized Laplacian Eigenmap can represent the data efficiently with much lower dimensions than PCA



(a) KNN (b) Naive Bayes (c) Logistic regression

FIGURE 5.9: AUC Comparisons of DCLDE data

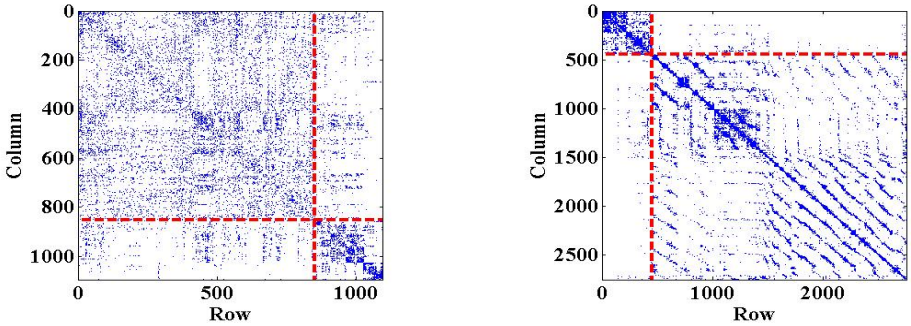
and MDS. This means that the nonlinear mapping can capture the intrinsic structure of the signal, and that the whale vocalization reside on a low dimensional nonlinear manifold.



(a) KNN (b) Naive Bayes (c) Logistic regression

FIGURE 5.10: AUC Comparisons of Mobysound data

Reordering the signal based on the sign of the Fielder vector, the corresponding adjacency matrix created by the Laplacian Eigenmap is shown in Fig. 5.11.



(a) DCLDE data (b) Mobysound data

FIGURE 5.11: Adjacency matrix of Laplacian Eigenmap

We can see that if two clusters are determined through the sign of the Fielder vector, the overall weight between data within the same cluster is larger than the weight between data from different clusters.

5.4 Conclusion

In this chapter we apply PCA, MDS, ISOMAP and Laplacian Eigenmap to perform dimension reduction on whale data. The nonlinear methods can efficiently represent the data with lower dimensions and capture the intrinsic physical information of the whale data. The fact that the nonlinear methods outperform the linear methods is consistent with the low dimensional nonlinear structure of whale vocalizations' signal model.

6

PCANet and DCTNet for Acoustic Signals Feature Extraction

We introduce the use of PCANet for acoustic signal classification, and propose DCTNet as an efficient alternative. The eigenfunctions of the local sample covariance matrix are used as filterbanks for convolution and feature extraction. For acoustic signals, these functions are well approximated by the Discrete Cosine Transform (DCT), and the convolutions become a short time DCT. This connection means that the output of the first layer of PCANet and DCTNet are essentially time-frequency representations. The second layer of DCTNet will give us features similar to linear frequency spectral coefficients (LFSC). Experimental results show that the whale vocalizations are well separated by using the features produced by both PCANet and DCTNet, and the word error rate of DCTNet in speech recognition tasks is similar to Mel-frequency spectral coefficients (MFSC), suggesting that PCANet and DCTNet can well represent the acoustic signal information.

6.1 Introduction

The power of multi-layer convolutional networks for learning features has been established in audio signal processing, such as speech recognition and music classification [74, 75, 76, 77]. The idea of a convolutional network is to convolve signals with filters, and use the obtained features for classification. The scattering transform, proposed by Andén and Mallat [78] showed that the Mel-Frequency Spectral Coefficients (MFSC) can be viewed as time-averaged wavelet convolutions, and the use of a two layer scattering transform led to improvements in classification accuracy over a single layer.

PCANet [123] was recently proposed for image classification. In this framework, filters are learned from the data as principal components at the local "image patch" level. Despite its simplicity, PCANet was shown to match and in some cases improve upon state of the art performance in a variety of image classification benchmarks.

In this chapter, we translate the PCANet framework into the world of acoustic signal processing. We present experimental evidence that the obtained PCANet filters are well approximated by the Discrete Cosine Transform (DCT) basis functions [124, 125]. We thus propose DCTNet as an efficient extension, in which PCA filters are simply replaced with fixed DCT filters. The process of our PCANet and DCTNet is shown in Fig. 6.1.

We relate PCANet and DCTNet to spectral feature representation methods for acoustic signals, such as the short time Fourier transform (STFT), spectrogram and MFSCs. In particular, each DCTNet layer is essentially a short time DCT. Just as the scattering transform cascades log-scale (Mel-scale) spectral coefficients, DCTNet cascades linear scale spectral coefficients. More technical details can be found in Section 6.2 to Section 6.4.

We use the DCLDE whale vocalization data [120] and Aurora 4 speech cor-

pus [126] for experiments. Results show that using the two-layer DCTNet and PCANet features can well separate the whale vocalizations, and the DCTNet achieves state-of-the-art performance in the whale vocalization classification task. The word error rate of using the two-layer DCTNet is similar to the MFSC in speech recognition tasks, suggesting that the DCTNet can represent the acoustic feature well.

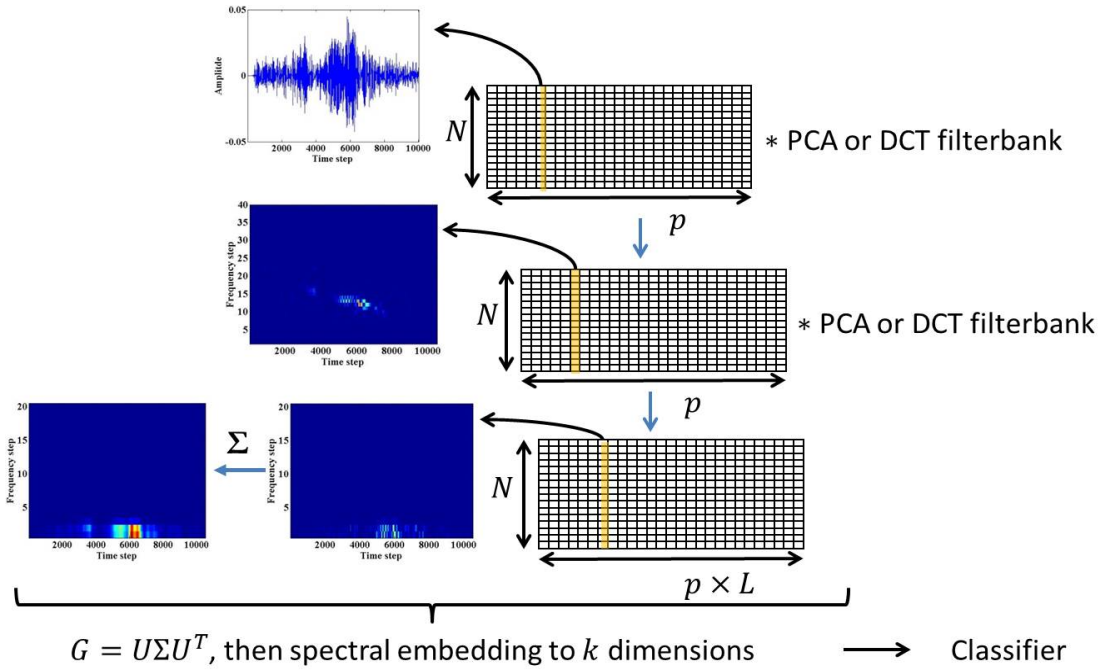


FIGURE 6.1: PCANet and DCTNet Process. The input is the time series of an acoustic signal. After convolving with DCT and PCA filterbanks, we have the short time PCA and short time DCT of the signal. After the second convolution, we have linear scale spectral coefficients. We then use them for spectral clustering and classification.

6.2 Eigenfunctions of Toeplitz Matrix

Suppose a signal $\mathbf{x} = (x(1), x(2), \cdots, x(N))$, we construct a Hankel matrix:

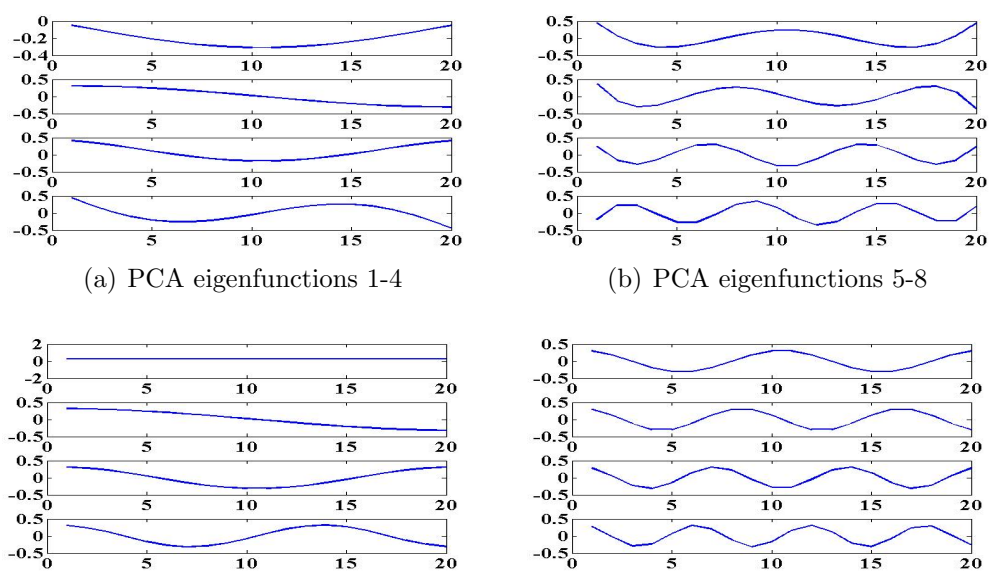
$$\mathbf{X} = \begin{bmatrix} x(1) & x(2) & \cdots & x(N-M+1) \\ x(2) & x(3) & \cdots & x(N-M+2) \\ \vdots & \vdots & \ddots & \vdots \\ x(M) & x(M+1) & \cdots & x(N) \end{bmatrix},$$

where $M < N$. When the first column $x(1), \cdots, x(M)$ and the last column $x(N - M + 1), \cdots, x(N)$ are zeros, let $\rho_j = \sum_i x(i)x(i+j)$, the sample covariance matrix:

$$\mathbf{XX^T} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{M-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{M-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{M-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{M-1} & \rho_{M-2} & \rho_{M-3} & \cdots & 1 \end{bmatrix}$$

is a Toeplitz matrix.

When the autocorrelation of the signal decays fast, the discrete cosine transform (DCT) basis function can well approximate the eigen-functions of the Toeplitz matrix [124, 125]. The comparisons of top eight eigenfunctions of DCT and PCA of whale vocalization data is shown in Fig. 6.2.



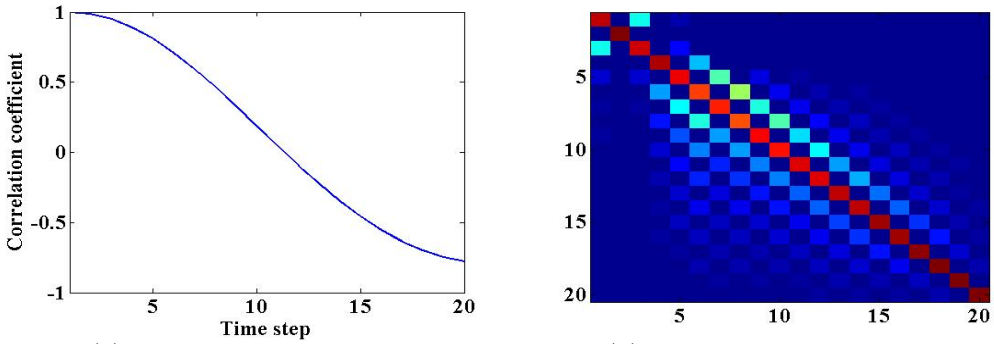
(a) PCA eigenfunctions 1-4 (b) PCA eigenfunctions 5-8

(c) DCT eigenfunctions 1-4 (d) DCT eigenfunctions 5-8

FIGURE 6.2: Comparisons of top eight DCT eigenfunctions and PCA eigenfunctions of whale vocalization data

The autocorrelation of signal and the correlation of DCT eigenfunctions and PCA eigenfunctions are shown in Fig. 6.3. We can see that the approximation of DCT to PCA eigenfunctions in terms of autocorrelation depends on the structure of data.

The error bound of using DCT eigenfunctions to diagonalize the Toeplitz matrix is determined by the autocorrelation coefficients of the signal [125].



(a) Autocorrelation of signal (b) Eigenfunctions correlation

FIGURE 6.3: Signal autocorrelation and eigenfunctions correlation

We can also interpret the DCT eigenfunctions in the Toeplitz-Fourier framework. Since the Toeplitz matrix T can be represented as the sum of circulant matrix C and skew circulant matrix S, that is $T = C + S$. The eigenfunctions of circulant matrix is Fourier eigenfunctions. We can view the skew circulant matrix as an error, and optimized energy of the circulant matrix within the given Toeplitz matrix. Applying the Chan's optimal circulant preconditioner [127] for the Toeplitz matrix, that is $C = \arg\min_{C'} ||C' - T||_F^2$. By doing this, we can optimize the energy of circulant matrix within the the Toeplitz matrix, and use the DCT eigenfunctions as a approximation of the Toeplitz matrix.

6.3 Short Time PCA and Short Time DCT

The discrete time STFT can be written as [83]:

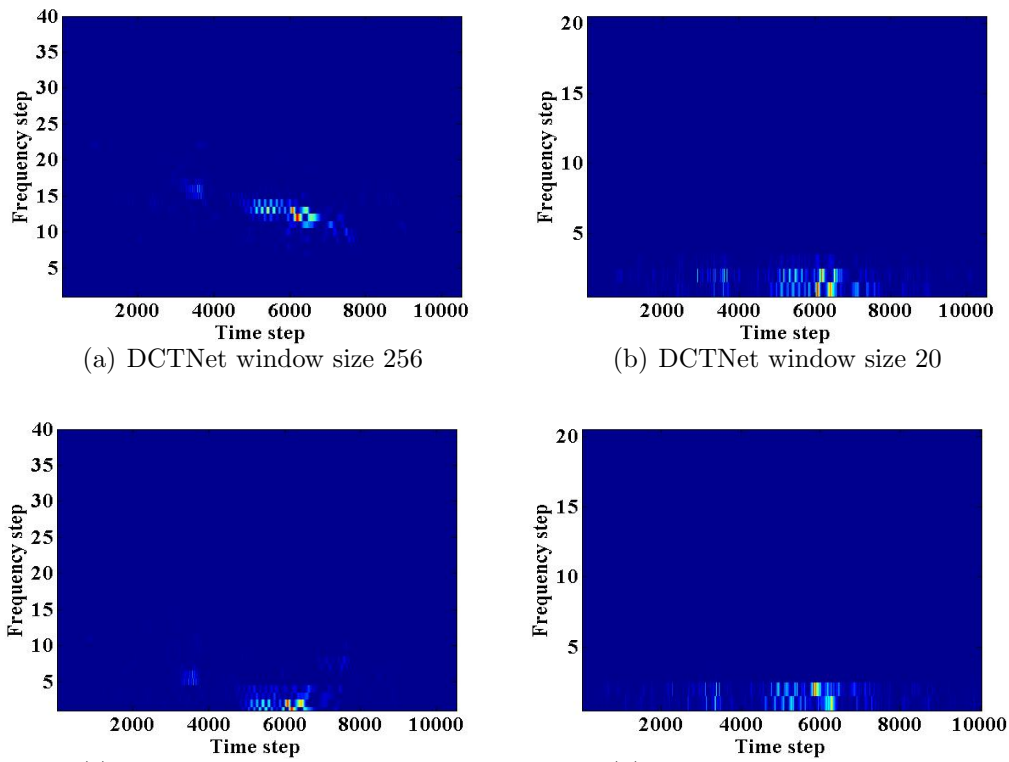
$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x(m-n)w(n)\exp(-j\omega(m-n))$$

$$= \exp(-j\omega m) \sum_{n=-\infty}^{\infty} (w(n)\exp(j\omega n))x(m-n)$$

$$= \exp(-j\omega m)[(w(m)\exp(j\omega m)) * x(m)]$$

where w is the window function, ω is the angular frequency. It can be viewed as the modulation of band-pass filter. For the PCANet and DCTNet, we replace the fourier basis function $\{\exp(j\omega m)\}$ with PCA eigenfunctions and DCT eigenfunctions, obtaining a short time PCA and short time DCT of the signal respectively.

Plots of short time PCA and DCT (output of the first layer) are shown in Fig. 6.4 for different window lengths. We use the DCLDE blue whale vocalization data as an example. DCT filterbanks are a natural choice since they are time-frequency representations, whereas time-frequency representation and resolution may be lost when using PCA filterbans, especially when the PCA eigenfunctions cannot be approximated by the DCT. We can see the comparison of Fig. 6.4(a) and Fig. 6.4(c), PCA fail to represent the time-frequency content of the signal.



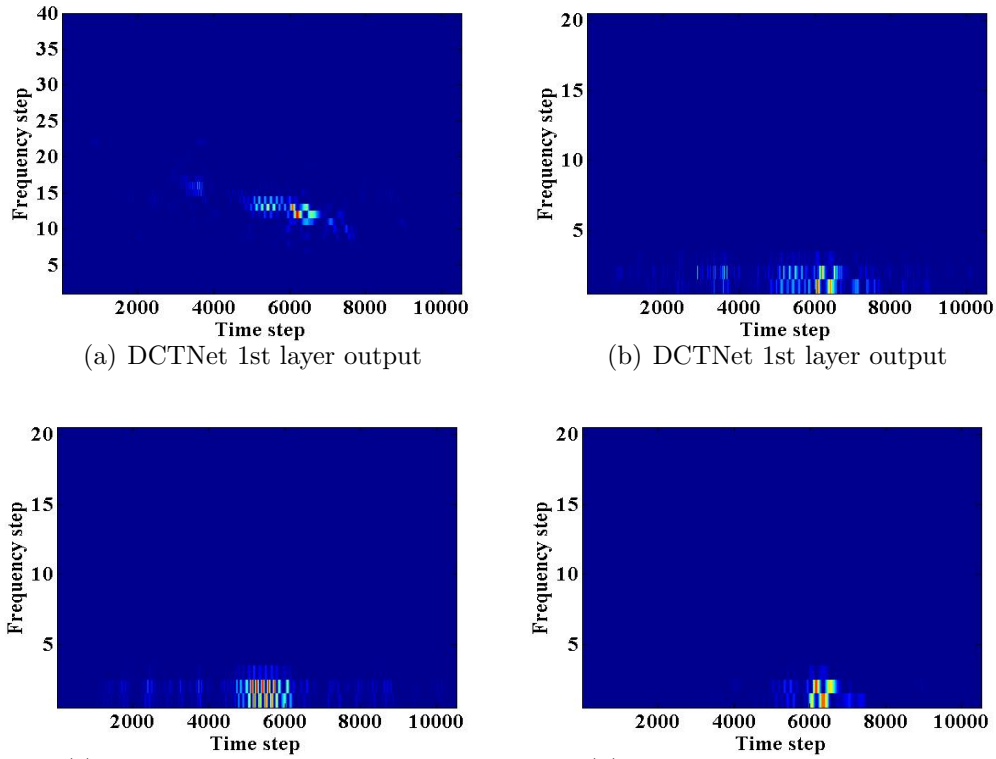
(a) DCTNet window size 256 (b) DCTNet window size 20

(c) PCANet window size 256 (d) PCANet window size 20

FIGURE 6.4: Plots of the first layer output

6.4 Linear Frequency Spectrogram

After obtaining the first layer, we view each row of output as a separate signal, and convolve it with a new PCA filterbank, or DCT filterbank. We thus end up with multiple new short time PCA and short time DCT, which can capture the dynamic structure of the signal. We choose a smaller window compared with the first layer window size for the filterbanks, so we have a finer scale representation (second layer) inside coarser scale representation (first layer). Plots of second layer short time DCT are shown in Fig. 6.5.



(a) DCTNet 1st layer output (b) DCTNet 1st layer output

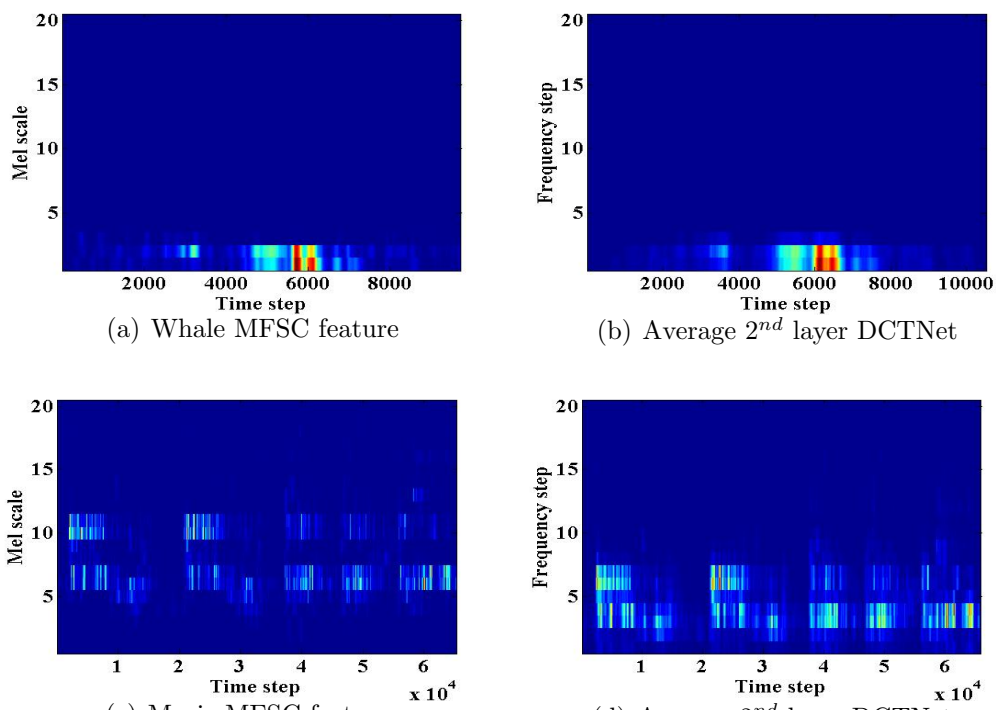
(c) DCTNet 2nd layer output 1 (d) DCTNet 2nd layer output 2

FIGURE 6.5: Comparisons of the first layer and the second layer outputs. (a) shows DCTNet first layer output with window size 256; (b) shows DCTNet first layer output with window size 20; (c) and (d) show DCTNet second layer output with window size 256 at the first layer, and window size 20 at the second layer. (c) shows the signal component at frequency step 14 of (a), while (d) shows the signal component at frequency step 12 of (a).

Comparing Fig. 6.5(b), and Fig. 6.5(c) and (d), we can see that the second layer

DCTNet outputs can reveal more detailed and dynamic signal components then just using first layer output as Fig. 6.5(b). We choose first layer window size be 256, in order to have a good time-frequency resolution to visualize the whale sounds, while the second layer window size be 20, in order to have a finer time resolution of the first layer output.

Taking an average to the second layer outputs, we arrive at linear frequency spectrogram like features, which have the stability to time-warping deformation. The features are similar to MFSC, which extract 12 to 20 coefficients after binning the spectrogram acoustic feature.



(a) Whale MFSC feature (b) Average 2nd layer DCTNet

(c) Music MFSC feature (d) Average 2nd layer DCTNet

FIGURE 6.6: MFSC and the second layer of DCTNet

As Fig 6.6 shows, we use whale vocalization data and Handel's "Messiah" music data to illustrate the similarity of the obtained feature to MFSC coefficients. The frequency range of this whale data is below 200Hz, and the Mel-scale in low frequency can be viewed as a linear scale, so we can see the similarity with the MFSC plot and the average of the DCTNet second layer coefficients. For the music data, its frequency

range is from 0 Hz to 1500 Hz, so there is a difference in Mel-scale and linear scale, leading to a shift in scale in the MFSC features compared to the average of the DCTNet second layer coefficients.

6.5 Experimental Results

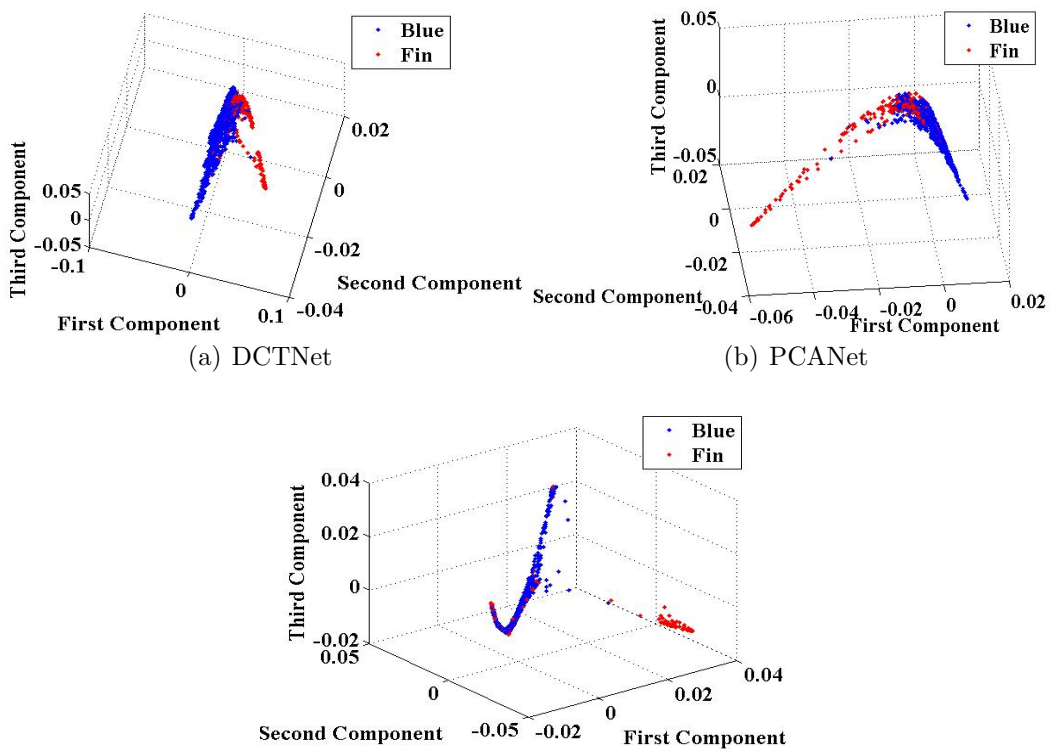
6.5.1 Dataset

We use whale vocalization data and speech data to examine the performance of PCANet and DCTNet. For whale vocalization test, we use the 2015 DCLDE blue whale D call data and fin whale 40Hz call data [120] for experiments. There are 851 blue whale calls and 244 fin whale calls of the same sampling frequency 2000Hz. The results are shown and analyzed in Section 6.5.2. For speech recognition test, we use the Aurora 4 corpus [126], which is derived from the Wall Street Journal 5000-word closed vocabulary transcription task. The training set contains 15 hours of speech, and it contains clean speech and speech corrupted by one of six different noises (street traffic, train station, car, babble, restaurant, airport) at 5-15 dB SNR. The test set has 300 utterances from 8 speakers. The sampling rate of this dataset is 16 kHz. The preliminary result is shown in Section 6.5.3.

6.5.2 Spectral Clustering

We use the MATLAB toolbox for the scattering transform [128] with two layers, and $Q_1 = 8$, $Q_2 = 1$, and $T = 120ms$ for the experiments. And for both the DCTNet and PCANet, we use window size 256 for the first layer, and window size 20 for the second layer. The size of the window for the first layer does not have much influence on spectral clustering and classification results for DCTNet and PCANet in the experiments. In order to have a better time-frequency resolution for DCTNet, we choose the window length to be 256 for the first layer. The window size for second layer should be smaller than the first layer, in order to have a finer time scale, and in order to bin frequency to obtain the linear frequency spectrogram like features. We choose a window size of 20 for the second layer of both DCTNet and PCANet.

We use the Laplacian Eigenmap to visualize the three dimensional distribution of features obtained from two-layer DCTNet, PCANet and the scattering transform. The plots are shown in Fig. 6.7. We use 3 nearest neighbors to create the kernel of the Laplacian Eigenmap. We can see that the two-layer DCTNet and PCANet can well separate the blue whale and fin whale data, while the scattering transform fails to separate the data in three dimensions.



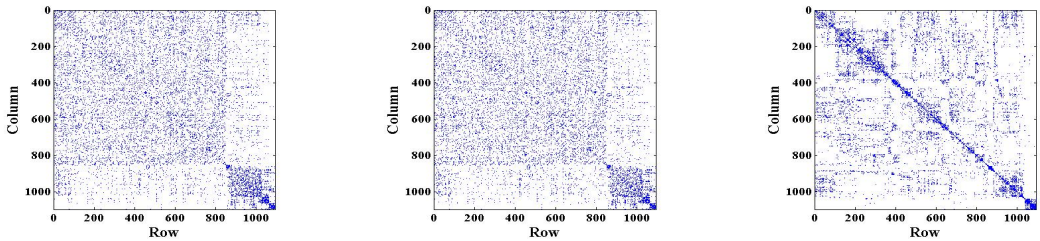
(a) DCTNet (b) PCANet

(c) Scattering transform

FIGURE 6.7: Whale vocalizations data three dimensional mapping

We can examine the adjacency matrices created by the Laplacian Eigenmap. Since it is a binary classification, we use the Fielder value [122] for spectral re-ordering. We can see that there are two blocks in the adjacency matrices for DCTNet and PCANet, and the blue whale and fin whale data are well separated.

Based on the features obtained from PCANet, DCTNet and the scattering transform, the Laplacian Eigenmap is used for dimensional reduction, to compress the features to dimension 1 to 20. We use the kNN classifier ($k = 3$) to evaluate the



(a) DCTNet (b) PCANet (c) Scattering transform

FIGURE 6.8: Comparison of Adjacency matrices

separation of the data. The AUC (area under the curve) versus dimension (1 to 20) plot is shown in Fig. 6.9, and the AUC values for dimension 20 are shown in Table 6.1. We use 5-fold cross validation to generate the plot, that is 681 blue whale vocalizations and 196 fin whale vocalizations for training, and 170 blue whale calls and 48 fin whale calls for testing. We can see that using both DCTNet and PCANet features can achieve high classification rate on the whale vocalization data.

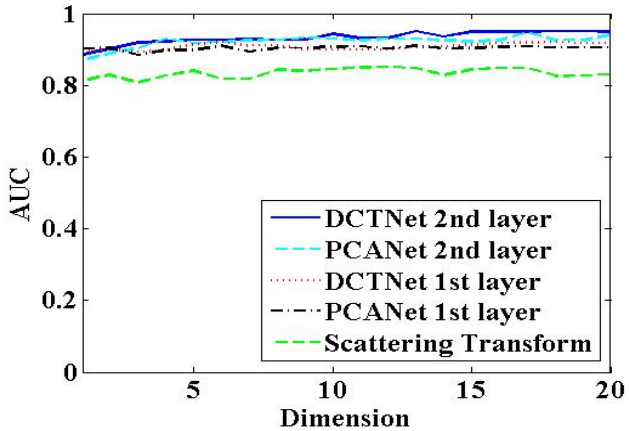


FIGURE 6.9: AUC comparisons

Table 6.1: Area Under the Curves (AUCs) of DCLDE Data Classification

DCTNet 2nd layer	PCANet 2nd layer	DCTNet 1st layer	PCANet 1st layer	Scattering Transform
0.9513	0.9404	0.9200	0.9079	0.8500

6.5.3 Speech Recognition Test

For speech recognition task, we use window size 256 and keep 160 components on the first layer, and use window size 40 and keep all the components on the second layer for DCTNet. We take a *log* for the second layer DCTNet features and apply them to the 7-layer DNN-HMM system [75]. We compare the DCTNet feature with the MFCC feature and MFSC feature. For the MFCC and MFSC feature, we use 12th ordered mel-frequency cepstral coefficients and spectral coefficients, plus differential (delta) and acceleration (delta-delta) features. The preliminary results are shown in Table 6.2.

Table 6.2: Word Error Rate (WER) of Aurora 4 Speech Corpus

Methods	*log*-DCTNet	*log*-MFSC	MFCC
WER	14.9%	13.5%	15.3%

Compared with the whale vocalization data above, speech signals are more dynamic and have more variation. Taking a *log* on the DCTNet output can better represent mel-scale human listening acoustic feature. The MFCC takes DCT for the *log*-MFSC feature. The DCT step can be view as a compression step, and will result in information loss. Therefore, the MFCC features are not as good as the *log*-MFSC features in this task. We have downsampled the *log*-DCTNet output to reduce the feature dimension, therefore, there is a loss of information. Table 6.2 shows that the DCTNet feature is similar to the MFSC, and muti-layer convolution network is able to reveal acoustic content of speech.

6.6 Conclusion

In this chapter, we use the PCANet and propose the DCTNet for acoustic signal classification. We have shown that each layer of the DCTNet and the PCANet is essentially a short time DCT or a short time PCA, respectively, and can reveal different time-frequency content of the signal. Using both the PCANet and the DCTNet as features can achieve high classification rate in the whale vocalization

data, and the DCTNet feature achieves state-of-art performance. The word error rate of the DCTNet is similar to the MFSC in the speech recognition task, suggesting that the convolutional network is able to reveal acoustic content of speech signals.

7

Conclusion and Future Research

Conclusion

This dissertation focuses on the detection and classification of whale vocalizations in the time-frequency plane. In detection (Chapter 3), we have derived and evaluated the probability distributions of a signal's spectrogram, and the Short Time Fourier Transform (STFT) of the signal. We proceed to obtain the corresponding likelihood ratio for detection in the ocean environment both in the case in which the sound speed profile is known exactly, and also when it is uncertain. With the phase information retained, the detector based on the STFT outperforms detectors based on the spectrogram, but it is more sensitive to environmental uncertainty.

In classification, we aim to find a good representation for whale vocalizations. We can extract polynomial phase coefficients of whale vocalizations in the energy spectrogram of the STFT, or the energy distribution of a bilinear time-frequency transform. In this dissertation, we use the Weyl transform (Chapter 4), a transform closely related to the Wigner Ville distribution and the ambiguity function, for feature extraction. The Weyl transform is able to capture chirp rate information. Therefore, with a two dimensional feature set, we are able to represent linear chirp

like whale vocalizations globally. Experimental results show that, on our collected data set, the feature set obtained from the Weyl transform outperforms both the MFCC and the chirplet transform in whale acoustic classification.

The fact that we can represent whale vocalizations based on their polynomial coefficients can be well explained by examining the geometry of the data, that the signal lies on a manifold parameterized by polynomial coefficients. In this dissertation, we use ISOMAP and Laplacian Eigenmap for nonlinear mapping of high dimensional whale acoustic data (Chapter 5), to examine its intrinsic dimension and intrinsic structure. Experimental results show that the nonlinear manifold mapping methods outperform the linear mapping methods, such as PCA and MDS, pointing to the nonlinearity of whale vocalizations.

Since the STFT can be interpreted as passing the signal into a set of filterbanks, we make the connection with convolutional networks, which are fundamental building blocks in the current hot topic of deep learning. We build each layer with either a PCA filter bank (PCANet), and a DCT filter bank (DCTNet) for acoustic feature extraction (Chapter 6). With DCTNet, each layer has a different time-frequency scale representation, and from this, one can extract different physical information. Experimental results show that using both PCANet and DCTNet features can achieve high classification rate in the whale vocalization data, and using the DCTNet feature achieves state-of-art performance. The word error rate of the DCTNet feature is similar to the MFSC in speech recognition tasks, suggesting that the convolutional network is able to reveal acoustic content of speech signals.

Future research

One of the most important challenges related to whale acoustics classification is to find the best features for capturing their information content. Compared with complex human vocalizations, whale calls are relatively simple, so the study of whale

vocalization enables us to have a better understanding for acoustic signal representation in general. By exploiting the data geometry, we can have a better nonlinear mapping for the acoustic signal. Note that for ISOMAP, we have to sample densely to approximate the geodesic distance. This is computationally expensive and the convergence criteria is not necessarily guaranteed (Chapter 5). Finding a better way to represent the geodesic distance is an open area for investigation.

To enhance the computational efficiency of manifold mapping is also of research interest. The computational cost of sampling in ISOMAP is expensive, as stated above. The computational cost of Laplacian Eigenmap is also expensive when the dataset is large. The eigenvalues and eigenvectors of the graph Laplacian have to be computed in order to perform the mapping, which is computationally expensive when the number of data points is large. These factors prohibit the application of nonlinear manifold mapping methods to big data.

For multilayer feature extraction (Chapter 6), we have already discovered that the first layer of DCTNet is actually a short time DCT, and the second layer output can generate features similar to linear frequency spectrogram coefficients (LFSC). The third and fourth layer output are delta LFSC and delta-delta LFSC. The approach for higher order layers and the physical meaning behind them remain to be explored.

Besides signal representation, interpreting the underwater acoustic propagation model using differential geometry is also an extremely interesting direction for future exploration. By studying that, we can have a better understanding of sound propagation and better track the whales and other underwater objects.

We use the ray theory model as an approximation to the sound propagation model [39, 129]. The geodesic structure that determines the eigenray in the high-frequency limit in ray theory is identical to field equations, which is related to the Jacobi field [129]. The convergence of geodesic is equivalent to caustic formation in the acoustic field. With the tools and methods in differential geometry, we can

measure the geometric transmission loss and caustic location without repeatedly solving the ray equation.

Appendix A

Derivation of the Short Time Fourier Transform Detector

A.1 Matched Ocean

According to eq. (3.20), we can see that $\mathbf{X}$ is of dimension $(2(B+1)\times(J+1))\times 1$. Because we use the Gaussian distribution as our noise model, $\mathbf{X}$ follows a multivariate normal distribution under both H_0 and H_1 hypotheses.

Under the H_0 hypothesis, we have

$$\mathbb{E}(\mathbf{X}) = \mathbf{0}_{2(B+1)\times(J+1))\times 1}. \tag{A.1}$$

The covariance matrix $\mathbf{C_x}$ of $\mathbf{X}$ can be represented as [130]

$$\mathbf{C_x} = \begin{bmatrix} \mathbf{C_{UU}} & \mathbf{C_{UV}} \\ \mathbf{C_{VU}} & \mathbf{C_{VV}} \end{bmatrix}, \tag{A.2}$$

where,

$$\mathbf{C_{UU}} = \begin{bmatrix} Var(U[0,0]) & \cdots & Cov(U[0,0],U[J,B]) \\ \vdots & \ddots & \vdots \\ Cov(U[J,B],U[0,0]) & \cdots & Var(U[J,B]) \end{bmatrix},$$

$$\mathbf{C_{VV}} = \begin{bmatrix} Var(V[0,0]) & \cdots & Cov(V[0,0],V[J,B]) \\ \vdots & \ddots & \vdots \\ Cov(V[J,B],V[0,0]) & \cdots & Var(V[J,B]) \end{bmatrix},$$

$$\mathbf{C_{UV}} = \begin{bmatrix} Cov(U[0,0], V[0,0]) & \cdots & Cov(U[0,0], V[J,B]) \\ \vdots & \ddots & \vdots \\ Cov(U[J,B], V[0,0]) & \cdots & Cov(U[J,B], V[J,B]) \end{bmatrix},$$

$$\mathbf{C_{VU}} = \begin{bmatrix} Cov(V[0,0], U[0,0]) & \cdots & Cov(V[0,0], U[J,B]) \\ \vdots & \ddots & \vdots \\ Cov(V[J,B], U[0,0]) & \cdots & Cov(V[J,B], U[J,B]) \end{bmatrix}$$

For $a_1, a_2 = 0, 1, \cdots, J$ and $b_1, b_2 = 0, 1, \cdots, B$, applying the rectangular window function $w[i] = 1,$ for $i = 0, \cdots, M-1$, we have

$$\begin{aligned} &Cov(U[a_1, b_1], U[a_2, b_2]) \\ =&\sigma_n^2 \sum_{i_1=0}^{M-1} \sum_{i_2=0}^{M-1} \delta[a_1 D + i_1 - a_2 D - i_2] \cos\left(\frac{2\pi b_1 i_1}{M}\right) \cos\left(\frac{2\pi b_2 i_2}{M}\right) \end{aligned}$$

$$\begin{aligned} &Cov(V[a_1, b_1], V[a_2, b_2]) \\ =&\sigma_n^2 \sum_{i_1=0}^{M-1} \sum_{i_2=0}^{M-1} \delta[a_1 D + i_1 - a_2 D - i_2] \sin\left(\frac{2\pi b_1 i_1}{M}\right) \sin\left(\frac{2\pi b_2 i_2}{M}\right) \end{aligned}$$

$$\begin{aligned} &Cov(U[a_1, b_1], V[a_2, b_2]) \\ =& -\sigma_n^2 \sum_{i_1=0}^{M-1} \sum_{i_2=0}^{M-1} \delta[a_1 D + i_1 - a_2 D - i_2] \cos\left(\frac{2\pi b_1 i_1}{M}\right) \sin\left(\frac{2\pi b_2 i_2}{M}\right) \end{aligned}$$

$$\begin{aligned} &Cov(V[a_1, b_1], U[a_2, b_2]) \\ =& -\sigma_n^2 \sum_{i_1=0}^{M-1} \sum_{i_2=0}^{M-1} \delta[a_1 D + i_1 - a_2 D - i_2] \sin\left(\frac{2\pi b_1 i_1}{M}\right) \cos\left(\frac{2\pi b_2 i_2}{M}\right). \end{aligned}$$

When the covariance matrix is singular, we apply spectral decomposition to the covariance matrix

$$\mathbf{C_x} = Q \Lambda Q^T$$

where Q is a unitary orthogonal matrix, and Λ is a diagonal matrix. The diagonal elements in Λ are all real-valued. Suppose there are k non-zero eigenvalues in Λ,

then the covariance matrix $\mathbf{C_x}$ is:

$$\begin{aligned}\mathbf{C_x} =& Q\Lambda Q^T \\ =& [Q_1 \;\; Q_2]\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}\begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} \\ =& Q_1\Lambda_1 Q_1^T + Q_2\Lambda_2 Q_2^T\end{aligned}$$

where Λ_1 and Q_1 corresponds to the k non-zero eigenvalues components, and Λ_2 and Q_2 corresponds to the zero components.

When $\mathbf{C_x}$ is singular, the probability density function for $\mathbf{X}$ does not exist. Mapping $\mathbf{X}$ into a subspace formed by Q_1, the probability density function for $Q_1^T\mathbf{X}$ is

$$p(Q_1^T\mathbf{X}) = \frac{1}{(2\pi)^{(B+1)(J+1)}\det^{1/2}(\Lambda_1)}\exp\left(-\frac{1}{2}\mathbf{X^T}Q_1\Lambda_1^{-1}Q_1^T\mathbf{X}\right). \tag{A.3}$$

Under the H_1 hypothesis, letting

$$\mu = \mathbb{E}(\mathbf{X}),$$

we know that

$$\begin{aligned}\mu =& [\sum_{i=0}^{M-1} y[0\cdot D+i]\cos(2\pi 0\cdot i/M),\ldots,\sum_{i=0}^{M-1} y[J\cdot D+i]\cos(2\pi(M-1)\cdot i/M), \\ & \sum_{i=0}^{M-1} y[0\cdot D+i]\sin(-2\pi 0\cdot i/M),\ldots,\sum_{i=0}^{M-1} y[J\cdot D+i]\sin(-2\pi(M-1)\cdot i/M)]^T\end{aligned}$$

and,

$$\mathbb{E}(Q_1^T\mathbf{X}) = Q_1^T\mu.$$

It can be proved that the covariance matrix $\mathbf{C_x}$ under the H_1 hypothesis is the same as the one under the H_0 hypothesis. Therefore, the probability density function under the H_1 hypothesis for $Q_1^T\mathbf{X}$ is

$$p(Q_1^T\mathbf{X}) = \frac{1}{(2\pi)^{(B+1)(J+1)}\det^{1/2}(\Lambda_1)}\exp\left(-\frac{1}{2}(\mathbf{X}-\mu)^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T(\mathbf{X}-\mu)\right). \tag{A.4}$$

Based on eq. (A.3) and eq. (A.4), we find the likelihood ratio based on STFT data, when the environmental parameters and source signal is known exactly, to be

$$\begin{aligned}\lambda =& \frac{p(Q_1^T\mathbf{X}|H_1)}{p(Q_1^T\mathbf{X}|H_0)} \\ =& \exp\left(-\frac{1}{2}(\mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu - 2\mathbf{X}^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu)\right).\end{aligned} \tag{A.5}$$

Therefore, the log likelihood ratio is

$$\ln\lambda = -\frac{1}{2}\left(\mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu - 2\mathbf{X}^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu\right). \tag{A.6}$$

Because $\mathbf{X}$ follows a multivariate normal distribution under both hypotheses, $Q_1^T\mathbf{X}$ also follows a multivariate normal distribution. We can derive the log likelihood distribution under both hypothesis, and derive the analytic solution for the ROC plots:

$$\begin{aligned}\mu_1 =& \mathbb{E}(\ln\lambda|H_1) = \frac{1}{2}(\mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu) \\ \sigma_1^2 =& Var(\ln\lambda|H_1) \\ =& (Q_1\Lambda_1^{-1}Q_1^T\mu)^T Cov(\mathbf{X})Q_1\Lambda_1^{-1}Q_1^T\mu \\ =& (Q_1\Lambda_1^{-1}Q_1^T\mu)^T(Q_1\Lambda_1Q_1^T + Q_2\Lambda_2Q_2^T)Q_1\Lambda_1^{-1}Q_1^T\mu \\ =& \mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu \\ \mu_0 =& \mathbb{E}(\ln\lambda|H_0) = -\frac{1}{2}(\mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu) \\ \sigma_0^2 =& Var(\ln\lambda|H_0) = Var(\ln\lambda|H_1) = \mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu.\end{aligned}$$

Therefore,

$$\ln\lambda|H_1 \sim \mathcal{N}\left(\frac{1}{2}(\mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu),\ \mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu\right) \tag{A.7}$$

$$\ln\lambda|H_0 \sim \mathcal{N}\left(-\frac{1}{2}(\mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu),\ \mu^{\mathbf{T}}Q_1\Lambda_1^{-1}Q_1^T\mu\right). \tag{A.8}$$

The probability of correct detection P_D and the probability of false alarm P_F in this case are [103]:

$$P_D = \int_{\ln\beta}^{\infty} p(\ln\lambda|H_1)d(\ln\lambda) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\int_{\ln\beta}^{\infty} e^{-\frac{(\ln\lambda-\mu_1)^2}{2\sigma_1^2}} d(\ln\lambda)$$

$$P_F = \int_{\ln\beta}^{\infty} p(\ln\lambda|H_0)d(\ln\lambda) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\int_{\ln\beta}^{\infty} e^{-\frac{(\ln\lambda-\mu_0)^2}{2\sigma_0^2}} d(\ln\lambda).$$

Since $\sigma_1^2 = \sigma_0^2$, the standard deviation in this case is positive, so $\sigma_1 = \sigma_0$. Letting $\sigma = \sigma_1 = \sigma_0$, $z = \frac{(\ln\lambda-\mu_1)}{\sigma}$, $z' = \frac{(\ln\lambda-\mu_0)}{\sigma}$, $\sigma' = (\ln\beta - \mu_0)/\sigma$, so $dz = \sigma d(\ln\lambda)$, $dz' = \sigma d(\ln\lambda)$, we have

$$P_F = \frac{1}{\sqrt{2\pi}}\int_{\sigma'}^{\infty} e^{-\frac{z'^2}{2}} dz', \tag{A.9}$$

$$P_D = \frac{1}{\sqrt{2\pi}}\int_{\sigma'-\frac{\mu_1-\mu_0}{\sigma}}^{\infty} e^{-\frac{z^2}{2}} dz, \tag{A.10}$$

and the separation parameter, or detection index, d in this case is

$$d = \frac{\mu_1 - \mu_0}{\sigma},$$

where,

$$\begin{aligned} d^2 &= \frac{(\mu_1-\mu_0)^2}{(\sigma)^2} \\ &= \frac{(\mu^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu)^2}{\mu^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu} = \mu^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu. \end{aligned} \tag{A.11}$$

A.2 Mean Ocean

The following is a proof that when the uncertainty of the sound speed profile is small, we can theoretically approximate the likelihood ratio using eq. (3.26). Suppose p_k is the probability of the sound speed profile of the kth path; $\mu_{\mathbf{k}}$ is the propagated signal corresponding to the kth sound speed profile, and $\mu_{\mathbf{m}}$ is the propagated signal of

the mean sound speed profile. According to eq. (3.3), we can see that the mapping of the sound speed and propagated signal is linear, so we have $\mu_{\mathbf{m}} = \sum_{k=1}^{P} p_k \mu_{\mathbf{k}}$. According to eq. (3.23), the geometric mean of the likelihood ratio is

$$\begin{aligned}\prod_{k=1}^{P} \lambda_k^{p_k} &= \exp\left(-\frac{1}{2}\left(\sum_{k=1}^{P} p_k \left(\mu_{\mathbf{k}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{k}} - 2\mathbf{X}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{k}}\right)\right)\right) \\ &= \exp\left(\left(-\frac{1}{2}\sum_{k=1}^{P} p_k \mu_{\mathbf{k}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{k}}\right) + \mathbf{X}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{m}}\right)\end{aligned}$$

For the term $-\frac{1}{2}\sum_{k=1}^{P} p_k \mu_{\mathbf{k}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{k}}$, notice that:

$$\begin{aligned}&\sum_{k=1}^{P} p_k \mu_{\mathbf{k}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{k}} \\ =&\sum_{k=1}^{P} p_k \mu_{\mathbf{k}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T (\mu_{\mathbf{k}} - \mu_{\mathbf{m}}) \\ =&\sum_{k=1}^{P} p_k (\mu_{\mathbf{k}}^{\mathbf{T}} - \mu_{\mathbf{m}}^{\mathbf{T}}) Q_1 \Lambda_1^{-1} Q_1^T (\mu_{\mathbf{k}} - \mu_{\mathbf{m}}) + \sum_{k=1}^{P} p_k \mu_{\mathbf{m}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T (\mu_{\mathbf{k}} - \mu_{\mathbf{m}}) \\ =&\sum_{k=1}^{P} p_k (\mu_{\mathbf{k}}^{\mathbf{T}} - \mu_{\mathbf{m}}^{\mathbf{T}}) Q_1 \Lambda_1^{-1} Q_1^T (\mu_{\mathbf{k}} - \mu_{\mathbf{m}}) \\ \leqslant& C \sum_{k=1}^{P} p_k ||\mu_{\mathbf{k}} - \mu_{\mathbf{m}}||^2 = CVar(\mu_{\mathbf{k}})\end{aligned}$$

where C is the largest eigenvalue of Λ_1^{-1}. When the variance of μ_k is small, we can approximate $\sum_{k=1}^{P} p_k \mu_{\mathbf{k}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{k}}$ with $\sum_{k=1}^{P} p_k \mu_{\mathbf{m}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{m}}$. Therefore,

$$\prod_{k=1}^{P} \lambda_k^{p_k} \approx \exp\left(-\frac{1}{2}\left(\mu_{\mathbf{m}}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{m}} - 2\mathbf{X}^{\mathbf{T}} Q_1 \Lambda_1^{-1} Q_1^T \mu_{\mathbf{m}}\right)\right).$$

We can see that $\mu_{\mathbf{m}}$ in eq. (3.26) is the geometric mean of the likelihood ratio.

Bibliography

[1] S. Pompa, P. R. Ehrlich, and G. Ceballos, "Global distribution and conservation of marine mammals," *Proceedings of the National Academy of Sciences*, vol. 108, no. 33, pp. 13600–13605, 2011.

[2] T. A. Jefferson, M. A. Webber, and R. L. Pitman, *Marine Mammals of the World: A Comprehensive Guide to Their Identification*. Academic Press, 2011.

[3] Marine Mammal Taxonomy List, "List of marine mammal species & subspecies,"

[4] W. M. Zimmer, *Passive acoustic monitoring of cetaceans*. Cambridge University Press, 2011.

[5] Nature Spotlights, "Acoustic pollution and marine mammals,"

[6] C. Welch, "Smart and fast marine mammals are guarding our military bases,"

[7] J. Berg, "Sound production in isolated human larynges," *Annals of the New York Academy of Sciences*, vol. 155, no. 1, pp. 18–27, 1968.

[8] K. S. Norris, "The evolution of acoustic mechanisms in odontocete cetaceans," *Evolution and environment*, pp. 297–324, 1968.

[9] D. Wartzok and D. R. Ketten, "Marine mammal sensory systems," *Biology of marine mammals*, vol. 1, p. 117, 1999.

[10] N. Gandilhon, O. Adam, D. Cazau, J. T. Laitman, and J. S. Reidenberg, "Two new theoretical roles of the laryngeal sac of humpback whales," *Marine Mammal Science*, vol. 31, no. 2, pp. 774–781, 2015.

[11] Discovery of Sound in the Sea, "How do marine mammals produce sounds?,"

[12] F. Mandal, *Textbook of Animal Behavior, Third Edition*. 2015.

[13] V. M. Janik, L. S. Sayigh, and R. Wells, "Signature whistle shape conveys identity information to bottlenose dolphins," *Proceedings of the National Academy of Sciences*, vol. 103, no. 21, pp. 8293–8297, 2006.

[14] M. J. Noad, D. H. Cato, M. Bryden, M.-N. Jenner, and K. C. S. Jenner, "Cultural revolution in whale songs," *Nature*, vol. 408, no. 6812, pp. 537–537, 2000.

[15] M. D. Beecher, "Spectrographic analysis of animal vocalizations: implications of the uncertainty principle," *Bioacoustics*, vol. 1, no. 2-3, pp. 187–208, 1988.

[16] I. R. Urazghildiiev and C. W. Clark, "Acoustic detection of North Atlantic right whale contact calls using the generalized likelihood ratio test," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1956–1963, 2006.

[17] M. Goodwin and M. Vetterli, "Time-frequency signal models for music analysis, transformation, and synthesis," in *Time-Frequency and Time-Scale Analysis, 1996., Proceedings of the IEEE-SP International Symposium on*, pp. 133–136, IEEE, 1996.

[18] L. Cohen, *Time-frequency analysis*, vol. 1. Prentice hall, 1995.

[19] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[20] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, no. 7, pp. 674–693, 1989.

[21] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *Information Theory, IEEE Transactions on*, vol. 36, no. 5, pp. 961–1005, 1990.

[22] G. Strang, "Wavelets and dilation equations: A brief introduction," *SIAM review*, vol. 31, no. 4, pp. 614–627, 1989.

[23] P. M. Woodward and I. Davies, "Xcii. a theory of radar information," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 41, no. 321, pp. 1001–1017, 1950.

[24] P. M. Woodward, *Probability and Information Theory, with Applications to Radar*. Pergamon Press, 1953.

[25] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Physical Review*, vol. 40, no. 5, p. 749, 1932.

[26] J. d. Ville *et al.*, "Théorie et applications de la notion de signal analytique," *Cables et transmission*, vol. 2, no. 1, pp. 61–74, 1948.

[27] S. Mann and S. Haykin, "The chirplet transform: A generalization of gabors logon transform," in *Vision Interface*, vol. 91, pp. 205–212, 1991.

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

[29] M. Bahoura and Y. Simard, "Chirplet transform applied to simulated and real blue whale (balaenoptera musculus) calls," in *Image and Signal Processing*, pp. 296–303, Springer, 2008.

[30] M. Bahoura and Y. Simard, "Blue whale calls classification using short-time Fourier and wavelet packet transforms and artificial neural network," *Digital Signal Processing*, vol. 20, no. 4, pp. 1256–1263, 2010.

[31] O. Adam, "Advantages of the Hilbert Huang transform for marine mammals signals analysis," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2965–2973, 2006.

[32] P. Flandrin, *Time-frequency/time-scale analysis*, vol. 10. Academic press, 1998.

[33] P. Flandrin and P. Borgnat, "Time-frequency energy distributions meet compressed sensing," *Signal Processing, IEEE Transactions on*, vol. 58, no. 6, pp. 2974–2982, 2010.

[34] D. Mellinger and C. Clark, "Mobysound: A reference archive for studying automatic recognition of marine mammal sounds," *Applied Acoustics*, vol. 67, no. 11, pp. 1226–1242, 2006.

[35] F. Auger, P. Flandrin, P. Gonçalvès, and O. Lemoine, "Time-frequency toolbox," *CNRS France-Rice University*, 1996.

[36] W. C. Cummings and D. Holliday, "Passive acoustic location of bowhead whales in a population census off point barrow, alaska," *The Journal of the Acoustical Society of America*, vol. 78, no. 4, pp. 1163–1169, 1985.

[37] L. E. Freitag and P. L. Tyack, "Passive acoustic localization of the atlantic bottlenose dolphin using whistles and echolocation clicks," *The Journal of the Acoustical Society of America*, vol. 93, no. 4, pp. 2197–2205, 1993.

[38] K. M. Stafford, C. G. Fox, and D. S. Clark, "Long-range acoustic detection and localization of blue whale calls in the northeast pacific ocean," *The Journal of the Acoustical Society of America*, vol. 104, no. 6, pp. 3616–3625, 1998.

[39] F. B. Jensen, *Computational ocean acoustics*. Springer Science & Business Media, 1994.

[40] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*. Simon & Schuster, 1992.

[41] R. P. Hodges, *Underwater acoustics: Analysis, design and performance of sonar*. John Wiley & Sons, 2011.

[42] D. Rouseff, D. R. Jackson, W. L. Fox, C. D. Jones, J. Ritcey, and D. R. Dowling, "Underwater acoustic communication by passive-phase conjugation: Theory and experimental results," *Oceanic Engineering, IEEE Journal of*, vol. 26, no. 4, pp. 821–831, 2001.

[43] P. C. Etter, *Underwater acoustic modeling and simulation*. CRC Press, 2013.

[44] M. B. Porter, "The bellhop manual and users guide: Preliminary draft," *Heat, Light, and Sound Research, Inc., La Jolla, CA, USA, Tech. Rep*, 2011.

[45] M. B. Porter, "The KRAKEN normal mode program," tech. rep., DTIC Document, 1992.

[46] M. Lopatka, A. Olivier, C. Laplanche, J. Zarzycki, and J.-F. Motsch, "An attractive alternative for sperm whale click detection using the wavelet transform in comparison to the fourier spectrogram," *Aquatic Mammals*, vol. 31, no. 4, p. 463, 2005.

[47] O. Adam, "Segmentation of killer whale vocalizations using the hilbert-huang transform," *Eurasip Journal on Advances in Signal Processing*, vol. 2008, p. 162, 2008.

[48] D. K. Mellinger and C. W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3518–3529, 2000.

[49] D. K. Mellinger and C. W. Clark, "Methods for automatic detection of mysticete sounds," *Marine & Freshwater Behaviour & Phy*, vol. 29, no. 1-4, pp. 163–181, 1997.

[50] D. Gillespie, "Detection and classification of right whale calls using an'edge'detector operating on a smoothed spectrogram," *Canadian Acoustics*, vol. 32, no. 2, pp. 39–47, 2004.

[51] D. K. Mellinger, S. W. Martin, R. P. Morrissey, L. Thomas, and J. J. Yosco, "A method for detecting whistles, moans, and other frequency contour sounds," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 4055–4061, 2011.

[52] J. R. Buck and P. L. Tyack, "A quantitative measure of similarity for tursiopstruncatus signature whistles," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2497–2506, 1993.

[53] S. Madhusudhana, E. M. Oleson, M. S. Soldevilla, M. Roch, J. Hildebrand, *et al.*, "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, 2008.

[54] S. Madhusudhana, E. M. Oleson, M. S. Soldevilla, M. Roch, J. Hildebrand, *et al.*, "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, 2008.

[55] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[56] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.

[57] K. K. Paliwal and L. D. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[58] M. Pechenizkiy, "The impact of feature extraction on the performance of a classifier: knn, naïve bayes and c4. 5," in *Advances in Artificial Intelligence*, pp. 268–279, Springer, 2005.

[59] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *The Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[60] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[61] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[62] J. Kruskal and M. Wish, *Multidimensional scaling*, vol. 11. Sage, 1978.

[63] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[64] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[65] J. Tenenbaum, V. De Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[66] R. Coifman and S. Lafon, "Diffusion maps," *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[67] J. C. O'Neill, P. Flandrin, and W. C. Karl, "Sparse representations with Chirplets via maximum likelihood estimation," 2000.

[68] B. Logan, "Mel frequency cepstral coefficients for music modeling.," in *ISMIR*, 2000.

[69] F. Pace, F. Benard, H. Glotin, O. Adam, and P. White, "Subunit definition and analysis for humpback whale call classification," *Applied Acoustics*, vol. 71, no. 11, pp. 1107–1112, 2010.

[70] M. A. Roch, M. S. Soldevilla, J. C. Burtenshaw, E. E. Henderson, and J. A. Hildebrand, "Gaussian mixture model classification of odontocetes in the southern california bight and the gulf of california," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1737–1748, 2007.

[71] M. A. Roch, H. Klinck, S. Baumann-Pickering, D. K. Mellinger, S. Qui, M. S. Soldevilla, and J. A. Hildebrand, "Classification of echolocation clicks from odontocetes in the southern california bight," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 467–475, 2011.

[72] X. C. Halkias, S. Paris, and H. Glotin, "Classification of mysticete sounds using machine learning techniques," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3496–3505, 2013.

[73] D. Nouri, "Using deep learning to listen for whales," *http://danielnouri.org/notes/2014/01/10/using-deep-learning-to-listen-for-whales/*.

[74] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, 1995.

[75] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[76] L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, and J. Williams, "Recent advances in deep learning for speech research at

microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8604–8608, IEEE, 2013.

[77] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, pp. 1096–1104, 2009.

[78] J. Andén and S. Mallat, "Deep scattering spectrum," *Signal Processing, IEEE Transactions on*, vol. 62, no. 16, pp. 4114–4128, 2014.

[79] A. Jansen and P. Niyogi, "A geometric perspective on speech sounds," *University of Chicago, Tech. Rep*, 2005.

[80] A. Errity and J. McKenna, "An investigation of manifold learning for speech analysis.," in *INTERSPEECH*, Citeseer, 2006.

[81] J. Kim, S. Lee, and S. Narayanan, "An exploratory study of manifolds of emotional speech," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 5142–5145, IEEE, 2010.

[82] K. Gröchenig, *Foundations of time-frequency analysis*. Springer Science & Business Media, 2013.

[83] A. Oppenheim, *Applications of digital signal processing*. Prentice-Hall signal processing series, Prentice-Hall, 1978.

[84] J. V. Bouvrie and T. Ezzat, "An incremental algorithm for signal reconstruction from short-time fourier transform magnitude.,"

[85] L. Cohen, "Wavelet moments and time-frequency analysis," in *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pp. 434–445, International Society for Optics and Photonics, 1999.

[86] S. Mann and S. Haykin, "The chirplet transform: Physical considerations," *Signal Processing, IEEE Transactions on*, vol. 43, no. 11, pp. 2745–2761, 1995.

[87] M. A. Richards, *Fundamentals of radar signal processing*. Tata McGraw-Hill Education, 2005.

[88] H. I. Choi and W. J. Williams, "Improved time-frequency representation of multicomponent signals using exponential kernels," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 6, pp. 862–871, 1989.

[89] M. Born and P. Jordan, "Zur quantenmechanik," *Zeitschrift für Physik*, vol. 34, no. 1, pp. 858–888, 1925.

[90] G. Casella and R. L. Berger, *Statistical inference*, vol. 2. Duxbury Pacific Grove, CA, 2002.

[91] H. L. Van Trees, *Detection, estimation, and modulation theory.* John Wiley & Sons, 2004.

[92] R. A. Altes, "Detection, estimation, and classification with spectrograms," *The Journal of the Acoustical Society of America*, vol. 67, no. 4, pp. 1232–1246, 1980.

[93] J. Huillery, F. Millioz, and N. Martin, "On the description of spectrogram probabilities with a chi-squared law," *Signal Processing, IEEE Transactions on*, vol. 56, no. 6, pp. 2249–2258, 2008.

[94] J. Huillery, F. Millioz, and N. Martin, "On the probability distributions of spectrogram coefficients for correlated gaussian process," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 3, pp. III–III, IEEE, 2006.

[95] J. Preisig, "Acoustic propagation considerations for underwater acoustic communications network development," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 11, no. 4, pp. 2–10, 2007.

[96] J. Shorey and L. Nolte, "Wideband optimal a posteriori probability source localization in an uncertain shallow ocean environment," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 355–361, 1998.

[97] P. J. Book and L. Nolte, "Narrow-band source localization in the presence of internal waves for 1000-km range and 25-hz acoustic frequency," *The Journal of the Acoustical Society of America*, vol. 101, no. 3, pp. 1336–1346, 1997.

[98] S. L. Tantum and L. W. Nolte, "On array design for matched-field processing," *The Journal of the Acoustical Society of America*, vol. 107, no. 4, pp. 2101–2111, 2000.

[99] L. Sha and L. W. Nolte, "Effects of environmental uncertainties on sonar detection performance prediction," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 1942–1953, 2005.

[100] M. Wazenski and D. Alexandrou, "Active, wideband detection and localization in an uncertain multipath environment," *The Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 1961–1970, 1997.

[101] V. Trygonis, E. Gerstein, J. Moir, and S. McCulloch, "Vocalization characteristics of north atlantic right whale surface active groups in the calving habitat, southeastern united states," *The Journal of the Acoustical Society of America*, vol. 134, no. 6, pp. 4518–4531, 2013.

[102] M. Siderius and M. B. Porter, "Modeling broadband ocean acoustic transmissions with time-varying sea surfaces," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 137–150, 2008.

[103] S. M. Kay, "Fundamentals of statistical signal processing: Detection theory, vol. 2," 1998.

[104] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence*, vol. 116, pp. 374–388, 1976.

[105] J. C. O'Neill and P. Flandrin, "Chirp hunting," in *Time-Frequency and Time-Scale Analysis, 1998. Proceedings of the IEEE-SP International Symposium on*, pp. 425–428, IEEE, 1998.

[106] L. Applebaum, S. D. Howard, S. Searle, and R. Calderbank, "Chirp sensing codes: Deterministic compressed sensing measurements for fast recovery," *Applied and Computational Harmonic Analysis*, vol. 26, no. 2, pp. 283–290, 2009.

[107] S. D. Howard, A. R. Calderbank, and W. Moran, "The finite Heisenberg-Weyl groups in radar and communications," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, 2006.

[108] S. Howard, A. Calderbank, and W. Moran, "Finite Heisenberg-Weyl groups and Golay complementary sequences," tech. rep., DTIC Document, 2006.

[109] M. Esfahanian, H. Zhuang, and N. Erdol, "Sparse representation for classification of dolphin whistles by type," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. EL1–EL7, 2014.

[110] E. J. Candes, P. R. Charlton, and H. Helgason, "Detecting highly oscillatory signals by chirplet path pursuit," *Applied and Computational Harmonic Analysis*, vol. 24, no. 1, pp. 14–40, 2008.

[111] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music," in *Seventh International Conference on Music Information Retrieval (ISMIR)*, 2006.

[112] E. J. Candes, "Multiscale chirplets and near-optimal recovery of chirps," tech. rep., Technical Report, Stanford University, 2002.

[113] E. Chassande-Mottin, A. Pai, *et al.*, "Discrete time and frequency Wigner-Ville distribution: Moyal's formula and aliasing," *IEEE Signal Processing Letters*, vol. 12, no. 7, p. 508, 2005.

[114] S. Peleg and B. Porat, "Estimation and classification of polynomial-phase signals," *Information Theory, IEEE Transactions on*, vol. 37, no. 2, pp. 422–430, 1991.

[115] S. Peleg and B. Friedlander, "The discrete polynomial-phase transform," *Signal Processing, IEEE Transactions on*, vol. 43, no. 8, pp. 1901–1914, 1995.

[116] M. Cheney and B. Borden, *Fundamentals of radar imaging*, vol. 79. SIAM, 2009.

[117] S. Barbarossa, A. Scaglione, , and G. B. Giannakis, "Product high-order ambiguity function for multicomponent polynomial-phase signal modeling," *IEEE Trans. on Signal Processing*, pp. 691–708, 1998.

[118] B. Weisburn and T. Parks, "Design of time frequency strip filters," in *Signals, Systems and Computers, 1995. 1995 Conference Record of the Twenty-Ninth Asilomar Conference on*, vol. 2, pp. 930–934 vol.2, Oct 1995.

[119] M. Brookes, *VOICEBOX: a MATLAB toolbox for speech processing.* 2003.

[120] Scripps Institution of Oceanography, "DCLDE conference low frequency data," *http://www.cetus.ucsd.edu/dclde/dataset.html*, 2015.

[121] Y. Yao, "A mathematical introduction to data science," 2004.

[122] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[123] T. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?," *arXiv preprint arXiv:1404.3606*, 2014.

[124] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *Computers, IEEE Transactions on*, vol. 100, no. 1, pp. 90–93, 1974.

[125] V. Sánchez, P. Garcia, A. Peinado, J. Segura, and A. Rubio, "Diagonalizing properties of the discrete cosine transforms," *Signal Processing, IEEE Transactions on*, vol. 43, no. 11, pp. 2631–2641, 1995.

[126] N. Parihar and J. Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," *Inst. for Signal and Information Process, Mississippi State University, Tech. Rep*, vol. 40, p. 94, 2002.

[127] T. Chan, "An optimal circulant preconditioner for toeplitz systems," *SIAM journal on scientific and statistical computing*, vol. 9, no. 4, pp. 766–771, 1988.

[128] J. Anden and S. Mallat, "Scatnet (v0.2)," *http://www.di.ens.fr/data/software/scatnet/*.

[129] D. R. Bergman, "Application of differential geometry to acoustics: Development of a generalized paraxial ray-trace procedure from geodesic deviation," tech. rep., DTIC Document, 2005.

[130] S. K. Sengijpta, "Fundamentals of statistical signal processing: Estimation theory," *Technometrics*, vol. 37, no. 4, pp. 465–466, 1995.

Biography

Yin Xian was born in Guangzhou, China on August 3, 1986. She obtained her Bachelor of Science in Electrical Engineering at Hunan University in June 2009, and her Master of Science in Electrical Engineering from Duke University in May 2013.