On Eigendecomposition-based Algorithms as Feature Extraction Techniques used with Hidden Markov Model for the Detection of Whale Vocalisations

by

Ayinde Mohammed Usman

Dissertation presented for the degree of
Doctor of Philosophy in Electrical and Electronic Engineering
in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof. D.J.J. Versfeld

March 2024

Declaration

By submitting this dissertation electronically, I, Ayinde Mohammed Usman, declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extend explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2024.

Abstract

Whales emit a variety of distinctive sound signals for communication, echolocation, and other social functions, which are gathered through passive acoustic monitoring (PAM). Different automated methods have been proposed in the literature for analysing PAM datasets to detect and classify whale species, including the use of the hidden Markov model (HMM). This thesis proposes eigendecomposition-based (ED) algorithms as feature extraction (FE) techniques used with HMM for the detection of whale vocalisations. Specifically, the principal components analysis (PCA) and the dynamic mode decomposition (DMD) are deployed to extract the latent underlying characteristics of whale signals in PAM datasets. In addition, enhanced FE techniques are proposed through the kernelisation of PCA and DMD.

The emerging ED-based hidden Markov models (ED-HMMs): PCA-HMM, kPCA-HMM, DMD-HMM, and kDMD-HMM are grouped according to the underlying algorithm deployed for the FE: the PC-based hidden Markov models (PC-HMMs) and the DMD-based hidden Markov model (DM-HMMs). Each of the models is tested on PAM datasets containing southern right whale (SRW) and humpback whale (HW) vocalisations. Their performances are evaluated using metrics such as the true positive rate (TPR), precision (PREC), error rate (ERR), and F_1 scores. Performance outcomes vary subject to different experimental conditions like the dimension of the feature vectors, the size of the training data, and the species vocalisations.

The models demonstrated good performance across different evaluation metrics. For the PC-HMMs, the kPCA-HMM did not only outperform the PCA-HMM in terms of TPR and PREC, but it also exhibited a lower ERR. However, the kCPA-HMM exhibits a higher computational cost when compared to the PCA-HMM. Similarly, for DM-HMMs, the kDMD-HMM outperformed the DMD-HMM in terms of TPR and PREC, and it also exhibited lower ERR as well as a lower computational cost. The comparison showed that PC-HMMs stabilised faster than DM-HMMs in terms of performance. Thus, the PC-HMMs are less complex than the DM-HMMs in terms of dimension. However, the DM-HMMs outperformed the PC-HMMs, albeit at higher

dimensions. The reliability of the developed models was confirmed with F_1 scores, as all the models achieved F_1 scores > 0.9 at their respective optimal dimensions.

Lastly, the results of the proposed ED-HMMs are compared with the existing FE techniques used with HMM in the literature for the detection of whale vocalisations. The ED-HMMs do outperform the existing HMM methods. A general observation is that every model displays better performance with an increase in the number of samples deployed for training. Hence, large window sizes are recommended for model training. The different experimental results showed that a model's performance must be evaluated on a species-to-species basis. It is also important that the training data be a subset of the datasets for testing, or at least using recordings from the same region. This is to avoid bias that may arise from the variation that does exist between the vocalisations of the same species. The ED-HMMs proposed in this study can be further tested on other whale vocalisations to confirm their robustness. Besides, they can be explored by researchers working on the automatic detection of other vocalising animal species.

Opsomming

Walvisse straal 'n verskeidenheid kenmerkende klankseine uit vir kommunikasie, eggolokalisering en ander sosiale funksies, wat deur passiewe akoestiese monitering (PAM) versamel word. Verskillende geoutomatiseerde metodes is in die literatuur voorgestel vir die ontleding van PAM-datastelle om walvisspesies op te spoor en te klassifiseer, insluitend die gebruik van die verborge Markov-model (HMM). Hierdie tesis stel eie-ontbinding-gebaseerde (ED) algoritmes voor as kenmerk ekstraksie (FE) tegnieke wat met HMM gebruik word vir die opsporing van walvisvokalisering. Spesifiek, die hoofkomponent-analise (PCA) en die dinamiese modus-ontbinding (DMD) word ontplooi om die latente onderliggende kenmerke van walvisseine in PAM-datastelle te onttrek. Boonop word verbeterde FE-tegnieke voorgestel deur die kernelisering van PCA en DMD.

Die opkomende ED-gebaseerde versteekte Markov-modelle (ED-HMMs): PCA-HMM, kPCA-HMM, DMD-HMM en kDMD-HMM word gegroepeer volgens die onderliggende algoritme wat vir die FE ontplooi is: die PC-gebaseerde versteekte Markov-modelle (PC-HMM's) en die DMD-gebaseerde versteekte Markov-model (DM-HMM's). Elkeen van die modelle word getoets op PAM-datastelle wat suidelike regterwalvis (SRW) en boggelrugwalvis (HW) vokalisering bevat. Hul prestasies word geëvalueer met behulp van maatstawwe soos die ware positiewe koers (TPR), presisie (PREC), foutkoers (ERR) en F_1 tellings. Prestasie-uitkomste wissel onderhewig aan verskillende eksperimentele toestande soos die dimensie van die kenmerkvektore, die grootte van die opleidingsdata en die spesievokalisering.

Die modelle het goeie prestasie oor verskillende evalueringsmaatstawwe getoon. Vir die PC-HMM's het die kPCA-HMM nie net beter gevaar as die PCA-HMM in terme van TPR en PREC nie, maar dit het ook 'n laer ERR getoon. Die kCPA-HMM toon egter 'n hoër berekeningskoste in vergelyking met die PCA-HMM. Net so, vir DM-HMM's, het die kDMD-HMM beter as die DMD-HMM gevaar in terme van TPR en PREC, en dit het ook laer ERR sowel as 'n laer berekeningskoste getoon.

Die vergelyking het getoon dat PC-HMM's vinniger gestabiliseer het as DM-HMM's in terme van werkverrigting. Dus, die PC-HMMs is minder kompleks as die DM-HMMs in terme van dimensie. Die DM-HMM's het egter beter as die PC-HMM's gevaar, alhoewel by hoër afmetings. Die betroubaarheid van die ontwikkelde modelle is bevestig met F_1 tellings, aangesien al die modelle F_1 tellings > 0.9 by hul onderskeie optimale afmetings behaal het.

Laastens word die resultate van die voorgestelde ED-HMMs vergelyk met die bestaande FE-tegnieke wat met HMM in die literatuur gebruik word vir die opsporing van walvisvokalisering. Die ED-HMM's presteer wel beter as die bestaande HMM-metodes. 'n Algemene waarneming is dat elke model beter prestasie toon met 'n toename in die aantal monsters wat vir opleiding ontplooi word. Daarom word groot vensterg-roottes aanbeveel vir modelopleiding. Die verskillende eksperimentele resultate het getoon dat 'n model se prestasie op 'n spesie-tot-spesie basis geëvalueer moet word. Dit is ook belangrik dat die opleidingsdata 'n subset van die datastelle vir toetsing is, of ten minste opnames van dieselfde streek gebruik. Die ED-HMM's wat in hierdie studie voorgestel word, kan verder getoets word op ander walvisvokalisering om hul robuustheid te bevestig. Boonop kan hulle ondersoek word deur navorsers wat werk aan die outomatiese opsporing van ander vokale dierspesies.

"Then which of the Blessings of your Lord will you deny?" Holy Qur'an, Chapter 55 Vs 13.

Alhamdu lillahi rabbi alAAalameen: All the praises and thanks be to Allah, the Lord of all that exists, for His bountless mercies on me.

I dedicate this work to the memory of my parents and all those whom Allah has used in making me the person I am today.

Acknowledgements

I sincerely appreciate and thank my supervisor, Prof. D.J.J. Versfeld, for his mentorship, patience, encouragement, and support throughout the period of my research. He was there for me when things got tough. Baie dankie, Prof. Versfeld!

The financial support of the National Research Foundation (NRF), South Africa, and the Nigerian government's NEEDS assessment through the University of Ilorin, Kwara State, Nigeria, is duly acknowledged. I appreciate Stellenbosch University for providing conducive environments for research. I thank my home university, the University of Ilorin, for giving me the opportunity to pursue my PhD programme.

I am grateful to all the amazing people I had the privilege of meeting in South Africa. A special thank-you to Dr. O.O. Ogundile, Dr. O.P. Babalola and family, Gamu Mamhende, AbdurRahman Suleiman, Nurayn Tiamiyu, Mpho Molapo, Dr. Sherif Isa, Hazel Alexander, Dr. Stephen David, Buyi, Idris Munir, Mukhtar, staff at SU Int'l Centre, Stellenbosch Gujjatul Islam Community, and ANSSU members.

I am deeply grateful for all the support from my family and friends, whose names are too numerous to list individually. Your encouragement means the world to me.

I extend my heartfelt gratitude to my mothers, Hajia Folashade Adegboyega-Usman and Hajia Falilat Quadri-Usman, as well as *Mummy* Mary Anwa-Lawal, and my father, Alhaji Abdulfatai Quadri. Your unwavering foresight, love, and sacrifices for me have been invaluable. Words can never adequately express my gratitude.

To my darling wife, Dr. Aisha Omowumi "A24," and our adorable children, Maryam "Slingo," Abdullah "Engr Budu," Muhammad "Haji Muha," and Abdus-Salam "Prof. Cutie," I thank you for your patience and support throughout this journey. You made it easy for me by staying strong while I was away. Your resilience and strength during my absence made this endeavour easier for me. I consider myself incredibly fortunate to have you all in my life; you hold a special place in my heart. My love for each of you is boundless and everlasting!

Contents

List of Figures

List of Tables

List of Abbreviations

AAM active acoustic monitoring

ACC accuracy

AM amplitude modulated

BW-alg Baum-Welch algorithm

CNN convolutional neural network

DCT discrete cosine transform

DMD dynamic mode decomposition

DMD-HMM DMD-based hidden Markov model

DM-HMMs dynamic modes-based hidden Markov models

DTW dynamic time warping

ED eigendecomposition-based

ED-HMMs ED-based hidden Markov models

EMD empirical mode decomposition

EMD-HMM empirical mode decomposition-based hidden Markov model

ERR error rate

EM-alg expectation-maximisation algorithm

FE feature extraction

FFT fast Fourier transform

FM frequency modulated

FPR false positive rate

GMM Gaussian mixture model

HHT Hilbert Huang transform

HMM hidden Markov model

HW humpback whales

kDMD kernel dynamic mode decomposition

kDMD-HMM kernel DMD-based hidden Markov model

kPCA kernel principal component analysis

kPCA-HMM kernel PCA-based hidden Markov model

LPC linear prediction coefficient

LPC-HMM LPC-based hidden Markov model

MATLAB Matrix Laboratory

MF matched filtering

MFCC Mel-scale frequency cepstral coefficient

MFCC-HMM MFCC-based hidden Markov model

ML machine learning

NN neural network

NRW North Atlantic right whales

PAM passive acoustic monitoring

PCA principal component analysis

PCA-HMM PCA-based hidden Markov model

PCs principal components

PC-HMMs PCs-based hidden Markov models

PDF probability density function

PREC precision

RKHS reproducing the kernel Hilbert spaces

SNR signal-to-noise ratio

SPCC spectrogram cross-correlation

SRW southern right whales

STFT short-time Fourier transform

SVD singular value decomposition

SVM support-vector machine

TPR true positive rate

Vit-alg Viterbi algorithm

WT wavelet transform

List of Notations

- β_i - Expansion coefficients
- $\ddot{c}$ - IMF from EMD operation
- $\mathbf{E}$ - Emission distribution probabilities or Gaussian emission distribution
- $\mathbf{\Sigma}$ - Matrix of singular values of $\mathbf{X}$
- ξ - Covariance matrix of Gaussian emission distribution
- $\boldsymbol{F}$ - Feature matrix
- $\bar{\mathbf{F}}(t)$ - DMD solution for all time in future
- $\mathbf{f}$ - Feature vector
- $\tilde{\mathbf{G}}$ - Projected $\mathbf{X}$ in p-low-dimensional space
- $\mathbf{h}$ - Selected signal from sampled PAM recordings, $\mathbf{s}$
- $\mathit{\Pi} = (\tau, Tr, \mathbf{E})$ - HMM parameters
- i - Lower limit, index or counter variable of a series, unless otherwise defined
- k - Upper limit of a series
- K - Kernel function
- m - Number of columns of a data matrix, $\mathbf{X}$, unless otherwise defined
- $\tilde{\boldsymbol{M}}$ - DMD modes
- n - Number of rows of a data matrix, $\mathbf{X}$, unless otherwise defined
- N - Number of states in an HMM
- Q - HMM observation sequence
- p - Number of columns selected for HMM experiment
- $\bar{\mathbf{P}}$ - K-means objective function

- $\ddot{r}$ - Residue from EMD operation
- $\mathbb{R}$ - Set of real numbers
- $\bar{\mathbf{s}}$ - Sampled PAM recordings
- $\ddot{S}(t, f)$ - SPCC function
- $\mathcal{S}$ - Individual HMM states
- $\mathcal{S}_t$ - Active state at time t
- $\ddot{s}$ - Wavelet scaling parameter
- t - Time, unless otherwise defined
- T - Length of HMM observation sequence
- Tr - Transition probability matrix of HMM, with element tr_{ij} denoting transition from state i to state j
- τ - Start state probability in an HMM
- $\mathbf{U}$ - Matrix of left singular vectors of $\mathbf{X}$
- μ - Mean of Gaussian emission distribution
- $\ddot{u}$ - Wavelet translating parameter
- $\mathbf{V}$ - Matrix of right singular vectors of $\mathbf{X}$
- ϑ - Mixture weight of Gaussian emission distribution
- $\mathcal{W}$ - Window size or numbers of samples in an observation
- $\mathbf{x}$ - Vectors in a data matrix
- X - Signal or data (could be in raw form), unless otherwise defined
- $\mathbf{X}$ - Data matrix representation
- $\ddot{X}$ - STFT of a signal
- Υ - Eigenvalues

- Ψ - Eigenvectors
- $\mathbf{Z}$ - DMD best-fit linear operator
- $\ddot{\psi}(t)$ - Wavelet function
- $*$ - Transpose of a matrix such that $\mathbf{X}^*$ is the transpose of $\mathbf{X}$
- $\Theta(\mathbf{x})$ - Transformation of x into feature space
- $\dagger$ - Moore-Penrose pseudo-inverse

Thesis Output

The findings, results, and analyses from this study have contributed in part or in full to the following peer-reviewed publications:

[1] **A. M. Usman** and D. J. J. Versfeld, "Principal components-based hidden Markov model for automatic detection of whale vocalisations," *Journal of Marine Systems Volume 242*, 2024. https://doi.org/10.1016/j.jmarsys.2023.103941.

[2] **A. M. Usman** and D. J. J. Versfeld, "Detection of baleen whale species using kernel dynamic mode decomposition-based feature extraction with a hidden Markov model," *Ecological Informatics Volume 71 (101766)*, 2022. https://doi.org/10.1016/j.ecoinf.2022.101766.

[3] **A. M. Usman**, O. O. Ogundile and D. J. J. Versfeld, "Review of automatic detection and classification techniques for cetacean vocalization," *IEEE Access Volume 8 (2020)*, 2020. https://doi.org/10.1109/ACCESS.2020.3000477.

[4] O. O. Ogundile, **A. M. Usman**, O. P. Babalola and D. J. J. Versfeld, "Dynamic mode decomposition: A feature extraction technique based hidden Markov model for detection of Mysticetes' vocalisations," *Ecological Informatics Volume 63 (101306)*, 2021. https://doi.org/10.1016/j.ecoinf.2021.101306.

[5] O. P. Babalola, **A. M. Usman**, O. O. Ogundile and D. J. J. Versfeld, "Detection of Bryde's whale short pulse calls using time domain features with hidden Markov models," *SAIEE Africa Research Journal Volume 112*, 2021. https://doi.org/10.23919/SAIEE.2021.9340533.

[6] O. O. Ogundile, **A. M. Usman** and D. J. J. Versfeld, "An empirical mode decomposition based hidden Markov model approach for detection of Bryde's whale pulse calls," *the Acoustical Society of America Volume 147 (EL125-EL131)*, 2020. https://doi.org/10.1121/10.0000717.

[7] O. O. Ogundile, **A. M. Usman**, O. P. Babalola and D. J. J. Versfeld, "A hidden Markov model with selective time domain feature extraction to detect inshore

Bryde's whale short pulse calls," *Ecological Informatics Volume 57 (101087)*, 2020. https://doi.org/10.1016/j.ecoinf.2020.101087.

Chapter 1

Introduction

Whales fall under the cetacean taxonomy of marine mammals. According to the Society for Marine Mammalogy's committee on taxonomy, there are currently 92 identified species in the cetacean taxonomy [1]. The species identification is by no means complete, as new species are still being discovered [1, 2]. The species are categorised into two suborders - the *Odontocetes* or toothed whales, and the *Mysticetes* or baleen whales [2]. The toothed whales range in size from small to medium, with the exception of the sperm whale, which reaches lengths of about 18 m [2]. The physical characteristics of toothed whales include the presence of a single blowhole, homodont teeth, an asymmetrical skull, a thin-walled "pan bone" at the posterior end of the mandible, a complex system of nasal sacs, and the melon - a fatty organ in their forehead area [2, 3]. Currently, 77 toothed whales species have been identified, which include the sperm whales, killer whales, and all species of dolphins and porpoises [1]. The baleen whales (also called toothless whales) are generally big, with the females growing larger than the males. The smallest baleen species is the pygmy right whale, measuring less than 7 m in length. The biggest baleen whale species is the blue whale (the largest ever known animal), measuring up to 33 m in length and about 145 000 kg in weight [2, 3]. Baleen whales have a double blowhole, a single symmetrical skull, and lack a bony mandibular symphysis [3]. In place of teeth, the upper jaw of baleen whales is hung with hardened plates of keratin with fringes on the inside,

called "baleen plates". The fringes are used to filter food from water [2]. There are currently 15 identified baleen whale species, including the bowhead whales, Southern right whales (SRW), North Atlantic right whales (NRW), humpback whales, Bryde's whales and blue whales [1].

Whales are of concern to the general public, animal conservationists, and ecosystem managers because of their economic and other significance. For instance, the whale-watching industry, which involves over 87 nations and territories, has witnessed high demands annually [4]. The whale-watching ventures give tourists the opportunity to closely observe, touch, swim with, and feed the animals. This tourism sector generates over US$2 billion annually [5], thus creating employment opportunities for thousands of people while generating revenue for governments. South Africa is one of the leading destinations for whale-watching in the world [6]. Another significance of whales is that they contribute to the maintenance of healthy marine ecosystems. In addition, some whales serve as sentry species for the state of marine ecosystems [7]. Whale species are also deployed for security purposes by the US Navy [8].

However, increasing human anthropogenic activities have considerably altered the soundscape in oceans. This has led to constant threats to ocean mammals due to their reliance on sound for navigation, communication, avoidance of predators, recognition of prey, and proper functioning within their ecological space. Whales are among the marine mammals that face these threats. The threats stem from human activities such as shipping, marine exploration, geographic seismic surveys, commercial whaling, naval sonar actions, and climate change effects [9–11]. The negative effects of these activities include (a) physical injury; (b) physiological dysfunction: permanent or temporary loss of hearing sensitivity; (c) behavioural modification: decrease in exploration efficiency or inefficient use of the environment; separation of mother-calf pairs; (d) masking: difficulty in recognising crucial sounds as a result of increase in background noise; (e) avoidance and displacement from critical feeding and breeding grounds; (f) decrease in reproduction rate [11–13].

Consequently, whales have continued to gain the attention of researchers, who have

been proposing various solutions to mitigate the threats they face within their ecological space. The solutions are based on ecological informatics studies of the various whale species. They include reliable estimations of whale population density [14], measurements of range and seasonal occurrence, and determinations of the species population structures [15]. There have been reported increases/recoveries in the population of some whale species in recent times. These achievements can be attributed to management interventions, thus highlighting conservation successes from the research efforts of the scientific community [16].

1.1 Motivation and Research Question

The accurate detection and classification of different whale species plays a crucial role in supporting marine ecologists' conservation efforts, as it provides a better understanding of whale ecology and contributes to effective conservation strategies. Whale datasets are gathered through the passive acoustic monitoring (PAM) system. Several engineering methods with mixed outcomes have been presented over the years for the analysis of PAM datasets for the accurate detection and classification of whale species. So far, no single method is capable of detecting and classifying all species of whales [17]. Nonetheless, detection methods such as the hidden Markov model (HMM) have shown robustness in detecting various types of whale species [18–23], with possibilities for further improvement. HMM is one of the methods used for the detection of bioacoustic signals. The model is flexible and allows the categorisation of sounds from a series of specified observations [20]. The performance of HMM in terms of sensitivity, precision, and error rate will depend on the quality of the feature vectors fed into the model [19, 20, 24, 25]. This indicates that the quality of feature vectors influences HMM performance.

Mel-scale frequency cepstral coefficient (MFCC) [26] and linear prediction coefficient (LPC) [27] are the commonly used feature extraction (FE) techniques for HMM in the literature. However, recent research has found FE techniques that outperform existing MFCC and LPC techniques [22, 25]. One of the drawbacks of MFCC and

LPC-based FE techniques is their sensitivity to noise due to their dependency on spectral form [28, 29]. This sensitivity can result in a reduced ability to handle noisy data effectively. Also, MFCC efficiency is influenced by the number of filters [29]. The above-mentioned limitations can lead to increased computational complexity and reduced overall performance outputs for the HMM.

Therefore, the goal of this research is to design new feature extraction (FE) techniques that can be used with the HMM for the detection of whale vocalisations. The proposed FE techniques are aimed to improve the performance of the HMM. The proposed FE techniques are eigendecomposition-based (ED) algorithms. These algorithms are effective for reducing signals to their first few core characteristics. Often, a few core characteristics are sufficient to capture the intrinsic features of the signal. Therefore, the ED algorithms project signals onto a smaller subspace from which the statistical properties of the signals can be efficiently modelled. Consequently, two ED algorithms, principal component analysis (PCA) and dynamic mode decomposition (DMD), will be deployed as FE techniques in this research. Furthermore, kernel methods are explored with each of the algorithms (PCA and DMD) to enhance the feature extraction process. In the study, the reliability and adaptability of PCA and DMD algorithms (including their kernel versions) to extract features from PAM datasets and transform the obtained features into feature vectors appropriate for HMM will be investigated.

In this research, we adopt a novel approach by introducing PCA and DMD as FE techniques specifically tailored for PAM recordings containing whale vocalisations. The PCA is the foundation of dimensionality reduction techniques and remains one of the techniques suitable for extracting inherent features from datasets. PCA essentially reduces high-dimensional data to a lower number of dimensions while retaining the important information that explains the original data [30]. Although PCA has been around for a while and has been used for extracting features in different fields such as image processing, disease modelling, and denoising, its use is not one-size-fits-all [31]. In other words, the characteristics of the data, the formats of the data, and the goals to be achieved, among others, will determine how the principal components

will be utilised. In the literature, there are many modifications and adaptations to the use of PCA for the analysis of different data types, subject to the specifics of the goals to be achieved [30–32]. This research uniquely introduces PCA as an FE technique for PAM recordings containing whale vocalisations. The novel approach of this study would be to deploy the principal components (PCs) to be derived from the PCA and use them to project the original datasets onto a low-dimensional space while preserving their intrinsic features.

Furthermore, the DMD is a data-driven algorithm that can generate reliable eigendecomposition of non-stationary data into spatiotemporal coherent patterns that will distinguish prominent structures of the data [33]. The DMD does not rely on the mathematical equations governing a system; rather, it relies on the raw data to identify structures that describe the behaviour of the system. This makes DMD attractive for the analysis of different data types. DMD extracts information that structurally describes the inherent physical properties that dominate the entire data. In the literature, the DMD algorithm is primarily used for diagnostics, state estimate and future-state prediction, and control. We introduce it as an innovative tool for FE. The data-driven framework of the DMD, which eliminates the need to understand or have prior knowledge of the underlying equations governing the data being modelled, is one of the motivations for the exploration of the algorithm as a feature extraction technique in this research.

Additionally, kernel methods are incorporated into both PCA and DMD algorithms to account for the non-linear characteristics that may exist in the whale datasets. Whale vocalisations often exhibit non-linear characteristics such as frequency jumps, subharmonics, and biphonation [34]. Kernel methods enable algorithms to handle the non-linear relationship between variables by transforming the data into a higher-dimensional feature space in order to effectively capture the non-linear patterns in the data.

The integration of these ED algorithms, along with their respective kernel methods, into HMM in this study aims to enhance the performance while reducing the computational load for the detection of whale vocalisations. PCA and DMD, along with

kernel methods, will be investigated for their reliability and adaptability in extracting features from PAM datasets and transforming these features into suitable feature vectors for HMM analysis. Accordingly, in this thesis, the broad research question is stated as:

How can eigendecomposition-based (ED) algorithms, along with kernel methods, be utilised as feature extraction (FE) techniques to improve the performance of hidden Markov model (HMM) for the detection of whale vocalisations in passive acoustics monitoring (PAM) datasets?

1.2 Research Hypotheses

The PCA reduces high-dimensional data to a lower number of dimensions while retaining the important information that explains the original data. These dimensions are hierarchically ordered from the most to the least important statistical variation in the datasets. The variables within these dimensions are referred to as principal components (PCs). Similarly, the DMD breaks data into spatiotemporal coherent patterns that will distinguish the prominent features in the datasets. The spatiotemporal coherent patterns are represented as modes. The PCs from PCA and the modes from DMD are key parameters under which this study will be investigated. Also, kernel methods are introduced into each of the algorithms for enhanced feature extraction. Therefore, the following four key hypotheses are made in order to start investigations and answer the research question of this study:

Hypothesis 1 (H1). The PCs computed using PCA can be transformed into suitable feature vectors to be used with HMM for the detection of whale vocalisations.

Hypothesis 2 (H2). The modes computed using DMD can be transformed into suitable feature vectors to be used with HMM for the detection of whale vocalisations.

Hypothesis 3 (H3). The use of kernel methods with algorithms in Hypotheses 1 and 2 will produce enhanced feature vectors for HMM for the detection of whale vocalisations.

Hypothesis 4 (H4). The emerging ED-based hidden Markov models (ED-HMMs) from Hypotheses 1–3 will exhibit better performance when compared with existing FE techniques used the HMM in the literature for the detection of whale vocalisations.

1.3 Research Aim and Objectives

The aim of this study is to develop eigendecomposition-based feature extraction algorithms to enhance the performance of hidden Markov model for the detection of whale vocalisations. To achieve this aim, the following objectives are highlighted:

1. *To investigate the different FE techniques as well as the detection and classification methods in the literature for whale vocalisations.*

2. *To investigate and analyse ED algorithms as suitable FE techniques to be used with HMM for the detection of whale vocalisations.*

3. *To compute PCs using PCA and transform the computed PCs to feature vectors to be used with HMM for the detection of whale vocalisations.*

4. *To compute modes using DMD and transform the computed modes as feature vectors to be used with HMM for the detection of whale vocalisations.*

5. *To introduce kernel methods to the algorithms in Items 3 and 4 to obtain enhanced feature vectors to be used with HMM for the detection of whale vocalisations.*

6. *To model and simulate the proposed eigendecomposition-based hidden Markov models (ED-HMMs) in (3)-(5) using Matrix Laboratory (MATLAB) software.*

7. *To perform a comparison study of the results of the proposed ED-HMMs with existing FE techniques HMMs in the literature for the detection of whale vocalisations.*

1.4 Contributions

The contributions of this research are summarised as follows:

- Despite the considerable research that has been dedicated to the design of different methods for the detection and classification of whale vocalisations, no published review exists that gives a comprehensive overview and aggregation of the existing methods. Therefore, in this study, we commenced with a comprehensive review of the different detection and classification methods, which culminated in a review journal article.

- Introduction of ED algorithms based on PCA and DMD as feature extraction techniques used with HMM for the detection of whale vocalisations. Hitherto, to the best of the author's knowledge, these algorithms have not been used exclusively for feature extraction of whale vocalisations.

- Although PCA has been deployed for a variety of purposes in different fields of application, it has not been used for the exclusive derivation of features for whale vocalisations, particularly in conjunction with HMM. Besides, the design and formulation of our PCA framework were uniquely done to derive feature vectors appropriate for HMM for the detection of whale vocalisations.

- The DMD algorithm is popularly used for the analysis of complex flows and spatiotemporal patterns across various fields such as fluid mechanics, disease modelling, and control systems. This research pioneers a novel utilisation of DMD as an FE technique. Unlike its conventional applications, where DMD is primarily used for the study and view of the dynamic behaviour of a system, this research explores its potential as a technique for extracting features that can be easily adapted with machine learning (ML) models for detection or classification purposes, thus extending its utility beyond its traditional domains to other areas of research.

- Whale vocalisations do exhibit non-linear characteristics, which PCA and DMD may not properly capture. Therefore, the potentials of kernel methods for solving non-linear problems are explored with each of the techniques (PCA and DMD) to account for the non-linear subspace that may exist in whale vocalisations. The kernel method is uniquely incorporated into each of the techniques to further deepen the feature extraction process. While we acknowledge the existence and utility of kernel PCA for other uses, this research introduces the kernel method into the DMD architecture.

- This research introduces an innovative framework for matrix formulation that adapts to the window size, regardless of the sampling rate, thus providing flexibility in the analysis of audio signals.

- The developed FE techniques in this study can be used with different ML models for the detection and classification of a variety of data types.

- The models developed in this thesis can be adopted by researchers working on automatic detection of other vocalising animal species.

- The findings, results, and analyses from this study have contributed in part or in full to the following peer-reviewed publications:

  [1] **A. M. Usman** and D. J. J. Versfeld, "Principal components-based hidden Markov model for automatic detection of whale vocalisations," *Journal of Marine Systems Volume 242 (103941)*, 2024. https://doi.org/10.1016/j.jmarsys.2023.103941.

  • **A. M. Usman**: Idea conceptualisation, design and formulation of frameworks for transforming PCs from PCA and kPCA for feature extraction, methodology, and experimental formulations and execution, evaluation of results, coding, paper draft, paper writing and revision.

  • D. J. J. Versfeld: Idea conceptualisation, supervision, review and editing.

  [2] **A. M. Usman** and D. J. J. Versfeld, "Detection of baleen whale species using kernel dynamic mode decomposition-based feature extraction with a hidden Markov model," *Ecological Informatics Volume 71 (101766)*, 2022.

https://doi.org/10.1016/j.ecoinf.2022.101766.

• **A. M. Usman**: Idea conceptualisation, design and formulation of kernel into the DMD architecture for feature extraction, methodology, and experimental formulations and execution, evaluation of results, coding, paper draft, paper writing and revision.

• D. J. J. Versfeld: Idea conceptualisation, supervision, review and editing.

[3] **A. M. Usman**, O. O. Ogundile and D. J. J. Versfeld, "Review of automatic detection and classification techniques for cetacean vocalization," *IEEE Access Volume 8 (2020) 105181-105206*, 2020. https://doi.org/10.1109/ACCESS.2020.3000477.

• **A. M. Usman**: Idea conceptualisation, literature review, paper draft, paper writing and revision.

• O. O. Ogundile: Idea conceptualisation, review and editing.

• D. J. J. Versfeld: Idea conceptualisation, supervision, review and editing.

[4] O. O. Ogundile, **A. M. Usman**, O. P. Babalola and D. J. J. Versfeld, "Dynamic mode decomposition: A feature extraction technique based hidden Markov model for detection of Mysticetes' vocalisations," *Ecological Informatics Volume 63*, 2021. https://doi.org/10.1016/j.ecoinf.2021.101306.

• O. O. Ogundile: Idea conceptualisation, coding, evaluation of results, paper writing and revision.

• **A. M. Usman**: Idea conceptualisation, literature review, design and formulation of the DMD modes for feature extraction, methodology, experimental formulations, and paper draft and editing.

• O. P. Babalola: Experimental formulations and execution, paper draft and editing.

• D. J. J. Versfeld: Idea conceptualisation, supervision, review and editing.

[5] O. P. Babalola, **A. M. Usman**, O. O. Ogundile and D. J. J. Versfeld, "Detection of Bryde's whale short pulse calls using time domain features with hidden Markov models," *SAIEE Africa Research Journal Volume 112 (2021) 15-23*, 2021. https://doi.org/10.23919/SAIEE.2021.9340533.

• O. P. Babalola: Idea conceptualisation, design of methodology for feature extraction, coding, execution, evaluation of results, paper writing and revision.

• **A. M. Usman**: Idea conceptualisation, literature review, design of methodology for feature extraction, methodology, experimental formulations, and paper draft and editing.

• O. O. Ogundile: Idea conceptualisation, experimental formulations and execution, and paper draft and editing.

• D. J. J. Versfeld: Idea conceptualisation, supervision and guidance at the design and formulation phases, review and editing.

[6] O. O. Ogundile, **A. M. Usman** and D. J. J. Versfeld, "An empirical mode decomposition based hidden Markov model approach for detection of Bryde's whale pulse calls," *the Acoustical Society of America Volume 147 (EL125-EL131)*, 2020. https://doi.org/10.1121/10.0000717.

• O. O. Ogundile: Idea conceptualisation, design of methodology for feature extraction, coding, evaluation of results, paper writing and revision.

• **A. M. Usman**: Idea conceptualisation, literature review, experimental formulations, and paper draft and editing.

• D. J. J. Versfeld: Idea conceptualisation, supervision, review and editing.

[7] O. O. Ogundile, **A. M. Usman**, O. P. Babalola and D. J. J. Versfeld, "A hidden Markov model with selective time domain feature extraction to detect inshore Bryde's whale short pulse calls," *Ecological Informatics Volume 57 (101087)*, 2020. https://doi.org/10.1016/j.ecoinf.2020.101087.

• O. O. Ogundile: Idea conceptualisation, design of methodology for feature extraction, coding, evaluation of results, paper writing and revision.

• **A. M. Usman**: Idea conceptualisation, literature review, methodology, experimental formulations, and paper draft and editing.

• O. P. Babalola: Experimental formulations and execution, paper draft and editing.

• D. J. J. Versfeld: Idea conceptualisation, supervision and guidance at the design and formulation phases, review and editing.

1.5 Thesis Overview

Chapter 1 gives a general introduction to the theme of this thesis. The chapter starts with an overview of whale species, their taxonomy, their importance, and the threats they face from human anthropogenic activities. The motivation and research question (Section 1.1), the formulated research hypotheses (Section 1.2), the research aim and objectives to achieving the aim (Section 1.3), as well as the contributions (Section 1.4) of the research are presented, while Section 1.5 concludes this chapter.

Chapter 2 gives detailed background and relevant literature for this study. The chapter starts by giving insight into whale vocalisations (Section 2.2) and how the vocalisations aid the monitoring and observation of whales within their ecological space (Section 2.3). An overview of the detection and classification process is presented in Section 2.4. The data recording and preprocessing stages of the detection and classification process are briefly discussed in Section 2.5. The existing FE techniques used with the different detection and classification methods are reviewed in Section 2.6. The existing methods for the detection and classification of whale vocalisations are equally reviewed in Section 2.7. Furthermore, the performance metrics used to analyse and compare detection and classification techniques are discussed in Section 2.8. The chapter concludes with summaries of findings from the reviewed literature and highlights of some of the reviewed studies (Section 2.9).

The ED-FE techniques are presented in Chapter 3 with a general introduction in Section 3.1. The process of discretising raw PAM datasets and transforming them into matrices for further analysis is detailed in Section 3.2, while Section 3.3 provides an overview of the ED algorithms. Next, in Section 3.4, SVD is introduced as a central operation in most ED algorithms. Section 3.5 presents the process of computing PCs, while Section 3.6 details the proposed PCs-based feature vectors for HMM. The PC-based feature vectors process is concluded with an illustrative example in Section 3.7. The DMD is introduced in Section 3.8 as well as the steps for computing the modes. The proposed DMD-based feature vectors for HMM is explained in Section 3.9. The

conclusion to the chapter is presented in Section 3.10, where the need to further deepen the FE process is motivated.

Chapter 4 introduces the application of kernel methods to the FE process. A general overview of the kernel methods is given in Section 4.2. The proposed enhanced PCs-based feature vectors for HMM is described in Section 4.3, while Section 4.4 explains the proposed enhanced DMD-based feature vectors for HMM. The conclusion to the enhanced FE process is presented in Section 4.5.

Chapter 5 describes the data and the experimental set-up for training and testing the developed models. Section 5.2 gives a detailed description of the datasets used to test the developed models. The summary of HMM as deployed in this study is explained in Section 5.3. The experimental set-up of the experiments is elaborated in Section 5.4, and the chapter is concluded in Section 5.5.

In Chapter 6, the results from the series of experiments conducted are discussed and analysed. The performance of the PCA-based hidden Markov model (PCA-HMM) and the kernel PCA-based hidden Markov model (kPCA-HMM) are evaluated in Section 6.2, as well as those of the DMD-HMM and the kDMD-HMM in Section 6.3. Furthermore, a comparative study of the performance of the developed ED-HMMs detection models is presented in Section 6.4, while the results from the proposed ED-HMMs are compared with the existing FE techniques used with HMM in the literature for the detection of whale species in Section 6.5. The worst-case time complexity analysis of the FE techniques is investigated in Section 6.6, and conclusions drawn from the analysis and discussion of the results are presented in Section 6.7.

Finally, Chapter 7 draws a conclusion to the study with a summary of the research presented in Section 7.1. The limitations of research are presented in Section 7.2, while Section 7.3 concludes this thesis by giving some recommendations for future research directions.

Chapter 2

Background and Literature

2.1 Introduction

This chapter focuses on the review of the existing literature on techniques for the detection and classification of cetacean species through vocalisations. The reviewed methods in this chapter are explained to an acceptable extent of understanding. However, more attention is given to the hidden Markov model (HMM), since it is the detection method this study seeks to improve. The chapter starts by giving insight into whale vocalisations and whale monitoring systems. The chapter goes on to review the feature extraction (FE) techniques as well as the detection and classification methods in the literature. Common performance metrics for detection and classification techniques are also defined. The chapter concludes with summaries of findings from the reviewed literature.

This review has been published in Usman *et al.* [17]. The goal of the review paper is to give a quick overview of the area of automatic detection and classification of cetacean vocalisations. Therefore, this chapter is structured in line with the published review paper. The details of the review paper are as follows:

A. M. Usman, O. O. Ogundile and D. J. J. Versfeld, "Review of automatic detection and classification techniques for cetacean vocalization," *IEEE Access Volume 8 (2020) 105181-105206*, 2020. https://doi.org/10.1109/ACCESS.2020.3000477.

2.2 Whale Vocalisations

Whales vocalise a variety of distinctive sounds for communication, echolocation, and other social functions. The vocalisations are non-stationary and have varying physical properties [13, 15]. The vocalisations cover a large frequency range and exhibit different characteristics, such as variation within and between species, species-specific, temporal and geographical differences [13, 34, 35]. Thus, making the vocalisations complex to analyse [36].

The whale vocalisations are generally categorised as clicks, whistles, pulsed calls, and songs [37, 38]. The click sounds are used for echolocation and foraging [37, 39]. The clicks are impulsive short pulses of substantial strength, lasting for microseconds and extremely directional. Whales produce multiple clicks at a time, with diverse cue rates ranging from 0.5 to 2 clicks per second. The whistle and pulsed calls are used for social undertakings. Whistles are relatively tonal, narrow band, and frequency modulated (FM) signals with a frequency range from 2 to 30 kHz depending on species [37, 40]. The pulsed calls are amplitude modulated (AM) signals and can sound like growls. Songs are the most complicated of the signals produced by the baleen suborder [13, 15]. A song is a series of individual calls organised into a hierarchical structure that can last for several minutes in some instances and even hours [13, 37]. Just a few whale species generate songs; examples include humpback whales, bowhead whales, blue whales, and fin whales [13, 15, 37]. The known frequency range of whale vocalisations is from 7 Hz to 100 kHz [37, 41]. Figure 2.1 shows a typical example of the waveform and spectrogram views of humpback whales (HW) vocalisations.

Whale vocalisations exhibit non-linear characteristics such as frequency jumps, subharmonics, and biphonation [34]. Frequency jumps emerge when the frequency of a

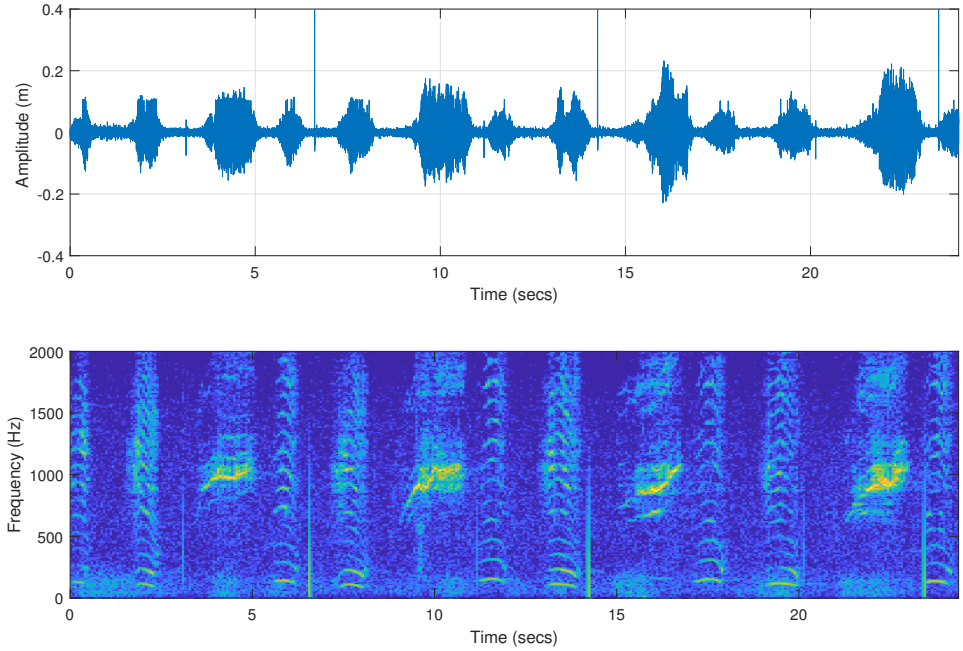


FIGURE 2.1: The waveform and spectrogram views of humpback whale vocalisations.

signal changes almost instantaneously. Subharmonics are additional spectral bands that exist below the fundamental frequency. The biphonation characteristic occurs when a whale species produces two different harmonically unrelated sounds at the same time. Whale vocalisations are usually represented in terms of their waveform (amplitude on the y-axis and time on the x-axis) and spectrogram (frequency on the y-axis and time on the x-axis).

2.3 Whale Monitoring and Observation

Traditionally, whales are visually monitored and observed to understand their ecology. But whale populations are often underestimated because they spend their entire lives in water. As a result, visual monitoring becomes insufficient for an accurate estimation of the whale population [15, 42]. Also, visual monitoring is hindered by environmental conditions such as remote topography, time of day, the short time whales spend on the surface, and whales high mobility rate [15, 35, 41].

Alternatively, acoustic monitoring is used for whale monitoring and observation. This approach is more effective for monitoring and observing whales from afar since whales use sounds as their primary means of communication and sensing [15, 18, 43]. Moreover, acoustic monitoring is not affected by the environmental conditions under which visual monitoring cannot be performed.

Acoustic monitoring can either be active or passive. In active acoustic monitoring (AAM), sound energy is transmitted and the returning signals are analysed. The AAM system is not commonly used for monitoring and observation since the method can upset the behaviour of whales due to its invasive mode of operation [41]. On the other hand, the passive acoustic monitoring (PAM) system uses underwater microphones (hydrophones) to capture whale vocalisations from the surrounding environment in a non-invasive manner [41]. PAM is widely used for marine mammal monitoring and observation [41, 44]. This system of monitoring has been proven to be an effective way to observe whales while remaining unobtrusive. Thus, PAM has become an important method for gathering data on whale species [17].

The accurate detection and classification of whale species are important to helping marine ecologists propose solutions to mitigate the threats faced by the species. Thereby, leading to a better understanding of whale ecology. The detection and classification of whale vocalisations can be performed manually or automatically. Manual detection and classification involve observing the recorded whale vocalisations on a spectrogram by an expert marine ecologist. Similarly, the vocalisations can be listened to by an expert marine ecologist to detect and classify them accordingly [20]. However, large volumes of data are often gathered during the PAM process, which can run for weeks, months, or even years. Thus, manual analysis of PAM data is strenuous, time-consuming, and prone to human error. Consequently, different automatic methods have been proposed over the years for the accurate detection and classification of whale species through their vocalisations [17].

2.4 Overview of the Detection and Classification Process

The application of robust automatic detection and classification methods to cetacean vocalisations can give understanding into their repertoire variation, individual vocal changeability, social setting relationships, and some other crucial cetacean behavioural questions [45]. The different detection and classification methods involve the use of traditional signal processing methods. The traditional methods include spectrogram cross-correlation (SPCC) [46], matched filtering (MF) [47], and dynamic time warping (DTW) [48]. In recent years, ML algorithms have become popular for the detection and classification of cetacean species. The flexibility of building models from data using ML algorithms provides more opportunities to find efficient ways to detect and classify cetacean species. The ML methods include support-vector machine (SVM) [49], Gaussian mixture model (GMM) [50], HMM [19, 20] and different classes of neural network (NN) [18], including deep learning [51]. The basic block diagram of the detection and classification process is shown in Figure 2.2.

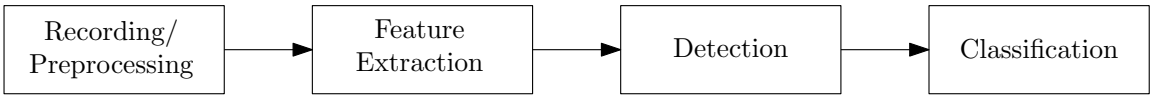


FIGURE 2.2: Block diagram of cetacean detection and classification stages.

Cetacean vocalisations are recorded through the use of hydrophones and preprocessed to clean the recorded data. Features of the species of interest are extracted using any reliable FE technique. Detection is the process of identifying the presence of the targeted whale signal in the dataset, while classification involves assigning the detected signals to a predefined category (species-specific) [52, 53]. The variability within and between the whale species vocalisations makes the classification stage very crucial. Classification helps in categorising the individual species. Therefore, having different techniques that automatically detect and categorise different species of cetaceans will aid comparisons both between and within species. Each of the stages in Figure 2.2 are explicitly discussed in the subsequent sections.

2.5 Data Recording and Preprocessing

During PAM, hydrophones are used to record cetacean vocalisations. Preprocessing is done to clean the recorded datasets. The preprocessing stage focuses on the removal of background noise that may be present in the recording through the application of appropriate filtering techniques and producing a time-frequency-amplitude representation of the recorded signal to form a dataset [18, 44]. This process includes the denoising done to clean and enhance the quality of the datasets [38]. This is followed by annotations of parts of the recording where the vocalisations of interest are located by a human expert who is assumed to know the property of the vocalisation of interest. The beginning and end points of a particular vocalisation class are identified in the recording file. A single identification can be regarded as a label. Multiple sets of such labels can be identified in a long recording. The different recognised labels are then used as samples for the training of the technique to be deployed for detection and classification. This is usually done via visual inspection of a spectrogram. However, a number of software are available to carry out the annotation seamlessly. Examples of existing software for annotation are *Sonic Visualiser*, and *Audacity* ™ [36].

2.6 Feature Extraction Techniques

Underwater signals contain different types of sounds. The sources of the sounds include environmental (rain, cracking of ice, estuaries, and so on), human anthropogenic activities (shipping, offshore exploration, geophysical seismic survey, and so on), and biological sounds (marine mammal sounds) [54]. Therefore, datasets gathered during the PAM process come with other sounds besides the sound of interest. Many of these sounds are not relevant to the detection and classification of interest, so they tend to reduce the performance of detectors and classifiers if used directly for

analysis. Hence, there is a need to extract useful information from the raw recordings. Feature extraction (FE) is the process of extracting relevant information from the data so as to enhance the performance of the detector and classifier.

The features are numerical representations of an aspect of the raw data that can be related to the sounds of interest. Feature engineering, an act of extracting features from raw data and transforming them into formats that are suitable for ML models, is a fundamental phase of the ML process. Extracting the right features will not only ease the modelling but will also lead to good model performance. The technique deployed in extracting features from datasets will depend on the structure or type of data and its adaptability to the ML algorithm in which it will be used. Several FE techniques exist, fitting into different detection and classification methods. The FE techniques discussed in this review have enjoyed wide application in different fields of study. However, the explanation of each in this study is tailored towards its application to cetacean signal analysis, as found in the literature.

2.6.1 Short-Time Fourier Transform

The short-time Fourier transform (STFT) method, also known as the windowed Fourier transform, is used for analysing non-stationary signals. It strives to solve the problem of loss of time information in the Fourier transform (FT) by introducing a sliding window $\bar{v}(t)$ passing through the whole signal $X(t)$. It presents the time-localised details of the non-stationary signal $X(t)$ by disclosing the changes in the frequency content as time progresses. It has been used as a FE tool on Phonocardiogram (PCG) signals [55, 56], vibration signals measured from rolling bearings and other machine components [57]. The STFT is obtained by multiplying the time signal $X(t)$ by a suitable sliding time window function $\bar{v}(t - \ddot{\tau})$, which is constructed to extract a part of the signal and thereafter obtain the FT. The location of the sliding window adds a time dimension and obtains a time-varying frequency analysis. The STFT is mathematically defined as [56, 57]:

$$\ddot{X}(\ddot{\tau},\omega) = \int_{-\infty}^{\infty} X(t)\bar{v}(t-\ddot{\tau})e^{-j\omega t}dt, \tag{2.1}$$

where ω is the frequency and $\ddot{\tau}$ is the delay of the window function. The STFT employs a sliding window to obtain a spectrogram which provides information on both the time and frequency of a signal. This information is, however, of limited resolution due to the fixed size of the sliding window.

The STFT is applied to underwater continuous-wave (CW)-like signals to extract feature vectors for GMM [58]. In [18], the STFT was used to create whistle contours in a denoised spectrogram of whistles of long-finned pilot whales and killer whales. Raw sound data that has been denoised was sequentially sliced into sound frames. The STFT coefficients for every single sound frame were computed to create visible contours of the whistles in the spectrogram. The sound frames containing the whistles were manually marked and labelled in preparation for their usage for training and testing of deep convolutional neural networks (CNNs) meant to detect and classify the whistles accordingly.

2.6.2 Wavelet Transform

The wavelet transform (WT) technique is good for describing features of non-stationary signals. The technique is used in many applications such as image processing, signal processing, communication systems, time-frequency analysis, and pattern recognition [59–63]. The WT decomposes a signal into wavelets of several scales in the time-domain with changing window sizes, with each scale representing a particular feature of the signal under review. This technique is centred on small wavelets with limited duration [64] developed as an alternative to solving time-frequency resolution problems associated with the STFT [65]. In contrast to the STFT, which uses a single window analysis, the WT uses long windows at low frequency and short windows at high frequency. It has three advantages: (1) it is more efficient for short-lived FE as related to cetacean signals; (2) it provides uniform resolution for all the scales; and (3) it extracts signals throughout the spectrum without the need for a dominant

frequency band. The wavelet basis function $\ddot{\psi}_{\ddot{s},\ddot{u}}(t)$ is defined as:

$$\ddot{\psi}_{\ddot{s},\ddot{u}}(t) = \frac{1}{\sqrt{|\ddot{s}|}}\ddot{\psi}\left(\frac{t-\ddot{u}}{\ddot{s}}\right), \quad \ddot{s} > 0,\ \ddot{u} \in \mathbb{R}, \tag{2.2}$$

where $\ddot{\psi}(t)$ is the mother wavelet function, $\ddot{s}$ is the scaling parameter which allows $\ddot{\psi}_{\ddot{s},\ddot{u}}(t)$ to expand or contract, and $\ddot{u}$ is the translating parameter (translation in time allowing time shifting of $\ddot{\psi}_{\ddot{s},\ddot{u}}(t)$). By convention, the wavelet function is configured to attain a balance between the time domain (limited distance) and the frequency domain (limited bandwidth). As the mother wavelet dilates and time-shifts (translate), a small scaling parameter $\ddot{s}$ leads to a high frequency wavelet function $\ddot{\psi}_{\ddot{s},\ddot{u}}(t)$ which gives good time resolution with poor frequency output, while a large scaling parameter $\ddot{s}$ leads to a low frequency wavelet function $\ddot{\psi}_{\ddot{s},\ddot{u}}(t)$ which gives poor time resolution with good frequency output [65, 66]. There are two kinds of WT: the continuous wavelet transform (CWT) and the discrete wavelet transform (DWT). Both can be applied to the cetacean signal FE process.

The CWT has been used as FE tools for sperm whale clicks [59, 67, 68]. The sounds emitted by sperm whales are distinctive, short-lived, and have a broadband spectrum [68]. A new FE technique based on the CWT approach was used in [38] to decompose identified (picked) clicks from denoised sounds of sperm whales and long-finned pilot whales, and a wavelet coefficient matrix was obtained from every single picked click. A feature extraction procedure built on the concept of the wavelet coefficient matrix was proposed, centred on the energy distribution and duration variance between the two whale clicks. The feature vector was derived from the scale (frequency) features and time feature achieved from each picked click. It gave improved time resolution and frequency resolution when compared with STFT and other time frequency transform methods.

2.6.3 Hilbert Huang Transform

The Hilbert Huang transform (HHT) is an alternative method for characterising bioacoustic signals. It enjoys a wide area of applications in bioacoustic signal characterisation, fault diagnosis in nuclear reactors, biomedical diagnosis, electrical machine condition monitoring, seismic studies, and financial application, among others [69]. It gives an improved result compared to conventional time-frequency analysis methods such as STFT and WT [70]. The method is entirely empirical, and it is implemented in two phases.

The first phase is the decomposition of the signals into some monocomponent signals called intrinsic mode function (IMF) using the empirical mode decomposition (EMD) algorithm. Each IMF represents a definite frequency range, and it indicates the time evolution of the components included within that band. The EMD operates in the time domain; it is adaptive and highly effective. The second phase is the Hilbert spectral analysis, which is the application of the Hilbert transform and the time-frequency representation associated with each IMF is performed. The time-frequency representation of the IMFs is important for the comprehension of the inherent structure of the analysed dataset [69–71].

The signal $X(t)$ given can be decomposed as:

$$X(t) = \sum_{i=1}^{k} \ddot{c}_i(t) + \ddot{r}(t), \tag{2.3}$$

where $\ddot{c}_i$ is the i-th IMF of the signal, and $\ddot{r}$ is the residue, which represents portions of the signal not decomposed by EMD. The Hilbert transform is used to derive the time-frequency representation from the modes, as proposed in [71]. The final presentation of the result is an energy-frequency-time distribution of the data.

HHT has been used for extracting features in a number of cetacean detection and classification work. HHT was used in the analysis of sperm whale clicks and killer whale clicks in [72] and [36] respectively. The original HHT technique is, however,

faced with three limitations. One is the generation of undesirable IMFs in the low-frequency region, which may lead to a misinterpretation of the result. Two, the first obtained IMF may cover too wide a frequency range such that the monocomponent attribute cannot be achieved; this limitation, however, is subject to the analysed signal. Three, the signals with low energy components cannot be separated via the EMD operation [73]. An improved HHT technique proposed in [73] established criteria for selection of IMF through the use of wavelet packet transform (WPT) as a preprocessor for decomposition of a signal into a set of narrow-band signals. Thus, frequency components with low energy are easily identified in different narrow bands. The EMD operation is then performed on these narrow-band signals with each derived IMF truly becoming monocomponent. Application of the WPT prior to the EMD operation would have avoided the other two identified limitations. This improved HHT technique was shown to detect underwater acoustic signals more effectively in [74].

2.6.4 Empirical Mode Decomposition

In some recent work [22, 75], the empirical mode decomposition (EMD) technique was used to extract features without the HHT applied. In [75], the IMFs generated from the EMD were used to obtain feature vectors. The sound sources detected were uniquely labelled and verified before being manually grouped into different categories. The labels were used to classify the detected sound sources. This new approach is a modern way to carry out detection and classification in the time domain depending entirely on EMD-type processing, eliminating the necessity to apply the Hilbert transform and manual labelling of pre-processed data by an expert. They claimed their approach can be applied to a number of transient sound sources (humpback whale songs, killer whale whistle, and beluga whale whistles). Also, in [22], the generated IMFs from the EMD process were used to form feature vectors which were fed into HMM to detect Bryde's whale pulsed calls.

2.6.5 Linear Prediction Coefficients

The linear prediction coefficient (LPC) is a signal analysis method used in speech coding, speech synthesis, speech recognition, speaker recognition and verification, and speech storage [76]. The LPCs expeditiously represent speech signals as short-time spectral information [27, 77]. The main concept of this method is that it predicts the value of the current sample signal $X(n)$ by a linear combination of past samples and then approximates the difference between the actual value and the predicted value as shown:

$$\hat{X}(n) = \sum_{i=1}^{k} a_i X(n-i), \tag{2.4}$$

$$\ddot{e}(n) = X(n) - \hat{X}(n), \tag{2.5}$$

where $X(n)$ is the current sample signal, $\hat{X}(n)$ is the predicted value, a_i are k-th order linear predictor coefficients, and $\ddot{e}(n)$ is the prediction error. The LPCs a_i are determined by minimising the sum squared errors over a given interval, that is, the actual speech samples and the linearly predicted ones as [76];

$$\sum_{n} \ddot{e}^2 = \sum_{n} \left(X(n) - \sum_{i=1}^{k} a_i X(n-i) \right)^2. \tag{2.6}$$

Differentiating Equation (2.6) with respect to $a_i, \quad i = 1, 2, \cdots, k$ yields:

$$\frac{\partial E}{\partial a_i} = \sum_{n=n_0}^{n_1} X(n)X(n-i) - \sum_{j=1}^{k} a_j \sum_{n=n_0}^{n_1} X(n)X(n-i)X(n)X(n-j) = 0,$$

leading to a set of k linear equations:

$$\sum_{i=1}^{k} a_i c_{ij} = c_j, \qquad j = 1, 2, \cdots, k,$$

where

$$c_{ij} = c_{ji} = \sum_{n=n_0}^{n_1} X(n-j)X(n-i).$$

The k unknown quantities can be solved for a_i as explained in [78].

Extracting features using LPC can be achieved in three steps as depicted in Figure 2.3: framing, windowing and computation of the LPC as explained in [28].

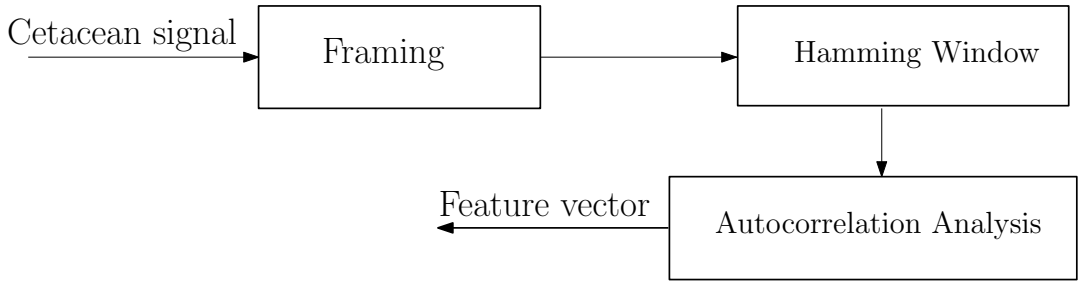


FIGURE 2.3: Steps for feature extraction in LPC technique.

LPC was used as one of the FE tools for the recognition of individual humpback whales in [79]. The extracted coefficients were tested on different classifier models in each of the works. The generated features were useful for the classifier; however, quantisation, stability, and interpolation are some of the drawbacks of LPC.

2.6.6 Mel-scale Frequency Cepstral Coefficients

The Mel-scale frequency cepstral coefficient (MFCC) is a widely used FE technique in signal processing. It has been used to extract features in speech recognition, image identification, gesture recognition, palm recognition, drone sound recognition, speaker identification, and cetacean vocalisation detection and classification [26, 80–83]. Features are extracted in the cepstral domain to build a feature vector set for every type of signal. The widespread use of MFCC is due to its robustness in capturing spectral and temporal features in sound signals. However, it is sensitive to noise due to its dependence on the spectral form. Features are extracted by transforming signals from the time domain into the frequency domain (mel frequency scale). Two filter types exist in MFCC, which are linearly set apart at low frequency below 1 kHz and a logarithm spacing above 1 kHz [29, 84].

Generally, FE using MFCC involves the following steps: framing, Hamming windowing, fast Fourier transform (FFT), the Mel-scale filter bank, logarithm operation,

and discrete cosine transform (DCT), as explicitly explained in [29, 82, 85]. A block diagram of these steps for extracting features using the MFCC technique is shown in Figure 2.4.

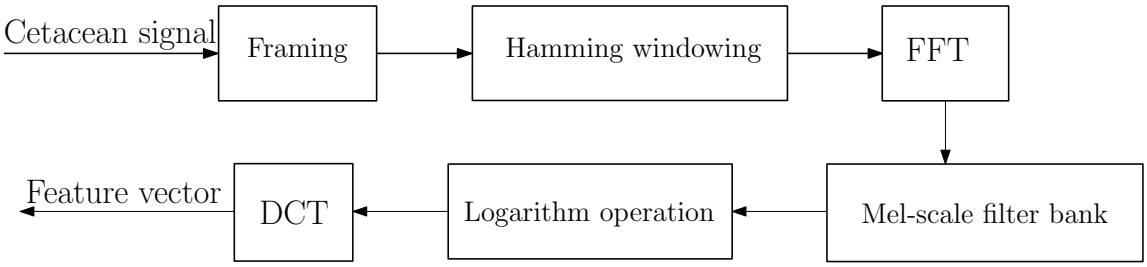


FIGURE 2.4: Steps for feature extraction in MFCC technique.

The corresponding value for frequency f is expressed in Hz as:

$$Mel_{(f)} = 1127 \ln\left(1 + \frac{f}{700}\right) \text{Hz}, \tag{2.7}$$

while the i-th mel-ceptral coefficient is expressed as:

$$\text{MFCC}_i = \sum_{i=1}^{k} X_i \cos\left(\frac{n(i-0.5)\pi}{k}\right), \qquad n = 1, 2, ..., k, \tag{2.8}$$

where k is the total number of cepstral coefficients, and X_i is the logarithmic energy of the i-th mel-spectrum band.

Typically, the first twelve coefficients are utilised to compose MFCC. The set of coefficients is the output feature vectors. Thus, each input acoustic signal is transformed into a sequence of feature vectors. The delta coefficients are included so as to demonstrate the dynamic features.

MFCC has been used as a FE tool in several cetacean vocalisation detection and classification processes. In most cases, it produced better performance than other feature extraction techniques due to its simplicity [86]. Three feature extraction techniques (LPC, cepstrum, and MFCC) were applied to extract essential features on individual HW vocalisations in [19]. The MFCC was shown to outperform the other two methods.

2.6.7 Other Feature Extraction Techniques

There are other FE techniques developed by some researchers who applied them in their work for the extraction of features from cetacean signals. These isolated and tested techniques are potential areas where further research can be done to determine their viability for global application to cetacean signals. Examples of such methods include the Bienenstock, Cooper, and Munro (BCM) theory used in [54] as a feature extraction tool in the design of a classifier. It was tested on sperm whale and porpoise signals. A network of BCM neurons was used to extract features from a wavelet representation. This is reported to have improved the classification accuracy of the signals. Recently, a simple but robust FE technique was proposed in [25] where three parameters-the mean, relative amplitude, and relative power/energy (MAP) from the signals were used to form feature vectors. These feature vectors were used with the HMM for the detection of Bryde's whale short pulse calls. The MAP feature vectors were empirically selected based on the observation of the calls to be detected. The result obtained presents enhanced sensitivity and false discovery rate performance, besides showing a low computational complexity in comparison to the LPC-HMM and the MFCC-HMM detectors. Others are the Teager energy operator (TEO) [87, 88], and the Weyl transform (WyT) [89].

A summary of the reviewed FE techniques in this study is given in Table 2.1. The examples of sound types that each technique has been used to analyse are indicated. The advantages and disadvantages of each of the techniques are highlighted, as are general remarks on the characteristics of the techniques.

2.7 Detection and Classification Methods

Different methods for the automatic detection and classification of cetacean signals have been developed over the years. Certain factors, such as the characteristics of background noise and intrusive sounds, the amount of variation in the species' sound, feature vectors, and whether a template parameter exists for the targeted signal, are

TABLE 2.1: Characteristics of feature extraction techniques.

FE Technique	Sound Type			Advantages	Disadvantages	Remark
	Whistles	*Clicks*	*Pulsed calls*			
STFT	✓	✓	✓	•Easy to implement. •Provide time-localized frequency information of the signal.	•Size of sliding window affects the resolution. •Specific vectors are required for signal decomposition. •Very sub-optimal for signals with varying signal-to-noise ratio (SNR). The SNR varies in signals during recording. •Not optimal when SNR is deficient or when signals are extremely short-lived.	•Uses sliding window to find spectrogram which gives information of both time and frequency. •The length of window limits the resolution in frequency.
WT	✓	✓	✓	•Is more efficient for short-lived signal than STFT. •Provides uniform resolution for all scales. •Domain frequency band is not required when extracting signal features over the entire spectrum. •Ability to view features at different scales or resolution.	•Specific vectors are required for signal decomposition. •Very sub-optimal for signals with varying SNR. The SNR varies in signals during recording. •Poor bias between signals with close high-frequency components.	•WT has been used as FE tools for sperm whale clicks by many authors. This is because the sperm whales clicks are distinctive, notably in their impulsive and brief shape.
HHT		✓		•Decomposes signal into small components. •Very adaptive. •Effective use of available data. •Gives better fine resolutions in time and frequency that leads to simple understanding of result than STFT and WT. •Performs better than other time-frequency methods for signals with varying SNR.	•Number of important IMFs is not known priori. •Components with close frequencies are difficult to screen. •Presence of high energy components in signal may lead to masking problems. •Generation of undesirable IMFs at low frequency may lead to misinterpretation of result. •Signal with low energy component cannot be separated via the EMD operation.	•An empirical based data-analysis technique. •Appropriate for detecting continuous short click signals in the presence of non-constant noise. •Outputs are similar to spectrogram but with higher spectro-temporal resolution than WT.
EMD		✓	✓	•A complete data-driven method. •Eliminates the application of Hilbert transform on the generated IMFs unlike traditional HHT. •Eliminates the preliminary human analysis.	•Presence of high energy components in signal may lead to masking problems. •Sensitive to ambient noise.	•A promising alternative to cetacean signal analysis due to its multi-species detection and classification ability. •Faster computational time. •Does not require manually analysed dataset.
LPC		✓	✓	•Shows higher performance for pulsed calls. •Easy to calculate. •Suitable for extracting features of sound whose source is not well understood. •Suitable for characterising low and mid frequency vocalisation.	•Features generated experience issues of quantisation, stability and interpolation. •Low performance in the presence of noise.	•Less complex than MFCC and easy to calculate.
MFCC		✓	✓	•Low computational complexity. •Suitable for characterising high frequency broadband and amplitude modulated sounds.	•Sensitive to noise due to it dependence on spectral form.	•Widely used feature extraction technique due to it low computation complexity.

considered when choosing a technique for the analysis of marine mammal sounds [14]. Due to variations in acoustic repertoire with respect to region or population, it is important that the model is trained with calls from the intended region for enhanced performance [90]. Among the widely used existing detection and classification methods are SPCC, MF, DTW, SVM, NN, GMM, and HMM. The performance of each technique is dependent on the species, the physical environment where the data are recorded, the size of the datasets available, and, more importantly, the feature vectors used with the technique. However, some techniques do not require a feature extraction process to carry out detection or classification. Such techniques have the ability to learn the inherent features needed for detection or classification during the training of the data.

2.7.1 Spectrogram Cross-Correlation

Spectrogram Cross-Correlation (SPCC) is a correlation technique that is popular among existing methods for bioacoustic detection and classification due to its simplicity of implementation. It only requires a single sound sample of the call type to be detected. It can be effectively applied to either continuous detection or isolated vocalisation tasks where recordings have been presegmented into separate files. A sliding window is applied across a long recording, with correlation peaks signifying target detection [45]. A spectrogram conveniently represents signals by interpreting them in a waveform as a non-negative function of instantaneous frequency and time.

Generally, an input sound signal is converted into a spectrogram through a conditioning procedure; level equalisation and normalisation are then applied. A template vocalisation is cross-correlated directly with a spectrogram [91, 92]:

$$\ddot{S}(t,f) = \left[\sum_{i=0}^{k-1} w(i)x(t+i)\exp\left(\frac{-2\pi jif}{k}\right)\right]^2, \tag{2.9}$$

where $w(i)$ is a windowing function, $X(t+i)$ is signal to be detected, and k is the length of the windowing function to produce an output function $d(t)$ which is the filter output or detection score:

$$d(t) = \sum_{t_1}\sum_{f} S(t+t_1,f)\hat{\mathbf{k}}(t_1,f), \tag{2.10}$$

where $\hat{\mathbf{k}}(t_1,f)$ is the time-frequency kernel function. A threshold is then applied and the times at which the detection function goes over the threshold are considered detection events (the presence of sounds of interest) [93–95].

The SPCC technique gives good performance in the following scenarios:

- For detection of call types when a relatively few instances of call types are known [94, 96].

- When the desired output is to minimise the number of missed calls (false negatives) [94].

SPCC, however, cannot be adjusted to variations in call duration and alignment and is also substantially influenced by changes in frequency, such as shifts triggered by vocal uniqueness among callers as well as high SNR [45, 92]. It is also subject to ocean acoustic propagation effects, that is, signal distortion as the sound propagates through the ocean medium.

The performance of the SPCC technique developed in [96] was compared with that of the MF and the HMM using a set of 114 song samples of bowhead whale end notes. Each of the techniques produced a recognition score for each of the 114 sounds in the sample set. A low detection threshold was chosen in order to determine if the score would be considered a detection event. The SPCC offers better performance than the MF. The SPCC was also compared with the NN technique, but with a larger dataset; the NN, however, performed better than the SPCC. The better performance of NN is due to its ability to manage time variation in bowhead whale vocalisation better than the SPCC, perhaps as a result of the large training set deployed in NN. The SPCC works fairly well with small training set; therefore, this technique can fit in well in situations where small data recordings are available.

SPCC performed best on short sound recordings when used for the automatic detection of North Pacific right whale calls in [46]. It detects 17 calls out of 18 call samples. However, the proportion of false and missed detections increased as the recording duration increased because longer-duration recordings contain a longer period of noise relative to the number of right whale calls present. This emphasises the fact that SPCC works better when a few datasets are available.

Mellinger and Clark [92] developed a detector for the automatic detection of bowhead whale vocalisations using SPCC and MF techniques. A set of bowhead whale vocalisations was analysed with both techniques. SPCC performed relatively better than MF in a substantially noisy recording. The MF, on the other hand, performs better when the noise present in the recording has a flat spectrum. The same set of

bowhead whale vocalisations was tested with the NN method. The NN performed better. However, SPCC and MF work better than NN when a few samples of the signal of interest are present in the data.

2.7.2 Matched Filtering

The matched filtering (MF) technique is similar to the SPCC; it is also a correlation-based technique. It is used in radar systems and image processing analysis [96]. The MF detects a signal of interest by maximising the SNR of the input signal with respect to the noise present in the recordings. This is followed by cross-correlating the sound signal with the known template signal. Generally, the objective of the MF technique is to detect the presence of sound $s(t)$ in the received signal $X(t)$, which is contaminated with additive noise $\ddot{n}(t)$ as defined in [97, 98]:

$$X(t) = s(t) + \ddot{n}(t). \tag{2.11}$$

With respect to cetacean detection, in Equation (2.11), $X(t)$ represents the dataset to be analysed for the presence of species of interest, $s(t)$ represents the known template signal associated with the species of interest, and $\ddot{n}(t)$ represents all other signals (such as shipping sounds or sounds from other species) present in the dataset. There is prior knowledge of the signal of the species of interest (which serves as a template), but the shape of such templates may vary within the species of interest. Due to this prior knowledge, detection can be achieved based on a matched filter. A matched filter can be implemented using a finite impulse response (FIR) digital filter with an impulsed response of $h(t) = s(t - t_0)$. In other words, the filter's impulse response is a time-reversed version of the template signal, maximising the output SNR when the input signal $X(t)$ is applied [99]. The resulting output of the filter $y(t)$ is expressed as:

$$y(t) = y_s(t) + y_{\ddot{n}}(t), \tag{2.12}$$

where the output signal, $y_s(t) = s(t) * h(t)$ is the result of the convolution of the signal with the filter response $h(t)$. Likewise, $y_{\ddot{n}}(t) = \ddot{n}(t) * h(t)$ is the noise output of $X(t)$

[97, 99]. The condition for optimal detection is that the output signal component $y_s(t)$ must be substantially greater than the output noise component $y_{\ddot{n}}(t)$. To satisfy this condition, the filter is to make the instantaneous power in the output signal $y_s(t)$, measured at time $t = t_0$ as large as possible compared to the average power of the output noise $y_{\ddot{n}}(t)$. This is equivalent to maximising the peak pulse SNR as shown in [97]:

$$r_0 = \frac{|y_s(t_0)|^2}{E[y_{\ddot{n}}^2(t)]}, \tag{2.13}$$

where E is the expectation operator.

The MF is an optimal detector if $\ddot{n}(t)$ is a white Gaussian noise random variable. In other words, it is an optimum detector for the detection of sound with white Gaussian background noise and a known signal [92, 99]. However, MF exhibits a poorer performance level than HMM and NN techniques when there is variation in the recordings, which can be as a result of harmonic interference such as ship noise [96]. MF also gives optimal performance for signals with known sources. MF has been applied for the detection of some cetacean species. The target sound of interest, selected by a human expert, serves as a template for the MF design. Any match with this template defines the signal of interest. Measured parameters such as mean frequency and bandwidth are used to synthesise a filter kernel representing the sound type [92]. Its strengths and weaknesses are similar to those of the SPCC [45]. Although it requires more effort to construct the pattern template, it is easy to implement. The stochastic matched filter (SMF) [100, 101] method, which is a better version of the traditional MF, has been proposed in analysing the PAM dataset containing cetacean signals. SMF is attractive due to its ability to both detect and classify. It was used in analysing recorded data in a noisy underwater environment containing Antarctic blue whales [101]. It has been shown to have improved performance when compared with MF.

MF and HMM techniques were applied to a set of 189 recorded underwater acoustic signals and white Gaussian noise in [47] to detect the presence of bowhead whale

notes. The performance of the two techniques was compared, where the HMM selected 97% and the MF selected only 84%. Both methods give almost the same misclassification output (noises that were wrongly classified as bowhead notes): HMM method 51%, and the MF method 49%. These noises are in the same frequency band as the bowhead notes and sometimes resemble a portion of a note.

2.7.3 Dynamic Time Warping

The dynamic time warping (DTW) technique measures the similarity between time temporal sequences, which may vary in speed. In other words, it serves to estimate the similarity between an unknown token and a reference template [102]. It has been explored in many areas of applications like speech recognition, handwriting and online signature matching, sign language recognition, gesture recognition, data mining and time series clustering (time series database search), computer vision and computer animation, surveillance, protein sequence alignment and chemical engineering, music, and signal processing [103].

Suppose we have two time series, a sequence $\mathcal{R}$ of length n and a sequence $\mathcal{G}$ of length m, where

$$\mathcal{R} = (r_1, r_2, ..., r_i, ..., r_n) \text{ and } \mathcal{G} = (g_1, g_2, ..., g_j, ..., g_m). \tag{2.14}$$

The optimal match between these two sequences can be found using DTW, an $n \times m$ distance matrix, $\mathcal{Z}$, where the (i-th, j-th) element of the matrix corresponds to the squared distances, $d(r_i, g_j) = (r_i - g_j)^2$, which is the alignment between points r_i and g_j. The path through the matrix that minimises the total cumulative distance between them is the optimal path. The shorter the distance, the more similarity between points, and vice versa [104]. The distance matrix, $\mathcal{Z}$, is solved to get a contiguous set of matrix $\mathcal{W}$ with elements $\mathcal{W} = (w_1, w_2, ..., w_k,)$ that represents a mapping between $\mathcal{R}$ and $\mathcal{G}$. The connection of each element is called the warping path (warping distance of dynamic time) $\mathcal{W}$. The k-th element of $\mathcal{W}$, $w_k = (i.j)_k$ is the alignment of the i-th point of series $\mathcal{R}$ and j-th point of series $\mathcal{G}$. The shortest path is defined as the DTW distance. There are several paths to be considered;

however, they are not randomly chosen but subject to some conditions, as explained in [103, 104]. The optimal path is the path that minimises the warping cost. Details on the working procedures of DTW are explained in [103].

DTW showed good performance results in the automatic classification of killer whale pulsed calls by presenting precise measurements of the differences in the calls [105]. The pulsed calls are complex sounds with many harmonics, which make their classification more challenging than whistles. Five calls with high SNR from previously classified sounds of captive killer whales from the Marineland of Antibes, France, were used for the experiment. DTW was used to relate the pulsed call contours' fundamental frequencies to all likely pairs of sounds, number by number. The sounds were classified into nine call types. However, preprocessing the measurement of the frequency contours was time-consuming. Brown and Miller in [106] broaden what was done in [105] by investigating the effectiveness of DTW algorithms on more natural recordings that include a diverse collection of species. The DTW was implemented using four different approaches: (1) the Ellis method, (2) the Sakoe-Chiba method, (3) the Itakura method, and (4) the Chai-Vercoe method on a large dataset. The four algorithms give good classification between 70% and 90% despite the presence of biphonic calls and over 100 calls. The results show the versatility of DTW for the analysis of cetacean signals. DTW can be used to observe the movement and habitat inclination of killer whales by tracking sounds heard from remote locations.

Ogundile and Versfeld, in [48], developed a detector using DTW and LPC algorithms for continuous weekly recording of Bryde's whale short pulsed calls. They formed templates from manually identified short pulsed calls from the datasets of each day's recording. The manually identified short pulse calls are from a small section of the recordings, while the remaining larger section (obviously containing other non-targeted sounds) is used to test the detector's performance. Each template has k numbers of samples of variable lengths l. The performance of the detector was substantiated for different values of k. The performance of the detector was tested using 6, 12, and 18 number of samples for each template. The best performance was achieved with 18 number of samples. This indicates that the higher the number of

samples, the better the performance of the detector. Also, the effect of background noise influences the detector's performance.

2.7.4 Support Vector Machine

The support vector machine (SVM) is one of the popularly known ML algorithms that is based on statistical learning concepts [107, 108]. It has enjoyed wide applications in diverse areas of study, such as text categorisation, state estimation, face recognition, image recognition, and cetacean signal classification, among others [107]. SVM is defined in [108] as systems that utilise the hypothesis space of linear functions in a high-dimensional feature space and are trained with a learning algorithm from optimisation theory that implements a learning preference derived from statistical learning theory. The SVM formulation uses the structural risk minimisation (SRM) principle [109], which minimises the upper bound on the expected risk. It has good classification ability through the use of the kernel function mapping technique [109]. The SVM classifier's performance is heavily subject to parameter selection and settings. A detailed explanation of the working principles of the SVM can be found in [107–109].

A multi-class SVM classifier was developed in [110] to classify vocalisations from beaked whales and species of small odontocetes. The classifier, which was dubbed class-specific SVM (CS-SVM), distinguishes among the classes of interest from a referenced class. The class selected is the one with the highest decision function with respect to the reference class. The CS-SVM was structured to recognise the existence of noise, which was treated as a common reference class. This is not the case in a single SVM. A workshop dataset comprising three species (beaked whale, short-fin pilot whale, and Risso's dolphin) of labelled and unlabelled training and test data, respectively, was used to train and test the classifier. A four-class (the three species and the noise) CS-SVM was created which were each trained and tested with approximately 250 signal-present labelled feature vectors and a similar number of noise-only

feature vectors. The training set was used in the optimisation step to find the optimal hyperplane for each class. The performance of the classifier was evaluated using the following metrics: P_{cc} = fraction correctly classified (signal present), P_{miss} = fraction misclassified, and P_{nse} = fraction of noise correctly classified. The classifier was then tested on unlabelled test files with the following average performance: $P_{cc} = 91.5\%$, $P_{miss} = 8.12\%$, and $P_{nse} = 96.7\%$. The performance of the classifier with an unlabelled dataset was reported not to be as good as the result obtained from a labelled dataset. This was attributed to the difference in the selected feature set. Therefore, the method can be further tested with a real-life data set that will indicate the location and method of recording of the data.

Humpback whale vocalisations were classified into song and non-song in [111] using three ML techniques: SVM, NN, and Naive Bayes classifier. Humpback whale vocalisations can be grouped into either songs or non-songs. The song vocalisations are a series of calls organised into a hierarchical structure that can go on for several minutes while non-song vocalisations include feeding cries, bow-shaped and downsweep or meow moans. 70-second-long signals from the data were used as input into the classifiers. The duration matches the hydrophone array signal recording time frame in each of the files in the dataset, which was recorded from the Gulf of Maine. The performance of each of the techniques was evaluated using the accuracy, receiver operating characteristics (ROC) curve, and area under the ROC curve (AUC). The performance of the three classifiers was compared; the MFCC-based SVM led with 94% accuracy, while the MFCC-NN led with 94.27% AUC. The authors intend to ascertain the generalisation of their approach by using data from other regions of the world.

2.7.5 Neural Network

Neural network (NN) is one of the mostly popularly used ML algorithms deployed in different fields to execute tasks such as classification, detection, pattern recognition, prediction and forecasting, and optimisation problems, among others [112, 113]. The

NN work like the biological neurons of the human brain by transforming inputs into outputs. Similar to the human brain, the NN is motivated by an activation function. The mathematical neuron calculates a weighted sum of its k inputs of signals X_i, where $i = 1, 2, \cdots, k$ and the output generated is 1 if this sum is above a definite threshold t, otherwise, the output will be 0. This is mathematically represented as:

$$\ddot{g} = \theta \left(\sum_{i=1}^{k} w_i X_i - t \right), \tag{2.15}$$

where θ is a unit step function at 0, w_i is the synapse weight associated with the i-th input. Equation (2.15) is a threshold activation function, also known as the McCulloch-Pitts model [114]. However, in ML, instead of the threshold, sigmoid σ is used as the activation function:

$$\sigma(X) = \frac{1}{1 + e^{-X}}. \tag{2.16}$$

A large positive value of X gives an output of the sigmoid function that is near 1. The output will be near zero when X is much smaller than 0. The mathematical concept of how NN operates can be found in [112, 113]. Many neurons are present in the NN; each has many weights. The NN learns through trials, and the weights can always be fine-tuned for the NN to learn (training stage).

The NN can be categorised into two categories: (a) feed-forward networks and (b) recurrent (or feedback) networks. Details of the categorisation of NN architectures can be found in [112]. There can also be different connection patterns of NN, and the connection patterns influence the behaviour of the networks. Each of the connection patterns has a specific advantage. The NN architectures are trained using suitable algorithms to learn about the dataset [112]. The ability of NN to learn inherent rules from the given collection of datasets is what makes NN useful to perform various tasks in various fields, which thus makes it appealing. The training of the networks is done with the aim of having outputs that are as close as possible to the desired outputs.

There are three training paradigms: supervised, unsupervised, and hybrid. For supervised learning, an *outside agent* provides the expected output to the networks for every input pattern. The weights are determined to permit the network to bring forth outputs as accurate as possible to the known expected outputs. However, unsupervised learning are self-organising maps (SOM) networks that do not require an external agent to dictate the pattern of their outcome but research the inherent structure in the data or relationships between patterns in the data and make these patterns into classes [112, 115]. In unsupervised NN learning, usually datasets are segregated into disjointed subcategories in such a way that patterns in the same category are as similar as possible and patterns in different groups are as dissimilar as possible [115]. Hybrid learning merges both supervised and unsupervised learning. Comprehensive tutorials on NN have been explained in [112, 113, 116]. NN's ability to look for characteristic correlations in the dataset and form classes based on these correlations make them good candidates for the classification of cetacean vocalisations [115].

A method was developed for the automatic categorisation of bioacoustic signals into biologically relevant categories in [117], using a combination of DTW and adaptive resonance theory (ART) NN, dubbed the ARTwarp algorithm. This method modified ART2 with the adaptation of DTW. DTW was used to compute the similarities between the frequency contours and the set of reference contours in order to ensure maximum overlap in the frequency domain. A dataset of four individuals bottlenose dolphin stereotyped whistles and field recordings of transient killer whale calls were randomly selected for testing the method. The problem addressed here is the categorisation of the signal rather than the usual classification (which is the process of ascribing a sound pattern to predefined categories). The 104 dolphin whistles were partitioned into 46 categories that were consistent with the known biological behavioural patterns of these dolphins.

Jiang *et al.* in [18] put forward a technique based on deep convolutional neural network (CNN) for the detection and classification of the whistles of long-finned pilot whales and killer whales. CNN is a class of deep feed-forward NN that uses

variations of multilayer perceptrons. The detection and classification models were configured together. The models were trained with tagged frame spectrograms and tagged whistle spectrograms, which contain features of each of the whale species. No specific time-frequency features were extracted directly. The data inputs to the models were the time-frequency spectrograms that qualified the entire information of the whistles. Therefore, the feature extraction pattern and the computed features were learned from the training data. The detection segment of the models admits the frame spectrograms of the unknown sounds as inputs and processes whether the corresponding frame spectrogram contains the whistles or not. The detected whistle spectrograms are measured and transmitted to the trained classification model, which in turn identifies the whale species. Their proposed method was able to attain a 97% detection rate and 95% classification rate. The method is said to be adaptable to other whale or dolphin species that produce whistles or other sounds.

A deep neural network-based detector was developed in [51] to detect the vocalisation of North Atlantic right whale calls. Two types of deep neural network architectures were used: CNN and RNN for the detector. They compared the performance of the developed detector with some existing traditional detection methods. The detector was reported to produce a lower magnitude of false positive rate while exhibiting a significantly increased true positive rate. The deployment of deep learning to develop algorithms for the detection and classification of cetacean signals is gaining more prominence, as seen in recent work [18, 39, 51, 118]. Deep learning, being an illustrative learning method where machines automatically learn the representations that are needed from the input raw data, is a promising area for improving on existing techniques for the detection and classification of cetacean species [118].

2.7.6 Gaussian Mixture Model

The Gaussian mixture model (GMM) is a classifier that uses the estimate of the probability density function (PDF) to model densities of different kinds of signals [119]. It has been applied to a broad range of applications, such as sound processing

and image processing. It has the ability to arbitrarily model any type of data distribution by modifying the parameters and number of Gaussian PDFs. A set of k Gaussian distributions of feature vectors $x = (x^1, x^2, \cdots, x^d)^w$ is represented as:

$$\text{pr}\left(x\middle|\mu,\xi\right) = \sum_{i=1}^{k} \vartheta_i \frac{1}{(2\pi)^{\frac{d}{2}}|\xi_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^w \xi^{-1}(x-\mu_i)}, \tag{2.17}$$

where μ_i is the mean, ξ is the covariance matrix of the Gaussian, $|\xi|$ is the determinant of ξ, ξ^{-1} is the inverse of ξ, d is the dimension of the vector, ϑ_i is the mixture coefficients (weight of the i-th Gaussian), and w denotes the transpose operator. The integral sum across the total feature space is 1:

$$\sum_{i=1}^{N} \vartheta_i = 1, \qquad 0 \leq \vartheta_i \leq 1.$$

The two parameters that directly determine the Gaussian distribution are the μ and the ξ. The expectation-maximisation algorithm (EM-alg) [120, 121] is used to obtain the best Gaussian mixture parameters for a given set of feature vectors (see Section 2.7.7.4 for more details on EM-alg). GMMs have been applied to the analysis of cetacean vocalisations. The feature vectors extracted are treated as probability distributions. Feature vectors of size d are assumed to be placed in different areas called clusters within the space when plotting the feature vectors. The number of the k mixture components is empirically chosen, depending on what is aimed to be achieved. In several research work, GMMs have been used to develop techniques for detecting and classifying the vocalisations of different species of cetaceans.

Peso and Cardenal-Lôpez [50] used cepstral coefficients extracted as feature vectors on GMMs to detect and classify sound recordings of underwater signals into one of four types: pulses, whistles, background noise, and combined whistles and pulses. The detection rate achieved was 87.5% for a 23.6% classification error rate (CER). The multiple signal classification (MUSIC) algorithm and an unpredictability measure were introduced as feature vector extractors to address the problem of modelling narrow-band high-frequency signals such as whistles using only cepstral coefficients

in the GMM classifier. This approach significantly improved the detection rate to 90.3% and reduced the CER to 18.1%.

Roch *et al.* [122] developed a GMM classifier to classify free-ranging delphinid vocalisations of four different species of odontocete (long-beaked, short-beaked common dolphins, bottlenose dolphins, and Pacific white-sided dolphins) from sounds recorded at the Southern California bight. The GMMs were trained with different mixtures, which ranged from 64 to 512. The classifier's accuracy increases with an increase in the number of mixtures per GMM. The optimal number of mixtures varies from species to species. The training data size also had a positive effect on the accuracy of the classifier. Different output was recorded when the mixture was retained at 256 with a 20 seconds test segment while the training data size was varied. The overall precision was impacted by the reduction in the amount of training data. For common dolphins, it was noted that the recognition rate was higher with a shorter amount of training data.

GMM and SVM techniques were used to build a classifier that distinguishes between clicks from three species of odontocetes: Risso's dolphin, Blainville's beaked whales, and short-finned pilot whales [87]. The experiments were structured in two parts: (1) to detect the specific clicks produced by species of target, and (2) to classify the set of clicks produced by particular species. Different GMMs; 2, 4, 8, 16, 32, and 64 were created. The 16 mixture models surpassed other mixture models in performance. This result was compared with the output of the SVM using the detection error tradeoff (DET) curve. The DET curve is said to be an efficient plot for highlighting the divergence between similar systems. The DET of the three species was plotted with the following equal error rates (EERs): Blainville's beaked whales-GMM 3.32%, SVM 5.54%, short-finned pilot whales-GMM 16.18%, SM 15.00% and Risso's dolphins-GMM 0.03%, SVM 0.07%.

2.7.7 Hidden Markov model

The HMM is generally used for sequential data analyses. The ubiquity of HMM is based on its rich mathematical frameworks, which serve as the theoretical foundation for deploying the model in a variety of applications [123]. HMMs have been adapted in different fields such as speech recognition [124], data compression [125], signature verification (pattern recognition) [126], bioacoustic analyses [45], and so on. HMMs are useful for explaining several complex, time-varying occurrences, such as whale vocalisations. The ability of the models to manage duration variability through non-linear time alignment is a major reason they are models of choice for the analysis of bioacoustic signals [45].

HMMs are statistical ML models built upon the Markov chain. The HMM is an extension of the first-order Markov chain (model). Every HMM is premised on the assumption that the events being observed rely on some hidden states that are not directly observable. The sequence of the hidden states through time is modelled using the Markov property. Thus, an HMM can be described as a probabilistic positioning identifier that allocates a label to every element in a sequence of observations with the goal of computing the probability distribution over the sequence and then selecting the best observation pattern. In other words, at time t, the HMM matches an active (hidden) state S_t to a known observation Q_t [127, 128].

2.7.7.1 HMM Components

The structure of an HMM can be illustrated using the following five components [123, 124, 129]:

1. A set of N (hidden) states, S, in the model. Each state is represented as $S = (S_1, S_2, \cdots, S_N)$, with S_t denoting the active state at time t.

2. A sequence of $Q = (Q_1, Q_2, \cdots, Q_T)$ observations of length T that match the output of the system being modelled.

3. The $N \times N$ transition probability matrix, Tr,

$$Tr = \begin{bmatrix} tr_{1,1} & tr_{1,2} & \dots & tr_{1,N} \\ tr_{2,1} & tr_{2,2} & \dots & tr_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ tr_{N,1} & tr_{N,2} & \dots & tr_{N,N} \end{bmatrix}, \tag{2.18}$$

whose elements (tr_{ij}) represent the switching probability from state i to state j at one time step where $i, j \in \{1, 2, \cdots N\}$, that is,

$$tr_{ij} = \text{pr}(\mathcal{S}_{t+1} = j \mid \mathcal{S}_t = i), \qquad \text{s.t.} \sum_{j=1}^{N} tr_{ij} = 1 \quad \forall i,$$

where $\mathcal{S}_t$ refers to the active state at time step t, $\mathcal{S}_{t+1} = j$ denotes the next event of state j at time $t+1$ given the current event of state i at time t. Each element is computed using the maximum likelihood estimate of the parameters.

4. The probability of an observation likelihood is called the emission distribution probabilities or Gaussian emission distribution. The emission distribution probability, $\mathbf{E} = \mathbf{E}_i(Q_t)$, indicates the probability of an observation Q_t being generated from a (hidden) state i, that is,

$$\mathbf{E}_i(Q_t) = \text{pr}(Q_t \mid \mathcal{S}_t = i). \tag{2.19}$$

5. The start state probability distribution, $\tau = \tau_1, \tau_2, ..., \tau_N$, denotes the start of the distribution process. The τ is represented by a row vector of size $1 \times N$ and at every stage of the process must sum up to 1; that is,

$$\sum_{i=1}^{N} \tau_i = 1,$$

where τ_i is the probability that the process is starting in state i. Some states j in the distribution may have $\tau_j = 0$, such states j cannot be the start states.

Given the above parameters, an HMM can be compactly represented as Π, where:

$$\Pi = (\tau, Tr, \mathbf{E}). \tag{2.20}$$

2.7.7.2 The Three HMM Problems

HMMs seek to obtain the best sequence of observations from the (hidden) states. Consequently, HMMs address three fundamental problems to provide the best (hidden) state sequence that led to the (physical) observations. The answers to the three problems enable the use of HMMs for real-world applications [123, 124, 127]. The three problems are:

Problem 1 : Evaluation of the probability of an observation sequence. Given the sequence of observations, $Q = (Q_1, Q_2, \cdots, Q_T)$ and the model $\Pi = (\tau, Tr, \mathbf{E})$, HMMs attempt to calculate the probability that the known model will produce the observation sequence, that is, $\text{pr}(Q \mid \Pi)$.

Problem 2 : Determining the model parameters. HMMs seek to refine the model parameters, $(\tau, Tr, \mathbf{E})$, to maximise the $\text{pr}(Q \mid \Pi)$.

Problem 3 : Finding the optimal state sequence. Given the sequence of observation, $Q = (Q_1, Q_2, \cdots, Q_T)$ and the model parameters, $\Pi = (\tau, Tr, \mathbf{E})$, HMMs find the corresponding state sequence that best explains the observation, Q.

In this study, the HMMs for the detection of whale vocalisations are designed to work in two stages: (1) the training stage and (2) the detection stage. **Problem 1** to **Problem 3** are deployed to address each of the stages. During the training stage, HMMs adopt a model for the underlying process that generates the sequence of observations (**Problem 1**) and then estimate the model parameters that will match the given data (**Problem 2**). For the detection stage, HMMs seek the optimal state sequence that best describes the observation sequence, Q, (**Problem 3**). The solutions to the three problems are presented below.

2.7.7.3 Evaluation of the Probability of the Observation Sequence

Problem 1 aims to calculate the probability of the observation sequence, Q, given the model Π, that is,

$$\text{pr}(Q \mid \Pi). \tag{2.21}$$

The basic approach to achieving this evaluation is to compute every possible state sequence of length T and then sum them up. However, this approach is found to be impracticable for many real-world applications due to the acute computational load that would be encountered. Therefore, the forward algorithm [123, 130] provides a more effective way to resolve **Problem 1**. The forward algorithm is a type of dynamic programming where intermediate values are stored in a table as the probability of the observation sequence is built up. The forward algorithm calculates the observation probability by adding the probabilities of all possible hidden state trails that can produce the observation sequence. This computation is performed in an effective way by implicitly folding each of the trails into a single forward trellis [124]. The forward algorithm described in [123, 124] is needed to solve Equation (2.21) as follows:

The forward variable $\alpha_t(j)$ is defined as:

$$\alpha_t(j) = \text{pr}(Q_1, Q_2, \cdots, Q_t, S_t = j \mid \Pi). \tag{2.22}$$

Given the model, Π, Equation (2.22) is the probability of the partial observation sequence, $Q = (Q_1, Q_2, \cdots, Q_t)$ till time t, $S_t = j$ indicating the event of state j at time t. This probability (Equation (2.22)) can be calculated by summing over the extensions of all the paths that leads to the current cell. Given state j at time t, the value $\alpha_t(j)$ is calculated as:

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) tr_{ij} \mathbf{E}_j(Q_t), \tag{2.23}$$

where $\alpha_{t-1}(i)$ is the previous forward path probability from the previous time step, tr_{ij} is the transition probability from previous state i to current state j and $\mathbf{E}_j(Q_t)$ is

the state observation likelihood of the observation symbol O_t given the current state j [124].

Inductively, $\alpha_t(j)$ can be solved as follows:

1. Initialisation:
$$\alpha_1(j) = \tau_j \mathbf{E}_j(O_1), \qquad 1 \leq j \leq N. \tag{2.24}$$

2. Induction:
$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) tr_{ij}\right] \mathbf{E}_j(O_{t+1}), \qquad 1 \leq j \leq N, \qquad 1 \leq t \leq T. \tag{2.25}$$

3. Termination:
$$\text{pr}(O \mid \Pi) = \sum_{i=1}^{N} \alpha_T(i). \tag{2.26}$$

The operation initialises the forward probabilities as the joint probability of state j and initial observation O_1. The induction step shows how state j can be reached at time $t+1$ from the N possible states, i, $1 \leq i \leq N$, at time t. Lastly, the termination step sums up the terminal forward variables $\alpha_T(i)$ to output $\text{pr}(O \mid \Pi)$. In practical terms, an underflow problem could occur with the repeated multiplication of the quantities in Equation (2.24), which are either probabilities or high-dimensional PDF values. However, this issue is resolved by introducing scaling factors to the $\alpha_t(j)$'s computations [123, 131].

2.7.7.4 Determining the Model Parameters

In an attempt to solve **Problem 2**, model parameters $\Pi = (\tau, Tr, \mathbf{E})$ are iteratively re-estimated to maximise the probability of the observation sequence. The observation sequence is used to 'train' the HMM in order to create a model that best fits the given data. Determining the model parameters is the most important and challenging step of the HMM process since it allows optimal adaptation of the model parameters to the observed training data. The forward-backward algorithm called

Baum-Welch algorithm (BW-alg) [132] is an efficient algorithm for iteratively computing the training parameters, Π. The BW-alg is a special type of EM-alg [133] that is deployed for iteratively finding the maximum-likelihood estimate of the parameters of the underlying distributions from the observation. The EM-alg starts by calculating an initial estimate for the probabilities, then uses that estimate to calculate a better estimate, and so on, till it converges [124]. The EM-alg performs a two-stage iteration process by alternating between an expectation (E) step and a maximisation (M) step [131]. The BW-alg requires forward (α) and backward ($\bar{\beta}$) probabilities. The backward probability, $\bar{\beta}$ is computed in a similar way to the forward probability, α. The backward variable $\bar{\beta}_t(i)$ is defined as:

$$\bar{\beta}_t(i) = \text{pr}(Q_{t+1}, Q_{t+2}, \cdots, Q_T \mid S_t = i, \Pi). \tag{2.27}$$

Given state i at time t and the model Π, Equation (2.27) represents the probability of the partial observation sequence from $t+1$ to the end. $\bar{\beta}_t(i)$ is inductively computed like the forward algorithm as follows:

1. Initialisation:

$$\bar{\beta}_T(i) = 1, \qquad 1 \leq i \leq N. \tag{2.28}$$

2. Induction:

$$\bar{\beta}_t(i) = \sum_{j=1}^{N} tr_{ij}\mathbf{E}_j(Q_{t+1})\bar{\beta}_{t+1}(j), \qquad 1 \leq i \leq N, \quad t = T-1, T-2, \cdots, 1. \tag{2.29}$$

3. Termination:

$$\text{pr}(Q \mid \Pi) = \sum_{j=1}^{N} \tau_j \mathbf{E}_j(Q_1)\bar{\beta}_1(j). \tag{2.30}$$

The operation initialises by randomly defining $\bar{\beta}_T(i)$ to be 1 for all i. The induction step computes the backward variable, $\bar{\beta}_t(i)$, which indicates the probability of being in a particular state at each time step and having observed the sequence from the next time step to the end. The termination step computes the backward variable

$\bar{\beta}_t(i)$ for each state at the first time step ($t = 1$) and completes the computation of the entire backward variable matrix. The forward (α) and backward ($\bar{\beta}$) probabilities are used to compute the state occupation probabilities, $\wp_t(i)$, and state transition probabilities, $\ell_t(i)$, which are important for the parameter estimation. The EM-alg performs the 2-step iterative operations that maximise the likelihood of the observed data to estimate the model parameters of the HMM [123, 124, 131]. Algorithm 1 shows the procedure for training the HMM parameter, Π.

The starting parameters for **E** depend on the structure of the distribution. The whale vocalisations are continuous distribution; thus, the selections are done randomly. However, HMMs are sensitive to flat or random starting values [22, 48, 131, 134]. Therefore, two ML tools are usually combined with the HMMs training stage for optimal initialisation of the emission distribution, **E**. The two ML tools are the K-means clustering algorithm and the GMM (discussed in Section 2.7.6). The K-means clustering algorithm [134, 135] is one of the oldest and most commonly used algorithms for data clustering. The algorithm is commonly used because of its ease of implementation, good convergence speed, scalability, and flexibility to sparse data [136]. In this study, both tools are serially combined with HMMs to initialise the emission distribution parameters, **E**.

Algorithm 1: BW-alg for Training HMM parameter, $\Pi = (\tau, Tr, \mathbf{E})$

Input : Q, $\Pi^{\mathbf{e}} = (\tau^{\mathbf{e}}, Tr^{\mathbf{e}}, \mathbf{E}^{\mathbf{e}})$, $\Pi^{\mathbf{e}}$ denotes the initial parameter

Output: $\Pi = (\tau, Tr, \mathbf{E})$

Initialisation: Initialise the model parameter $\Pi^{\mathbf{e}}$

repeat

E-step

Calculate: $\alpha_j(i)$ and $\bar{\beta}_t(i)$ using the forward-backward algorithm;

Compute: $\wp_t(i)$ for each time step and state;

Compute: $\ell_t(i)$ for each time step and pair of states;

M-step

Maximise: τ, Tr using the appropriate Lagrange multipliers;

Maximise: $\mathbf{E}$;

convergence;

repeat

goto: ***E-step*** ;

until *convergence*;

until *convergence*;

Obtain: optimised $\Pi = (\tau, Tr, \mathbf{E})$

2.7.7.5 Detection Stage

The detection stage of the HMM process is meant to provide a sequence of the states that best explains the given observation sequence of the model. The last of the three fundamental problems of HMM (**Problem 3**) is to find the optimal state sequence. This problem represents the detection stage of the HMM process.

Finding the Optimal State Sequence

Problem 3 of the HMM is to find the optimal state sequence, $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_T)$, given an observation sequence, $Q = (Q_1, Q_2, \cdots, Q_T)$ and a model, $\Pi = (\tau, Tr, \mathbf{E})$. This problem is solved using the Viterbi algorithm (Vit-alg) [123]. The Vit-alg is

a type of dynamic programming that is extensively used to solve estimation and detection problems in digital communications and signal processing [137]. The Vit-alg recursively tracks the states of a stochastic process and returns the most likely state sequence [138].

Given observation sequence, $Q = (Q_1, Q_2, \cdots, Q_T)$ and a model, $\Pi = (\tau, Tr, \mathbf{E})$, Vit-alg finds the optimal state sequence, $S = (S_1, S_2, \cdots, S_T)$ by defining the highest probability (best score) along a single path, $\varrho_t(j)$, at time, t as [123]:

$$\varrho_t(j) = \max_{S_1, S_2, \cdots, S_{t-1}} \text{pr}(S_1, S_2, \cdots, S_t = j, Q = (Q_1, Q_2, \cdots, Q_T) \mid \Pi). \tag{2.31}$$

Equation (2.31) accounts for the first t observations and ends in state j. Implicitly, the most probable path is represented by taking the maximum over all possible preceding state sequences.

Given that the probability of being in every state at time t had been computed, the best path probability can be computed by taking the most likely of the extensions of the paths that lead to the current path ϱ_{t+1}. For a given state j at time t, the value $\varrho_{t+1}(j)$ is computed as:

$$\varrho_{t+1}(j) = (\max_i \varrho_t(i) tr_{ij}) \mathbf{E}_j(Q_{t+1}), \tag{2.32}$$

where ϱ_t denotes the previous path probability from the previous time step, tr_{ij} is the transition probability from previous state i to current state j, and $\mathbf{E}_j(Q_{t+1})$ is the state observation likelihood of the observation symbol, given the current state j.

The state sequence is retrieved by keeping track of the argument that maximises Equation (2.32). This is done through the $\kappa_t(j)$ array:

$$\kappa_t(j) = \underset{1 \leq j \leq N}{\text{argmax}}(\varrho_{t-1}(i) tr_{ij}), \qquad 2 \leq t \leq T, \quad 1 \leq j \leq N. \tag{2.33}$$

The recursive procedure of the Vit-alg for finding the optimal state sequence can be summarised as follows:

1. Initialisation:

$$\varrho_1(j) = \tau_j \mathbf{E}_j(Q_1), \qquad 1 \leq j \leq N \tag{2.34}$$

$$\kappa_j = 0, \qquad 1 \leq j \leq N. \tag{2.35}$$

2. Recursion:

$$\varrho_t(j) = \max_{1 \leq i \leq N} (\varrho_{t-1}(i) tr_{ij}) \mathbf{E}_j(Q_t), \qquad 1 \leq j \leq N, \quad 2 \leq t \leq T \tag{2.36}$$

$$\kappa_t(j) = \operatorname*{argmax}_{1 \leq j \leq N} (\varrho_{t-1}(i) tr_{ij}), \qquad 2 \leq t \leq T, \quad 1 \leq j \leq N. \tag{2.37}$$

3. Termination:

$$\text{The best score:} \qquad \mathbf{V}^* = \max_{1 \leq i \leq N} (\varrho_T(i)\, Tr_{ij} \tag{2.38}$$

$$\text{The start of backtrace:} \qquad S_T^* = \operatorname*{argmax}_{1 \leq j \leq N} (\varrho_T(i)). \tag{2.39}$$

4. Path (state sequence) backtracking:

$$S_t^* = \kappa_{t+1}(S_{t+1}^*), \qquad t = T-1, T-2, \cdots, 1. \tag{2.40}$$

The Vit-alg is the same as the forward algorithm except that it takes the maximum over the previous path probabilities (Equation (2.36)), whereas the forward algorithm takes the sum (Equation (2.29)). Also, the Vit-alg has a component called 'back-pointers' that is absent in the forward algorithm. The reason for the backpointers in the Vit-alg is that it must generate a probability as well as the most likely state sequence while the forward algorithm must generate an observation likelihood [124].

In the detection stage of this research, the trained HMM parameters (the given model), $\mathit{\Pi} = (\tau, Tr, \mathbf{E})$, are used to refine the feature matrix, $\boldsymbol{F}$ of whale vocalisations. The Vit-alg then finds the optimal state sequence that best explains the given observation sequence.

HMMs have been used to analyse cetacean signals of different species due to their ability to manage duration variability through non-linear time alignment. It can also easily manage silence or delay within vocalisations [20, 45, 52]. HMMs have also been used to analyse other bioacoustic signals like those of fish [139], birds [86], and mammals [140]. HMMs offer good performance when applied to various bioacoustic signals across a diverse range of species for detection and classification tasks [45].

Putland *et al.* in [20] used the HMM method to detect Bryde's whale vocalisations and it proved to be effective despite the duration difference in Bryde's whale vocalisations and directly overlapping vessel sounds. Classification of individual calls in humpback whale songs was done using HMM in [19]. The size of the training data was varied. 50%, 25%, and 10% of the entire data were selected at different times as training data. The classification performance of each training data size differs. Their results show the best performance when 50% of the data was used for training and the remaining 50% for testing; the overall classification output was 94% as compared to when 25% and 10% were used as training data; the overall classification output was 90% and 78% respectively. This implies that the size of the training data has an effect on the classification performance; the larger the data used during training, the better the performance of the classifier. Although there are instances when lower training sizes give better classification, this can be attributed to the call types; that is, different sound types may require different training sizes. However, it has been shown in [19] that minimising the amount of training set, which reduces human efforts, can enhance the efficiency of the classifier because the computational load and time would have been reduced when running the algorithm. Thus, there exists an opportunity to choose between these two alternatives, subject to one's requirements, considering the give-and-take between time and human efforts needed when training and the performance result.

HMMs have been compared with other techniques used for cetacean vocalisation detection and classification. In most cases, the outcome of this comparison showed that HMM outperformed the other techniques. Instances where HMMs performance was compared with other techniques include HMM and MF in [47], HMM and GMM

in [141], HMM, SPCC and MF in [96]. Although the performance comparison is also subject to the FE technique deployed and how it is implemented with the HMMs [45]. HMM and MF methods were applied to a set of 189 recorded underwater acoustic signals and white Gaussian noise in [47] to detect the presence of bowhead whale notes. HMM was able to detect 97% while MF detected 84% of the bowhead notes present in the data. Both methods give almost the same false positive rate (noises that were wrongly classified as bowhead whale notes); HMM method 51% and MF method 49%. These noises are in the same frequency band as the bowhead whale songs and sometimes resemble a portion of a note. Datta and Sturtivant in [142] applied HMM for the classification of common dolphin signature whistles. The HMM was trained to represent members of the whistles class after feature extraction using MFCC.

2.7.8 Summary of the Techniques

In conclusion to this section, it must be noted that the list of techniques reviewed in this study represents the commonly used techniques for the detection and classification of cetacean species; it is by no means all the techniques. Table 2.2 shows a summary of the characteristics of the detection and classification techniques reviewed in this study. The conclusions to the information provided in the table are arrived at from the reviewed literature. The sound types applicable to each technique are stated. Some techniques fit into all the types of sounds emitted by different cetacean species, while others are only applicable to one or two types of the sound types. As stated earlier, the cetaceans consist of two suborders in their taxonomy: *Odontocete* and *Mysticete*. The species applicable to each of the techniques are grouped according to their taxonomy. In some of the techniques, an FE process is not required. Such techniques can learn to extract needed features through the training of the data. The advantages and disadvantages, as well as general remarks on each technique, are also included in the table.

TABLE 2.2: Summary of surveyed detection and classification techniques.

Technique	Sound Type			Applicable Taxonomy		FE Required	Advantages	Disadvantages	Remark
	Whistles	*Clicks*	*Pulsed calls*	*Odontocete*	*Mysticete*				
SPCC			✓		✓		•Easy to implement. •Works well when comparatively small training data is available. •Results are easy to interpret.	•Cannot adjust to variations in call duration and alignment. •Substantially influenced by frequency variation. •Negatively impacted by ocean acoustic propagation.	•Gives optimal performance for the detection of call types when a relatively few instances of call types are known. •Performs well when the desire output is to minimise the missed calls.
MF		✓	✓	✓	✓		•Optimum detector for signal with white Gaussian background noise. •Easy to implement. •Results are easy to interpret.	•Poor performance when there is even little disparity from sound to sound or harmonic interference such as ship noise. •Negatively impacted by ocean acoustic propagation.	•Optimal detector for known signal with the presence of white Gaussian noise.
DTW	✓	✓	✓	✓	✓		•High performance rate with few dataset. •Good at recognising individual call.	•Performs poorly with large dataset. •Takes longer time in preprocessing measurements of the frequency contours.	•Suitable in application when a specific call is to be recognised.
SVM		✓	✓	✓	✓	✓	•Presents good output for datasets with overlapping characteristics. •Training of dataset is simple.	•Poor performance with large datasets. •Slow computational time during training and testing.	•Performance is heavily subjected to parameter selection and settings. •Kernel function mapping technique enhances it classification ability.
NN	✓	✓	✓	✓	✓		•Highly adaptive and learn fast. •FE technique may not be necessary for deep learning NNs such as CNN.	•Requires large training dataset. •Faces problem of over-fitting if training goes on for too long whereby noise may be mistaken for sound of interest. •Data training can be slow due the necessarily large training data requirement.	•NNs associated operating time can be more due to requirement of large training dataset. •Performance improve with the availability of more data for training, however faces problem of over-fitting if training goes on for too long.

Technique	Sound Type			Applicable Taxonomy		FE Required	Advantages	Disadvantages	Remark
	Whistles	*Clicks*	*Pulsed calls*	*Odontocete*	*Mysticete*				
GMM	✓	✓	✓	✓	✓	✓	•Can efficiently model randomly complex distributions with multiple modes •An efficient classifier.	•Inability to take note of the sequential evolution reduces it classification ability.	•In GMM, the whole vocalisation is considered as an entity with exclusive properties that characterises each class. This reduces it classification capability in comparison to HMM. •Structurally equivalent to ergodic HMM.
HMM	✓	✓	✓	✓	✓	✓	•Wide availability of well developed tool-set. •Can seamlessly manage duration variability through non-linear time alignment. •Can efficiently manage silence or delay within vocalisations.	•More data are required to estimate the model parameters. •Sensitive to a flat start of the emission distribution parameters.	•Size of training data impacts on classification performance. However, there are instances where sound types also influence performance. •Minimising the amount of training can enhance the efficiency in terms of the classifier.

2.8 Output Parameters

The metrics used to determine the performance of each method developed and tested vary. Different methods have been developed for the detection and classification of cetacean species. Detector or classifier outputs are characterised by certain parameters that determine their performance level. However, the reporting style of performance measurements varies among authors. This variability can be attributed to the use of different techniques for analysing cetacean signals.

Like every scientific research project, the outcomes of research in this field are reported in terms of observable parameters. These parameters, or metrics, demonstrate the level of accuracy attained by the designed detector or classifier. Though there are a number of standard reporting metrics, some authors do adopt their own style

of reporting outputs. No method is 100% perfect; any method will produce false negatives or false positives [14], depending on the analysed signals, the FE technique, or types of detector and classifier.

The common metrics used in reporting the output of detection and classification processes are false positive rate (FPR), true positive rate (TPR), error rate (ERR), accuracy (ACC), and precision (PREC). These metrics rely on the four outcome parameters of a binary classification model. The four outcome parameters are defined as follows:

- False positives (FP): This is when a detector wrongly detects a signal as a signal of interest; that is, the number of sounds wrongly detected as sounds of interest.
- False negatives (FN): This occurs when the sounds of interest are missed. The false negatives are also known as *missed calls*.
- True positives (TP): This is the number of instances when a detector's output matches the sounds of interest as manually identified by the human expert.
- True negatives (TN): This is the correct prediction of the absence of sounds of interest by the detector.

The performance measuring metrics are defined as follows:

1. **False positive rate (FPR)**: This is the proportion of sounds of interest that are wrongly detected by the detector; it can be mathematically expressed as:

$$\text{FPR} = \frac{\text{FP}}{\text{TP} + \text{FP}}. \tag{2.41}$$

   The best FPR is 0.0 while the worst is 1.0.

2. **True positive rate (TPR)**: This is the proportion of sounds of interest that are correctly detected by the detector. The TPR is also known as sensitivity

or recall. It can be mathematically expressed as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.42}$$

The best TPR is 1.0 while the worst is 0.0.

3. **Error rate (ERR)**: This is computed as the total number of all wrong predictions (FP and FN) divided by the total number of prediction outcomes (FP, FN, TP, TN) of the dataset; it can be mathematically expressed as:

$$\text{ERR} = \frac{\text{FP} + \text{FN}}{\text{total number of predictions}}. \tag{2.43}$$

The best ERR is 0.0 while the worst is 1.0.

4. **Accuracy (ACC)**: This is computed as the total number of correct predictions (TP and TN) divided by the total number of prediction outcomes (FP, FN, TP, TN) of the dataset; it can be mathematically expressed as:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{total number of predictions}}. \tag{2.44}$$

The best ACC is 1.0 while the worst is 0.0.

5. **Precision (PREC)**: This is computed as the total number of correctly detected sound of interest divided by the total number of detection (both TP and FP) in the dataset; it can be mathematically expressed as:

$$\text{PREC} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.45}$$

The lower the number of FP predictions, the better the PREC value. Therefore, the best PREC value is 1.0, while the worst is 0.0.

Other metrics are used to further evaluate the performance of detection and classification models. These metrics rely on combinations of the above metrics with the goal of gaining more insights and drawing conclusions about the performance of models. They include:

6. **Receiver-Operating-Characteristics curve (ROC)**: This is a plot of the FPR on the x-axis against the TPR on the y-axis at various probability thresholds. It indicates the sensitivity and specificity of a model at different thresholds, thereby enabling the selection of a point that establishes the right trade-off.

7. **Area Under the Curve (AUC)**: This is also a graphical metric that is used to assess the overall performance of models, often associated with the ROC curve. The AUC represents the area under the ROC curve, ranging from $(0, 0)$ to $(1, 1)$. The combination of the ROC curve and AUC reveals how well a model can effectively differentiate between positive and negative predictions.

8. **Precision-Recall (PR) Curve**: This is another graphical metric that involves the plotting of the recall (TPR) on the x-axis against the precision (PREC) on the y-axis. It is particularly useful when working on imbalanced datasets, where one class is dominant over the other.

9. F_1 **Score**: This is defined as the harmonic mean of precision (PREC) and recall (TPR) and can be mathematically represented as:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{2.46}$$

   The best F_1 score is 1.0, while the worst is 0.0. The F_1 score integrates PREC and TPR into a single metric to provide more insight about the model's performance.

Each designed detector or classifier has a unique set of thresholds or sensitivity, subject to what it is intended to be achieved. For instance, in a survey of a relatively rare species such as the *Right whales*, the detector may be configured in such a way that there are as few *missed calls* (FN) as possible due to the availability of few datasets, but such a detector may have a large number of FP. On the other hand, in a survey of a common species with abundant datasets available, where an accurate index of detection or classification is important, the detector can be configured in such

a way that the sensitivity level is low so as to reduce the number of FP detections and achieve high TPR [41]. However, some authors, such as in [75], have been observed to report the output of their research using some other metrics to compare the performance of their detector or classifier. Though this does not really matter, in as much as there is clear reporting of the metrics used in determining the performance level of the detector or classifier.

Generally, it is common practice to adopt more than one metric to evaluate the performance of a model. This is because of the limitations (or biases) that are associated with each of the metrics and the characteristics of the dataset involved. For instance, the ACC performs poorly in evaluating the performance of an imbalanced dataset, while a metric like the F_1 score combines PREC and TPR into one metric to provide more insight about the model's performance on an imbalanced dataset. Moreover, the PR curve has been suggested to be more advantageous than the ROC curve because of its ability to work on a skewed dataset, owing to its independence on TN predictions [143]. In summary, the choice of performance metrics to be adopted to evaluate the performance of a model will depend on the specifics of the problems and the goal of the model.

2.9 Summary of Findings and Conclusion

The design of automatic techniques for the detection and classification of cetacean vocalisations has greatly assisted with the processing of large acoustic datasets that are acquired from long-term recordings during PAM. It has also helped eliminate human bias and errors associated with manual detection and classification. Thus, automatic detection and classification have led to faster computational times and greater consistency in comparison to manual detection and classification. Automatic detection and classification methods have assisted ecosystem managers in having a better understanding of the ecology of cetaceans.

It was observed that various factors influence the performance of the methods. These factors include the FE technique deployed, the species involved, recording quality, the duration of the recording, and the location of the recording. It was also observed that the procedures adopted in the implementation of the techniques by researchers varied. Some of the techniques may have more than one algorithm for their implementation. An example is the DTW method used for the classification of killer whale vocalisations in the work of Brown and Miller [106]. Four algorithms are used to implement the DTW, with each giving different outputs. Furthermore, researchers often use more than one method to carry out detection and classification on a given dataset and compare the performance of each method. An instance of such a scenario is in Garcia *et al.* [144], where five different classifiers- SVM, CNN, long-short-term memory (LSTM) network, logistic regression, and decision tree-are used to classify a large volume of fin whale vocalisations. The comparison is carried out using different metrics like ACC, PREC, and TPR. It has also been proven that the methods for the analysis of whale species signals can be very adaptive.

Despite the availability of different detection and classification methods, there are still some challenges facing this research area. One of the challenges is the comparison of the performance outputs among different methods, mainly due to the choice of the FE technique deployed. The FE techniques are often universal [44] in approach (e.g. MFCC, LPC, HHT) but some adjustments can be made to them to fit in properly with the characteristics of the signals to be analysed. Feature vectors play a pivotal role in the output of any detector/classifier [45]. Thus, the FE techniques must be carefully selected. Background noise as a result of harsh recording surroundings, the hydrophone type deployed, and sound propagation unpredictability can be inimical to the detection and classification process.

Furthermore, there is a need to make available more wide-ranging, expert-certified species sound catalogues, such as the one available on *MobySound*[1]. The collection of sound can be a laborious and expensive task for signal analysts to carry out. This can be addressed if experts (biologists) collaborate to develop harmonised catalogues for

[1]http://www.mobysound.org/mobysound.html

different species. These catalogues can lead to the launch of a centralised database that is well annotated (with information such as the date and time of recordings, location of recording, and so on) and can be made accessible to researchers working on the development of automatic detectors and classifiers. Such a database would greatly facilitate the work of signal analysts tasked with the design of detectors and classifiers, providing them with a valuable resource for reference and training.

More research efforts focusing on the potential of adopting techniques for analysing non-stationary signals in other field subject areas can lead to new discoveries. Continuous improvement in FE techniques will enhance the performance of existing detection and classification methods. There is a need for researchers to explore the opportunities of adapting FE techniques used in other research areas into this field. Prospective techniques, such as dynamic mode decomposition (DMD) [145], should be considered, as they possess the capability to extract feature vectors from non-stationary and non-linear signals. This study is focused on developing new FE techniques to be adapted to one of the leading methods (HMM) for the detection of whale vocalisations.

We surveyed several previous studies on automatic methods for the detection and classification of different cetacean species. Table 2.3 shows a summary of a few of the papers surveyed in this work. From the surveyed literature, we noted the tasks implemented: detection, classification or both. The recording method used is also stated: either fixed PAM or mobile PAM. The FE techniques, the detection and classification methods, as well as the species involved in the study. General remarks on what was done in each of the surveyed pieces of literature are given.

The hidden Markov model (HMM) is the central theme of this thesis since it is the detection method this study seeks to improve. Therefore, the next two chapters focus on the development of ED-FE techniques to be used with the HMM for improved performance.

TABLE 2.3: Summary of past work surveyed on detection and classification techniques.

Reference	**Task Performed**	**Recording method**	**FE Technique**	**Detection Classification Technique**	**Species Analysed**	**Data Collection Location**	**Sound Type**	**Remark**
[20]	•Detection	Fixed PAM	•MFCC	•HMM	•Bryde's whale.	•Hauraki Gulf, New Zealand.	•Pulsed calls.	Hidden Markov model Toolkit (HTK) was applied for the detection of Bryde's whale vocalisations. The method proved effective despite the duration difference in Bryde's whale vocalisation and overlapping passage vessel sounds. The work showed that HTK could further be explored for the analysis of signals from other species of cetaceans. Further work is needed to ascertain the robustness of HTK for the detection of other species of cetaceans that produce characteristic calls like the Bryde's whale.
[90]	•Detection •Classification	Mobile PAM	•STFT	•Threshold	•Short-beaked common dolphin. •Melon-headed whale. •Bottlenose dolphin.	•North Atlantic regions.	•Whistles.	The noise cancellation technique was applied to the spectrogram of data to search for connected regions that rose above a preset threshold. Detection takes place when energy within a stipulated frequency band has exceeded the threshold set. The classifier was designed to handle fragmented whistle detection or partially detected whistles. However, classifier performance deteriorates with an increase in the number of species in the dataset.
[50]	•Detection •Classification	Fixed PAM	•MFCC	•GMM	•Bottlenose dolphin. •Common dolphin. •Long-finned pilot whale.	•North-west Spanish coast.	•Pulsed calls. •Whistles	GMM used to develop a technique to detect sound recordings and classify them as one of the following four types: pulses, whistles, background noise, and combined whistles and pulses. A 23.6% classification error rate (CER) and 87.5% TPR were achieved with MFCC due to the narrow bandwidth in whistles; however, the introduction of the MUSIC algorithm and unpredictability measure improved the results to 18.1% CER and 90.3% TPR. The viability of the MUSIC algorithm and the unpredictability measure can be further explored in future research to analyse high-frequency signals with narrow bandwidths.
[18]	•Detection •Classification	Not stated	•STFT	•Deep CNN	•Killer whale. •Long-finned pilot whale.	•Canada. •Norway. •Mexico. •USA.	•Whistles.	A novel technique based on deep convolutional neural networks (CNN) was developed for the detection and classification of the whistles of killer whales and long-finned pilot whales. Though no feature extraction technique is used to extract features directly, STFT was used to create whistle contours in the annotated spectrogram. The data input to the models were the time-frequency spectrograms that qualified as the whole information of whistles. Therefore, the feature extraction structure and the computed features were learned directly from the training data. A graphic user interface (GUI) was developed to aid visualisation of the results of the detection and classification procedures. The accuracy of this method is, however, dependent on the size of the training data.
[146]	•Detection •Classification	Fixed PAM	•MFCC	•HMM	•Blue whale.	•Corcovado Gulf, Chile.	•Songs. •Pulsed calls.	A new approach to HMM for the detection and classification of blue whales is introduced using the Kaldi speech recognition toolkit. The Kaldi speech recognition toolkit looks promising for analysis of other cetacean signals.

TABLE 2.3: *(Continued.)* Summary of past work surveyed on detection and classification techniques.

Reference	**Task Performed**	**Recording method**	**FE Technique**	**Detection Classification Technique**	**Species Analysed**	**Data Collection Location**	**Sound Type**	**Remark**
[147]	•Classification	Mobile PAM	•MFCC	•GMM	•Risso's dolphin. •Bottlenose dolphin. •Pacific white-sided dolphin. •Cuvier's beaked whale.	•California, USA.	•Clicks.	Improved features fed into the GMM classifier for the classification of different odontocetes species using recordings of their echolocation clicks were proposed. The performance of the algorithm differs from species to species.
[19]	•Classification	Fixed PAM	•MFCC •LPC	•HMM	•Humpback whale.	•Ste. Marie Island, Madagascar.	•Songs.	HMM was applied to analyse humpback whale songs. The size of the training data directly influences the classification output. Large training sets lead to high performance. However, there are other instances where the sound type also influences the classification output. The data used has a high SNR which may not be so in continuous recording taken in the field; therefore, further work is needed to confirm the relativity of training data size and sound type to classification outputs in HMM.
[141]	•Classification	Not stated	•MFCC	•HMM •GMM	•Killer whale.	Not stated.	•Pulsed calls.	HMM and GMM were used to classify a set of 75 calls of Northern resident killer whales into call seven types. Their results establish that both GMMs and HMMs are effective in the task of automatic classification of killer whale call, though HMM performs better. The major dissimilarity between HMM and GMM is that in HMM, the sequential evolution of the sound is noted; therefore, it is able to illustrate the structure of the calls. This ability to note the sequential evolution of the sound enables it to have more information to clarify among the call types detected. GMM, on the other hand, considers the whole sound as an entity with exclusive properties that characterise each class. The details of the data used were not given; these could have helped in giving more insight into the work.
[22]	•Detection	Mobile PAM	•EMD	•HMM	•Bryde's whale.	•False Bay, South Africa.	•Pulsed calls.	The EMD method was used to extract features from Bryde's whale pulsed calls. These features were fed into the HMM to carry out detection. The EMD-based FE method enhances the performance efficiency of the HMM when compared with the existing LPC-HMM and MFCC-HMM.
[49]	•Detection •Classification	Not stated	•MFCC •DWT	•SVM	•North Atlantic Right whale.	•Florida, USA.	•Pulsed calls.	An integrated algorithm that combines two FE as well as SVM as classifiers is proposed for the detection and classification of North Atlantic right whale calls. A more precise and faster computational time was achieved with this algorithm.

TABLE 2.3: *(Continued.)* Summary of past work surveyed on detection and classification techniques.

Reference	**Task Performed**	**Recording method**	**FE Technique**	**Detection Classification Technique**	**Species Analysed**	**Data Collection Location**	**Sound Type**	**Remark**
[94]	•Detection	Fixed PAM	•STFT	•SPCC •NN	•Right whales.	•Jacksonville Florida, USA. •Massachusetts, USA.	•Pulsed calls.	The performances of SPCC and NN techniques were compared on right whale calls. Due to the influence of SNR on detector performance, testing was done separately for calls of different SNR. Results show NN performed better than the SPCC with only 6% error rate compared to the SPCC 26%. However, NN requires more datasets for training, while SPCC requires only a few datasets to give good performance. A good explanation of the detection process using the SPCC technique for cetacean sound analysis is given.
[96]	•Detection	Fixed PAM	•STFT •MFCC	•SPCC •HMM •NN •MF	•Bowhead whale.	•Alaska, USA.	•Songs.	The performance of three techniques (HMM, MF, and SPCC) were compared in the detection of bowhead whale songs. MF gives improved performance in the presence of Gaussian noise, while SPCC works fairly well with few datasets for training. The suitability of SPCC for extensive analysis of cetacean signals cannot be guaranteed due to its poor handling of time variation.
[115]	•Classification	Fixed PAM	•No FE used	•NN	•Killer whale.	•Sea Life Park, Hawaii USA.	•Whistles, •Clicks, •Pulsed calls.	Two types of NNs: (1) a competitive network and (2) a two-dimensional feature map were used for the classification of killer whale vocalisations. Both networks are trained with a combination of duty-cycle and peak-frequency input values, with both having complementary outputs; not withstanding, each of the networks has its own advantages. The competitive network was efficient in finding the minimum number of probable categories from the dataset, while the feature map was able to show additional properties such as the relative distribution of the feature space and topological correlations among categories. The unsupervised do not require any FE technique because the NN are able to detect patterns in their inputs.
[72]	•Detection •Classification	Fixed PAM	•HHT	•HHT	•Sperm whale.	•Strait of Gibraltar, Spain.	•Clicks.	HHT's ability to decompose non-stationary signals makes it one of the alternatives to time-frequency methods. It is used to detect sperm whale clicks in a continuous signal. HHT provides a better time-frequency localisation than STFT and WT, which leads to an easier interpretation of the result than STFT and WT. However, the method is faced with a masking problem in the presence of a high-energy component in the signal.
[148]	•Classification	Fixed PAM	•MFCC	•SVM	•Humpback whale.	•Chi-chi Island, Japan.	•Songs.	A SVM classifier centred on cepstral coefficient feature vectors was proposed for the classification of humpback whale songs. A 99% classification accuracy was attained with this algorithm compared to 88% classification accuracy attained in [149] where cepstral features are used on GMM classifiers.

TABLE 2.3: *(Continued.)* Summary of past work surveyed on detection and classification techniques.

Reference	Task Performed	Recording method	FE Technique	Detection Classification Technique	Species Analysed	Data Collection Location	Sound Type	Remark
[75]	•Detection •Classification	Not stated	•EMD	•EMD	•Beluga whale. •Sperm whale. •Humpback whale.	•Bering Sea, USA.	•Whistles, •Clicks, •Songs.	A novel method for detection and classification using only the EMD was proposed. The generated IMFs were assigned unique EMD identities. These unique labels are used to classify the detected sound sources. This new method approach is a modern way to carry out unsupervised detection and classification on transient cetacean signals in the time domain, depending entirely on EMD-type processing, eliminating the requirement to apply the Hilbert transform and manual labelling of pre-processed data by an expert. However, it performs poorly in the presence of extreme values in the signals. Therefore, this method can further be explored through the analysis of other species of cetaceans to ascertain its robustness, particularly with regard to data size.
[25]	•Detection	Mobile PAM	•MAP	•HMM	•Bryde's whale.	•False Bay, South Africa.	•Pulsed Calls.	A novel FE method was adapted with HMM for the detection of Bryde's whale pulsed calls. The new FE method draws on the strength of three simple but robust parameters: the mean, relative amplitude, and relative power (MAP) of the signals to be analysed. The selection of the parameters was based on empirical observation of the calls to be detected. The results, besides showing low computational complexity, also gave a higher sensitivity when compared with the existing LPC-HMM and MFCC-HMM detectors.
[38]	•Detection •Classification	Not stated	•CWT	•BP-NN	•Sperm whale. •Long-finned pilot whale.	•Not stated.	•Clicks.	Back propagation (BP) NN is used for the classification of clicks from complex overlapping sperm whales and long-finned pilot whale vocalisations found in the same dataset. The CWT approach was used to decompose picked clicks of sperm whales and long-finned pilot whales; afterwards, a wavelet coefficient matrix was obtained from every single picked click. It provided better time resolution and frequency resolution when compared with STFT and other time-frequency transform methods. However, the size of the data greatly influences the performance.
[67]	•Detection	Not stated	•CWT •STFT	•STWE	•Sperm whale.	•Strait of Gibraltar, Spain.	•Clicks.	A novel detection algorithm called Short-Time Windowed Energy (STWE) was proposed for the characterisation of a particular shape present in the time-frequency domain of sperm whale clicks. Further work is recommended on this technique to determine its viability for global application in cetacean signal analysis.

TABLE 2.3: *(Continued.)* Summary of past work surveyed on detection and classification techniques.

Reference	**Task Performed**	**Recording method**	**FE Technique**	**Detection Classification Technique**	**Species Analysed**	**Data Collection Location**	**Sound Type**	**Remark**
[87]	•Classification	Mobile PAM	•MFCC	•SVM •GMM •TEO	•Blainville's beaked whales. •Short-finned pilot whales. •Risso's dolphins.	•Gomera Island, Spain. •Canary Islands, Spain.	•Clicks.	Three cetacean species' clicks are classified using GMM and SVM. TEO was to locate individual clicks, while the MFCC was deployed in the construction of the feature vectors for the classifier. The results from each classifier were compared using the detection error tradeoff (DET) curve. A DET curve is a graphical plot for highlighting divergence between similar systems. However, for SVM, some clicks might have been wrongly classified due to the absence of a distributional tactic.
[122]	•Classification	Mobile PAM	•MFCC	•GMM	•Short and long-beaked dolphins. •Pacific dolphin. •Bottlenose dolphin.	•Southern California Bight, USA.	•Clicks, •Whistles, •Pulsed calls.	GMMs are trained with different mixtures, which vary from 64 to 512. The classifier's accuracy increases with an increase in the number of mixtures per GMM. The reduction in the amount of training data impacted the overall and mean species accuracy. The method developed here can be further tested on data from other locations to determine the impact of relativity on the species signals.
[150]	•Detection	Not stated	•No FE used	•Deep CNNs	•Sperm whale.	•Xiamen bay, China.	•Clicks.	A novel method using deep CNN methods is put forward to automatically detect echolocation clicks. The application of deep CNN led to a fast computational time and overcame the problems of fixed-size input. The method can be tested with datasets that contain real-life noise to ascertain its effectiveness.
[39]	•Detection •Classification	Not stated	•No FE used	•NNs	•Sperm whale.	•Eastern Caribbean, Dominica. •Galapagos Island, Ecuador.	•Clicks.	A robust NN technique using CNN and RNN approaches was used to build a detector and classifier for sperm whale echolocation clicks. The NNs are designed to understand the inherent features needed to carry out the detection and classification; no feature extraction technique was used. This method is promising for the analysis of large sperm whale clicks from raw recordings.
[101]	•Detection •Classification	Fixed PAM	•STFT	•SMF, •MF	•Blue whale.	•South west Indian Ocean.	•Pulsed Calls.	SMF gives reduced FPR despite the presence of background noise in the data compared to the traditional MF technique. However, an accurate estimation of the background noise is needed in order to maximise the SMF output.
[48]	•Detection	Mobile PAM		•DTW •LPC	•Bryde's whale.	•False bay, South Africa.	•Pulsed Calls.	A detector was developed for Bryde's whale short pulse calls using DTW and LPC. The performance of the DTW-based and LPC-based detector were compared. The LPC-based detector slightly outperform the DTW-based detector. The LPC-based detector is not novel because the LPC method has been hitherto used for feature extraction.
[142]	•Classification	Mobile PAM	•MFCC	•HMM	•Common dolphin.	•Not stated.	•Whistles.	The work demonstrates the use of HMM for the classification of common dolphin signature whistles. There is a need to further look into the suitability of this technique for the classification of signature whistles with a large dataset.

TABLE 2.3: *(Continued.)* Summary of past work surveyed on detection and classification techniques.

Reference	**Task Performed**	**Recording method**	**FE Technique**	**Detection Classification Technique**	**Species Analysed**	**Data Collection Location**	**Sound Type**	**Remark**
[117]	•Classification	Fixed PAM	•STFT	•DTW •NN	•Bottlenose dolphin. •Killer whale.	•Duisburg, Germany.	•Whistles, •Pulsed calls.	A novel approach called ARTwarp was proposed by combining DTW and ART2 NN techniques to categorise bioacoustic recordings containing bottlenose dolphin whistles and killer whale pulsed calls into biologically significant categories. However, the method can only address peculiarities of acoustic perception rather than the behaviour of the species. The method cannot classify individual signature whistles.
[110]	•Classification	Not stated	•MFCC	•SVM	•Blainville's beaked whale. •Short-fin pilot whale. •Risso's dolphin. •Sperm whale.	•Not stated.	•Clicks.	A novel classifier dubbed class-specific SVM (CS-SVM) was proposed. It is a multi-class SVM classifier that distinguishes among the classes of interest from a referenced class. The method can be further tested with real-life recording, as the location and method of recording the data used were not stated.
[105]	•Classification	Fixed PAM	•STFT	•DTW	•Killer whale.	•Marineland of Antibes, France.	•Pulsed calls.	DTW showed great performance in the automatic classification of killer whale pulsed calls by presenting precise measurements of differences in the calls. The dataset used here was from captive killer whales. Further work has been done using more natural recordings that include a diverse collection of species in [106] to test the robustness of this method.
[106]	•Classification	Fixed PAM	•STFT	•DTW	•Killer whale.	•Northern Resident, Canada.	•Pulsed calls.	This research is an extension of what was done in [105] to ascertain the robustness of DTW call classification. Four approaches were used to develop the DTW, which was tested on a larger dataset recorded on the open sea. The results achieved show the versatility of DTW for the analysis of killer whale signals. Further experiments can be carried out on the calls of other species of cetaceans.
[151]	•Detection	Fixed PAM	•MFCC	•GMM	•Blue whale.	•Guafo Island, Chile.	•Pulsed calls.	GMM was used for the detection of blue whale calls through the clustering of features into sets. The work demonstrated the strength of unsupervised GMM in detecting blue whale calls from recordings containing sounds from other sources by separating the calls into clusters. Other clustering approaches can be explored in later work to confirm the number of clusters that can be formed from a species' sound.

Chapter 3

Eigendecomposition-based Feature Extraction Techniques

3.1 Introduction

This chapter focuses on the development of the proposed eigendecomposition-based (ED) feature extraction (FE) techniques. The datasets gathered during the passive acoustic monitoring (PAM) process come with other sounds besides the whale vocalisations. Hence, there is a need to extract the characteristic features of sounds of interest. The features are numerical representations of an aspect of the raw data that can be related to the sounds of interest. Feature engineering, an act of extracting features from raw data and transforming them into formats that are suitable for machine learning (ML) models, is a fundamental phase of the ML process. In particular, the quality of the feature vectors used with the HMM influences its performance.

The technique deployed in extracting features from datasets depends on the structure of the data and its adaptability to the ML algorithm in which it will be used. Several FE techniques exist, fitting into different ML algorithms. In this chapter, two ED algorithms–principal component analysis (PCA) and dynamic mode decomposition (DMD)–are respectively introduced as techniques to extract features from

PAM datasets. The features are to be adapted with the HMM for the detection of whale vocalisations.

The schematic diagram in Figure 3.1 illustrates the comprehensive workflow proposed in this study for the detection of whale vocalisations. The diagram encompasses stages from PAM data collection and preprocessing to feature extraction using ED algorithms, detection using Hidden Markov Models (HMM), and subsequent result evaluation.

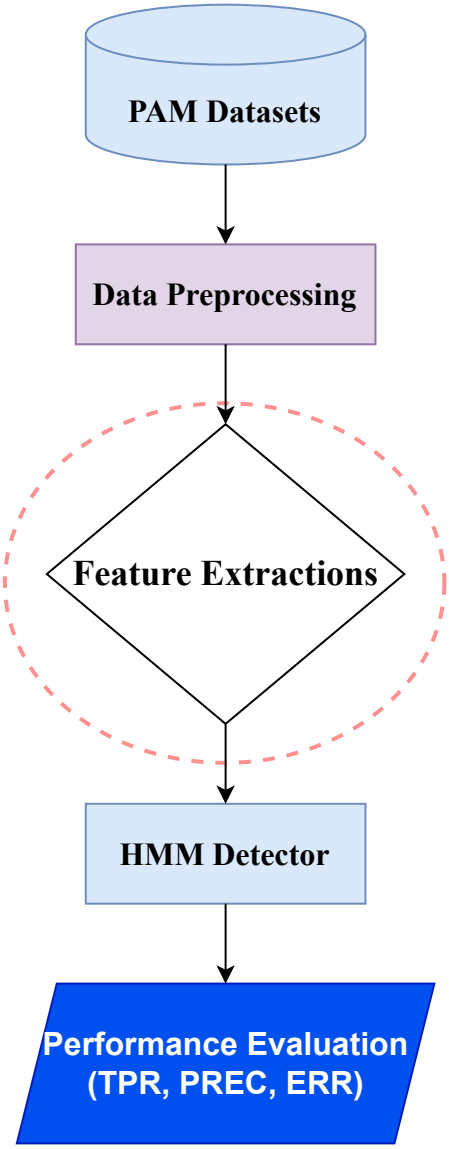


FIGURE 3.1: Schematic workflow of the proposed ED-HMM model for the detection of whale vocalisation.

3.2 Passive Acoustic Monitoring (PAM) Data

The PAM recordings containing whale vocalisations and other sounds in .WAV format are discretised into time intervals through sampling. The sampling results in a column vector denoted as:

$$\bar{\mathbf{s}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{bmatrix}. \tag{3.1}$$

Each element in $\bar{\mathbf{s}}$ represents a sample, which corresponds to the measured acoustic values at a specific point in time. The index L in Equation (3.1) refers to the index of the last sampled point. The sampled dataset, $\bar{\mathbf{s}}$, contains samples, which are measured at temporally equispaced instants in time.

For appropriate feature extraction, portions representing either whale vocalisation or noise are extracted from $\bar{\mathbf{s}}$ and represented as $\bar{\mathbf{r}}$. An extracted portion, $\bar{\mathbf{r}}$, consists of k consecutive samples and is defined as:

$$\bar{\mathbf{r}} = \begin{bmatrix} x_{(i)} \\ x_{(i+1)} \\ x_{(i+2)} \\ \vdots \\ x_{(i+k-1)} \end{bmatrix} \in \bar{\mathbf{s}}, \tag{3.2}$$

where i is the starting index of $\bar{\mathbf{r}}$ within $\bar{\mathbf{s}}$ and i is determined through annotation. Each $\bar{\mathbf{r}}$ represents either a vocalisation of interest or an unknown vocalisation (which are all classified as noise in this study). It is important to note that every $\bar{\mathbf{r}}$ taken from $\mathbf{s}$ must be a complete representation of either a vocalisation of interest or noise. The data matrix,

$$\mathbf{X} \in \mathbb{R}^{n \times m} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \\ | & | & \cdots & | \end{bmatrix}, \tag{3.3}$$

is then generated from $\bar{\mathbf{r}}$, where n represents the number of samples in an observation and m represents the number of observations taken. Note that there are k samples in $\bar{\mathbf{r}}$; therefore, h numbers of matrices

$$\mathbf{X}_y, \qquad y = 1,\ 2,\ 3, \cdots,\ h, \tag{3.4}$$

are generated from the elements in $\bar{\mathbf{r}}$. Each $\mathbf{X}_y$ will be used for the analysis of the techniques to be developed for the extraction of features from PAM recordings in this study. The value of n in Equation (3.3) is determined per window size, denoted as $\mathcal{W}$. The window size is an important parameter for analysing acoustic signals since it influences the temporal or frequency resolution of the signal [152]. The window size represents a specific number of samples and a duration within the signal, and it is influenced by the fundamental frequency, intensity, and changes of the signal. Generally, the choice of window size to be selected will depend on the specific analysis to be carried out. A larger window size gives better frequency resolution, enabling easier differentiation between the frequency content of the signal. However, it sacrifices time by not accurately capturing the rapid temporal changes in the signal. On the other hand, a smaller window size allows for better time resolution, which means that the rapid changes in the waveform of the signal are accurately captured. However, it sacrifices frequency resolution. Whale calls are generally short in duration when compared to sounds like music or speech [13, 152]. Therefore, in this study, we will select window sizes, $\mathcal{W}$, of 32, 64, and 128 for conducting our analyses.

3.3 Eigendecomposition

Eigendecomposition is the process of decomposing a data matrix, $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \\ | & | & \cdots & | \end{bmatrix},$$

to eigenvectors Ψ,

$$\Psi = \begin{bmatrix} | & | & \cdots & | \\ \Psi_1 & \Psi_2 & \cdots & \Psi_m \\ | & | & \cdots & | \end{bmatrix}, \tag{3.5}$$

and corresponding eigenvalues, Υ,

$$\Upsilon = \begin{bmatrix} \Upsilon_1 & 0 & \dots & 0 \\ 0 & \Upsilon_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Upsilon_m \end{bmatrix}. \tag{3.6}$$

Each column in Equation (3.5) denotes an eigenvector in $\mathbf{X}$ and Equation (3.6) is a diagonal matrix whose diagonal elements give the corresponding eigenvalues of the eigenvectors. The above equations can be expressed as:

$$\mathbf{X}\Psi = \Psi\Upsilon,$$

$$\mathbf{X} = \Psi\Upsilon\Psi^{-1}. \tag{3.7}$$

The decomposition of $\mathbf{X}$ in terms of Ψ and its corresponding Υ provides useful information about the properties of the matrix, simplifying the analysis of complex problems and thus enhancing the understanding of the behaviour of a system. Eigendecomposition is significant in ML because it is involved in optimisation problems and has a wide range of applications such as spectral analysis, quantum mechanics, structural engineering, stability analysis, image compression, probability theory, and stochastic processes. The eigendecomposition process reduces the computational load required to analyse $\mathbf{X}$. The application of singular value decomposition (SVD) is a key component in several ED algorithms because of its effectiveness for matrix factorisation.

3.4 Singular Value Decomposition

Singular value decomposition (SVD) is one of the first steps in data reduction algorithms and can be likened to a data-driven version of FFT. SVD offers a systematic approach to obtaining a low-dimensional approximation of a high-dimensional dataset in terms of the dominant patterns that explain the data [153]. SVD provides a numerically stable matrix decomposition that can be used for a variety of purposes [30]. Furthermore, SVD exists for any matrix, as such, it can be "tailored" to the specifics of the data being analysed [30, 154, 155].

In defining SVD, $\mathbf{X} \in \mathbb{R}^{n \times m}$ is decomposed as [154]:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*, \tag{3.8}$$

where $*$ represents the transpose, $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{m \times m}$ are unitary matrices with orthonormal columns. $\mathbf{\Sigma} \in \mathbb{R}^{n \times m}$ is a diagonal matrix with non-negative elements on the diagonal and zeros off the diagonal. The columns of $\mathbf{U}$ are the left singular vectors of $\mathbf{X}$, representing the information about the column space of $\mathbf{X}$. The vectors are hierarchically positioned, that is, $\mathbf{u}_1$ is more important than $\mathbf{u}_2$ and so on. The columns of $\mathbf{V}$ are the right singular vectors of $\mathbf{X}$, signifying the information about the row space of $\mathbf{X}$. The diagonal elements of $\mathbf{\Sigma} \in \mathbb{C}^{n \times m}$ are singular values, indicating the order of importance of $\mathbf{U}$ and $\mathbf{V}^*$. These elements are ordered in descending order, that is, $\Sigma_1 \geq \Sigma_2 \geq \cdots \geq \Sigma_m \geq 0$. Therefore, SVD in Equation (3.8), is the factorisation of $\mathbf{X}$ into a number of intrinsic components, each of which has a particular meaning in different applications.

When $n \geq m$, $\mathbf{\Sigma}$ contains no more than m non-zero elements; only the first m singular values are retained, and the extra rows of zeros in $\mathbf{\Sigma}$ are discarded. That is, $\mathbf{\Sigma}$ becomes an $m \times m$ matrix. This implies that only the m columns of $\mathbf{U}$ will be computed. This concept is referred to as *economy* SVD in the literature [30]. The *economy* SVD only computes the necessary number of singular values and corresponding singular vectors needed to exactly represent $\mathbf{X}$. Thus, the *economy* SVD

of $\mathbf{X}$ is defined as:

$$\mathbf{X} = \underbrace{\begin{bmatrix} | & | & \cdots & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_m \\ | & | & \cdots & | \end{bmatrix}}_{\mathbf{U}^{n\times m}} \underbrace{\begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_m \end{bmatrix}}_{\mathbf{\Sigma}^{m\times m}} \underbrace{\begin{bmatrix} - & - & - & \mathbf{v}_1^* & - & - & - \\ - & - & - & \mathbf{v}_2^* & - & - & - \\ & & & \vdots & & & \\ - & - & - & \mathbf{v}_m^* & - & - & - \end{bmatrix}}_{\mathbf{V}^{m\times m}}. \tag{3.9}$$

This property makes SVD relevant in dimensionality reduction techniques because it provides a hierarchy of low-rank representations of $\mathbf{X}$, thereby improving the execution time and reducing the storage requirements without compromising the accuracy of the decomposition of $\mathbf{X}$. The SVD is the key component of the proposed PCA and DMD-based FE techniques developed in this study.

3.5 Principal Component Analysis

Principal component analysis (PCA) is the foundation of dimensionality reduction techniques and remains one of the techniques suitable for extracting inherent features from datasets [30, 32]. PCA has been around for a long time since its invention by Karl Pearson in 1901 [156]; thus, it is an old technique, well established with several theories on how to implement it [30, 31]. The PCA approach described in this study follows a statistical interpretation of the SVD. The approach provides a hierarchical coordinate system to represent the statistical variation in the dataset.

PCA decomposes high-dimensional data, $\mathbf{X}$, into a lower-dimensional space containing the most statistically illustrative statistical variations of the original data. In other words, PCA reduces high-dimensional data to a lower number of dimensions while retaining the important information that explains the original data. These dimensions are hierarchically ordered from the most to the least important statistical variation in the datasets. The variables within these dimensions are referred to as principal components (PCs).

The PCA scheme performs an orthogonal transformation to convert a set of correlated variables into a new set of linearly uncorrelated variables called PCs. The scheme ensures that the first PC has the most variance, or variability, in the dataset; the second PC has the next most variance; and so on. Generally, PCA ensures that the variability in the dataset is captured within the first few PCs.

Given a dataset, $\mathbf{X}$, whose observations are arranged in column-vector format as:

$$\mathbf{X} \in \mathbb{R}^{n \times m} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \\ | & | & \cdots & | \end{bmatrix}, \tag{3.10}$$

where n represents the number of samples in an observation and m represents the number of observations taken. The computation of the PCs using the SVD approach involves the following steps:

Step 1 : Data Normalisation. The normalisation step ensures that all variables contribute equally to the analysis and facilitates the description of PCs along the directions of maximum variance. By bringing variables to the centre of the distribution, mean centring reduces bias and prevents variables with large mean values from dominating the analysis, thus promoting a more balanced representation of the data.

This step is important because PCA is sensitive to the mean value in the initial variables. For example, it is not uncommon for some variables to have larger values than others, potentially biasing the analysis. By subtracting the mean separately from each column, the influence of variables with large mean values is appropriately mitigated. Therefore, data normalisation is performed to ensure that the computed PCs accurately describe the underlying variance in the dataset rather than merely capturing the variables with big values.

Normalisation is done by computing the mean, $\bar{\mathbf{x}}$, of Equation (3.10) as defined by:

$$\bar{\mathbf{x}} = \frac{1}{n}\left(\sum_{i=1}^{n} x_{ij}\right), \tag{3.11}$$

where x_{ij} represents the elements at the i-th row and j-th column. A mean matrix, $\bar{\mathbf{X}}$, is computed by multiplying an all-ones column vector of size n with $\bar{\mathbf{x}}$ as defined by:

$$\bar{\mathbf{X}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \bar{\mathbf{x}}. \tag{3.12}$$

The mean matrix, $\bar{\mathbf{X}}$ is then subtracted from the data matrix, $\mathbf{X}$ as:

$$\bar{\mathbf{U}} = \mathbf{X} - \bar{\mathbf{X}}, \tag{3.13}$$

where $\bar{\mathbf{U}}$ is the mean centred data (normalised data) of the data matrix, $\mathbf{X}$ and all the variables in $\bar{\mathbf{U}}$ are on the same scale.

Step 2 : Computing the PCs with SVD: SVD provides a more numerically stable and efficient way of computing the PCs. Essentially, SVD saves us the time for some other 'sorting' needed to interpret the PC order of hierarchy. Therefore, the matrices of the PCs and their corresponding coefficients are computed by performing the SVD operation on $\bar{\mathbf{U}}$ as defined by:

$$\bar{\mathbf{U}} \in \mathbb{R}^{n \times m} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{*}, \tag{3.14}$$

In Equation (3.14), the columns of the right singular matrix, $\mathbf{V}$, represent the PCs, which indicate the direction of the axes with the most variance in the dataset, while the columns of the singular values, $\boldsymbol{\Sigma}$, are the respective

coefficients attached to the PC, indicating the amount of variance in each PC. The coefficients are sorted in descending order.

By default, in Equation (3.14), the number of PCs and the number of the coefficients are equal to the number of the original observations in $\mathbf{X}$. In other words, for an m-dimensional dataset, there will be m-eigenvectors and m-eigenvalues. Consequently, the PCs are represented in the form of an $m \times m$ matrix as:

$$\tilde{\mathbf{Pc}} = \begin{bmatrix} \tilde{P}c_{1,1} & \tilde{P}c_{1,2} & \dots & \tilde{P}c_{1,m} \\ \tilde{P}c_{2,1} & \tilde{P}c_{2,2} & \dots & \tilde{P}c_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{P}c_{m,1} & \tilde{P}c_{m,2} & \dots & \tilde{P}c_{m,m} \end{bmatrix}. \tag{3.15}$$

The computed PCs from PCA enable the interpretability of datasets in a lower-dimensional space while preserving their information contents. While some authors would argue that the PCA is simply a 'dimensionality reduction' tool for 'big data' before any analysis can be carried out, it is, at its core, an adaptable technique for extracting structure from high-dimensional datasets [32]. PCA has been used in different fields such as image processing, disease modelling, and denoising, among others; however, its use is not one-size-fits-all [31]. In other words, the characteristics of the data, the formats of the data, and the goals to be achieved, among others, will determine how the PCs will be utilised. The PCs computed from PCA do not rely on pre-defined basis functions; rather, they rely on the inherent nature, structure, and characteristics of the dataset. This makes the PCA an adaptive technique for data analysis. In the literature, there are many modifications and adaptations to the use of PCA for the analysis of different data types, subject to the specifics of the goals to be achieved [30–32]. Consequently, several variants of PCA exist across different disciplines. In this study, our focus is to deploy the computed PCs from PCA as part of the process of constructing feature vectors for HMM for the detection of whale vocalisations.

3.6 Proposed PC Feature Vectors for HMM

In this study, we are deploying the PCs computed from PCA as part of the process of constructing feature vectors for HMMs. The PCs-based feature vectors proposed are developed by projecting $\mathbf{X}$ using a selected number of $\tilde{\mathbf{Pc}}$. Therefore, a certain number of $\tilde{\mathbf{Pc}}$, denoted as p, are selected to project the original data, $\mathbf{X}$, into an $n \times p$ low-dimensional space represented as $\tilde{\mathbf{G}}$:

$$\tilde{\mathbf{G}} = \bar{\mathbf{U}}\tilde{\mathbf{Pc}},$$

where $\bar{\mathbf{U}} \in \mathbb{R}^{n \times m}$ and $\tilde{\mathbf{Pc}} \in \mathbb{R}^{m \times p}$. Thus, $\tilde{\mathbf{G}} \in \mathbb{R}^{n \times p}$ low-dimensional data is achieved. Consequently, the low-dimensional data, $\tilde{\mathbf{G}}$ is represented in the form of an $n \times p$ matrix as:

$$\tilde{\mathbf{G}} = \begin{bmatrix} \tilde{G}_{1,1} & \tilde{G}_{1,2} & \dots & \tilde{G}_{1,p} \\ \tilde{G}_{2,1} & \tilde{G}_{2,2} & \dots & \tilde{G}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{G}_{n,1} & \tilde{G}_{n,2} & \dots & \tilde{G}_{n,p} \end{bmatrix}. \tag{3.16}$$

Equation (3.16) represents the original data $\mathbf{X} \in \mathbb{R}^{n \times m}$ in $n \times p$ $\tilde{\mathbf{G}} \in \mathbb{R}^{n \times p}$. The matrix $\tilde{\mathbf{G}}$ will be used to construct a suitable matrix of feature vectors to be used with the HMM for the detection of whale vocalisations.

The $\tilde{\mathbf{Pc}}$ is computed for each $\mathbf{X}_y$ in Equation (3.4) as enumerated from **Step 1** to **Step 2**. A certain number of p are selected to project the computed $\tilde{\mathbf{Pc}}$ as defined in Equation (3.16). Thereafter, the individual feature vector, $\mathcal{F}_y$, is obtained by

computing the column-wise mean of Equation (3.16) obtained for each $\mathbf{X}_y$ as:

$$\begin{aligned}
\delta_{y,1} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{G}_{i1} \\
\delta_{y,2} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{G}_{i2} \\
&\vdots \\
\delta_{y,p} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{G}_{ip} \\
\mathcal{F}_y &= \begin{bmatrix}\delta_{y,1} & \delta_{y,2} & \cdots & \delta_{y,p}\end{bmatrix},
\end{aligned} \tag{3.17}$$

where $y = 1, 2, \cdots, h$. Thus, given $\bar{\mathbf{r}}$, the feature vectors derived for each $\mathbf{X}_y$ are restructured to form the feature matrix, $\boldsymbol{F}$, for the HMM as defined by:

$$\boldsymbol{F} = \begin{bmatrix}\delta_{1,1} & \delta_{1,2} & \dots & \delta_{1,p} \\ \delta_{2,1} & \delta_{2,2} & \dots & \delta_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{h,1} & \delta_{h,2} & \dots & \delta_{h,p}\end{bmatrix}. \tag{3.18}$$

Each $\boldsymbol{F}$ represents the feature vectors of a portion of the sampled datasets, $\bar{\mathbf{s}}$, which can either be vocalisation of interest or noise, as stated earlier. The number of PCs selected for $\tilde{\mathbf{G}}$ has a direct impact on the computational load and performance of the HMM. Thus, different values of p will be simulated per $\mathcal{W}$, during the experiment to determine the optimal p. A summary of the PCA-FE process described in this work is presented in Algorithm 2.

Algorithm 2: Feature Extraction with PCA

1: **Input:** $[\bar{\mathbf{s}}, \mathcal{W}, p]$

2: **Output:** $\boldsymbol{F}$

3: Decide the portion of $\bar{\mathbf{r}}$ to be taken from $\bar{\mathbf{s}}$ by annotation

4: Transform $\bar{\mathbf{r}}$ to h numbers of $\mathbf{X}$ with respect to $\mathcal{W}$ and m

5: Calculate the $\tilde{\mathbf{Pc}}$ from **Step 1** to **Step 2**

6: Calculate $\tilde{\mathbf{G}}$ from Equation (3.16)

7: Compute $\mathcal{F}_y$ from $\tilde{\mathbf{G}}$ as described in Equation (3.17)

8: **Do** horizontal concatenation for each $\mathcal{F}_y$ to obtain $\boldsymbol{F}$ for respective $\bar{\mathbf{r}}$ as described in Equation (3.18)

3.7 Numerical Example of PC Feature Vectors

Figure 3.2 shows an example of the waveform of a whale vocalisation.

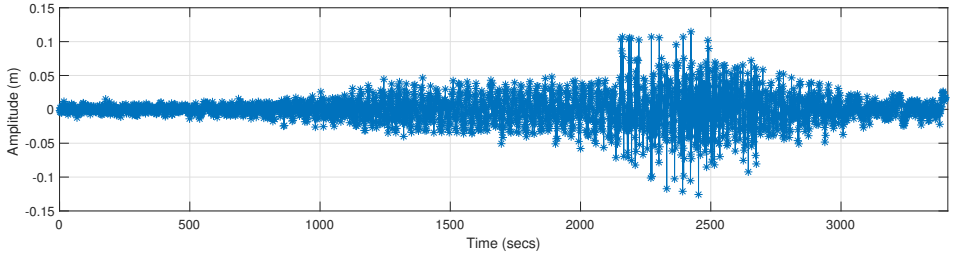


FIGURE 3.2: Waveform of a whale vocalisation.

Given that a portion of the whale vocalisation is to be analysed as shown in Figure 3.3(a),

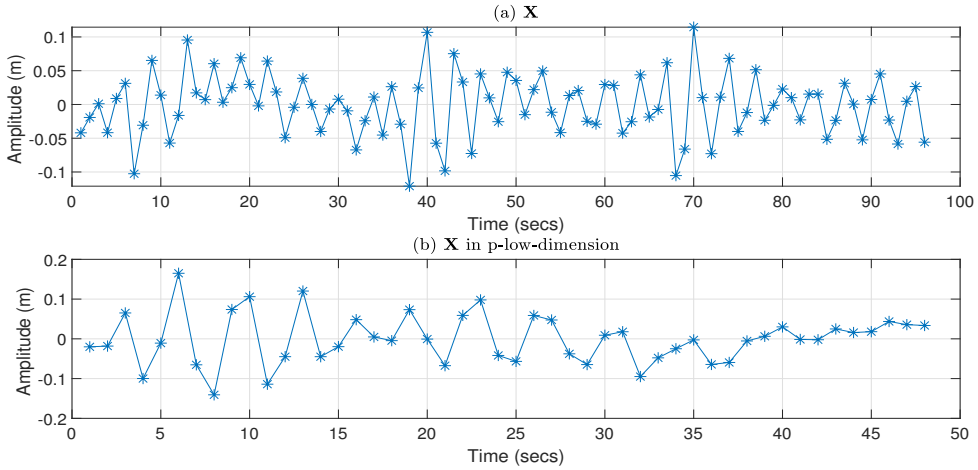


FIGURE 3.3: (a) Original $\mathbf{X}$ signal in high-dimensional space, (b) The projected $\mathbf{X}$ signal in low-dimensional space.

and its data matrix is represented as:

$$\mathbf{X} = \begin{bmatrix} -0.0420 & 0.0033 & -0.0241 & 0.0477 & -0.0186 & 0.0101 \\ -0.0192 & 0.0248 & 0.0110 & 0.0356 & -0.0074 & -0.0228 \\ 0.0011 & 0.0691 & -0.0453 & -0.0151 & 0.0618 & 0.0154 \\ -0.0418 & 0.0297 & 0.0263 & 0.0220 & -0.1053 & 0.0154 \\ 0.0089 & -0.0021 & -0.0292 & 0.0496 & -0.0661 & -0.0516 \\ 0.0313 & 0.0642 & -0.1210 & -0.0115 & 0.1145 & -0.0235 \\ -0.1026 & 0.0185 & 0.0245 & -0.0417 & 0.0102 & 0.0314 \\ -0.0308 & -0.0492 & 0.1068 & 0.0132 & -0.0729 & 0.0005 \\ 0.0652 & -0.0042 & -0.0573 & 0.0206 & 0.0108 & -0.0524 \\ 0.0139 & 0.0388 & -0.0983 & -0.0250 & 0.0683 & 0.0073 \\ -0.0572 & 0.0206 & 0.0752 & -0.0292 & -0.0400 & 0.0452 \\ -0.0162 & -0.0400 & 0.0333 & 0.0295 & -0.0120 & -0.0231 \\ 0.0955 & -0.0067 & -0.0726 & 0.0282 & 0.0515 & -0.0586 \\ 0.0172 & 0.0078 & 0.0452 & -0.0424 & -0.0235 & 0.0047 \\ 0.0072 & -0.0094 & 0.0098 & -0.0256 & -0.0014 & 0.0265 \\ 0.0604 & -0.0673 & -0.0255 & 0.0439 & 0.0228 & -0.0558 \end{bmatrix}. \quad (3.19)$$

The process of computing the PCs is as follows. Equation (3.19) is normalised using Equations (3.11)–(3.13) to derive the mean centred data $\bar{\mathbf{U}}$ as defined by:

$$\bar{\mathbf{U}} = \begin{bmatrix} -0.0414 & -0.0015 & \dots & 0.0183 \\ -0.0187 & 0.0200 & \dots & -0.0146 \\ \vdots & \vdots & \ddots & \vdots \\ 0.0610 & -0.0721 & \dots & -0.0476 \end{bmatrix}. \tag{3.20}$$

As earlier stated, SVD provides a more numerically stable and efficient approach for the computation of the PC. Therefore, the SVD operation is carried out on $\bar{\mathbf{U}}$ to obtain the PCs and the corresponding coefficients using the *economy* SVD as defined by:

$$\bar{\mathbf{U}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*,$$

where,

$$\mathbf{U} \in \mathbb{R}^{16\times 6} = \begin{bmatrix} -0.0587 & 0.0218 & -0.3658 & -0.3467 & 0.2472 & -0.4412 \\ -0.0531 & -0.0204 & -0.1906 & -0.1251 & -0.5933 & -0.0851 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -0.0574 & 0.0810 & 0.2732 & 0.1723 & 0.3796 & -0.2551 \\ 0.1409 & -0.4297 & 0.2572 & -0.3109 & 0.1772 & -0.0468 \end{bmatrix}, \tag{3.21}$$

is the left singular vectors,

$$\boldsymbol{\Sigma} \in \mathbb{R}^{6\times 6} = \begin{bmatrix} 0.3424 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2208 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1307 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0945 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0569 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0425 \end{bmatrix}, \tag{3.22}$$

is the singular values that represent the respective coefficients of the PCs in descending order and,

$$\mathbf{V}^* \in \mathbb{R}^{6\times 6} = \begin{bmatrix} 0.4140 & -0.4951 & 0.3423 & 0.6043 & -0.0346 & -0.3161 \\ 0.1598 & 0.4823 & -0.4421 & 0.4635 & -0.5707 & -0.0764 \\ -0.6695 & -0.0668 & 0.4641 & 0.0338 & -0.5568 & -0.1438 \\ 0.0208 & -0.4447 & -0.4865 & -0.4165 & -0.2548 & -0.5716 \\ 0.5608 & 0.3288 & 0.4769 & -0.4826 & -0.3060 & -0.1533 \\ -0.1999 & 0.4604 & 0.0881 & 0.1116 & 0.4522 & -0.7234 \end{bmatrix}, \tag{3.23}$$

is the right singular vectors representing matrix $\tilde{\mathbf{P}}\mathbf{c}$ of Equation (3.15). In other words, the columns of Equation (3.23) represent the PCs. Note that the PCs are hierarchically arranged according to the ordered structure of Equation (3.22), that is, the largest coefficient in Equation (3.22) (0.3424) indicates the position of the first PC (column one in Equation (3.23)). Figure 3.4 provides insights into the amount

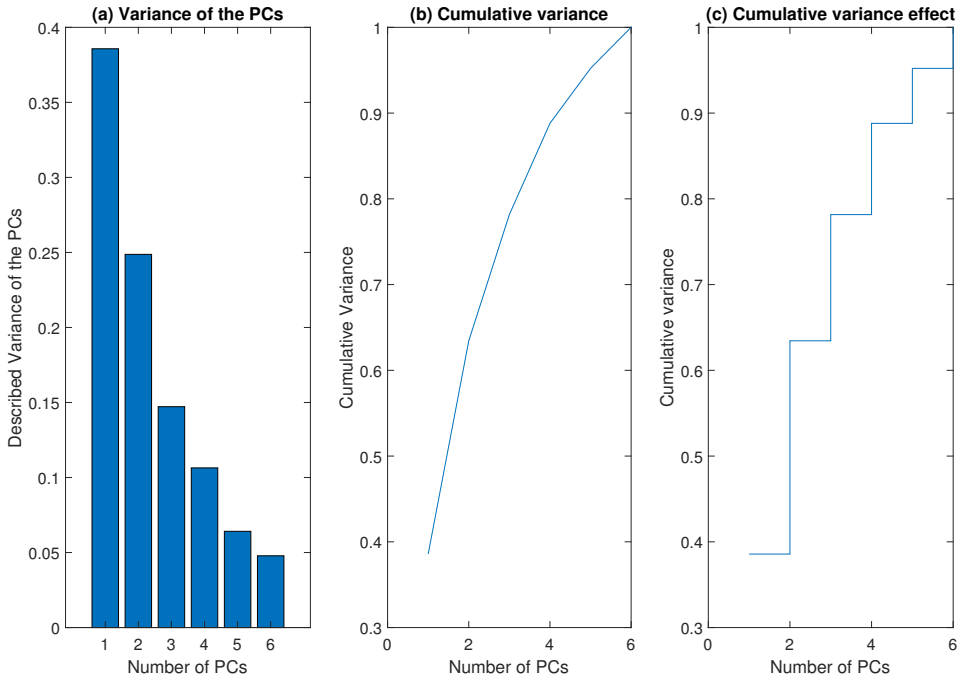


FIGURE 3.4: (a) Proportion of information captured by each PC, (b) cumulative variance described by a combination of PCs, and (c) incremental contribution of additional PC to the cumulative variance.

of information captured by each PC and the cumulative information captured by a combination of PCs.

Figure 3.4(a) shows the proportion of the described variance by each PC. The bar plot shows that the first PC accounts for 38.6% of the information in the original data, the second PC accounts for 24.9%, the third PC accounts for 14.7% and so on. It is evident that the first few PCs capture the majority of the variance, while the later PCs contribute less to the overall information. For example, while the first and second PCs account for 38.6% and 24.9% of the information, the fifth and sixth PCs account for only 6.4% and 4.8% respectively. Consequently, the dimension of the original data can be reduced while retaining the most important information since the top PCs account for the majority of the variance in the data. Hence, the reduction in dimension simplifies the data presentation, making subsequent analysis and modelling more efficient. The bar plot expounds on the importance of each PC in capturing the variance in the data.

The cumulative variance plot in Figure 3.4(b) shows the accrued ratio of variance described by a combination of PCs. The cumulative sum represents the proportion of the total variance described by the PC. The percentage of the cumulatively described variance is computed by dividing the cumulative sum by the sum of all eigenvalues. This information is important as it gives guidelines on the optimal number of PCs to be retained for analysis.

Furthermore, the stair plot in Figure 3.4(c) gives more illustration on the cumulative variance. It explains how much of the total variance is described by each addition of a PC. Each step of the stair plot indicates a step-like progression showing the incremental contribution of an additional PC to the cumulative variance. This helps in assessing the trade-off between dimensionality reduction and information retention. For instance, a combination of the first four PCs accounts for 89% of the information in the original data.

Therefore, mean centred data can be projected to low-dimensional space by selecting p-numbers of PCs to obtained the transformed data in a new feature space. For

instance, if $p = 4$, $\bar{\mathbf{U}} \in \mathbb{R}^{n \times m}$ is projected by the first-four PCs to low-dimension data as defined by:

$$\begin{aligned} \tilde{\mathbf{G}} \in \mathbb{R}^{n \times p} &= \bar{\mathbf{U}}\tilde{\mathbf{Pc}}, \\ &= \begin{bmatrix} -0.0201 & 0.0048 & -0.0478 & -0.0328 \\ -0.0182 & -0.0045 & -0.0249 & -0.0118 \\ \vdots & \vdots & \vdots & \vdots \\ -0.0196 & 0.0179 & 0.0357 & 0.0163 \\ 0.0482 & -0.0949 & 0.0336 & -0.0294 \end{bmatrix}. \end{aligned} \tag{3.24}$$

The low-dimension data $\tilde{\mathbf{G}} \in \mathbb{R}^{n \times p}$ is represented in Figure 3.3(b).

3.8 Dynamic Mode Decomposition

Dynamic mode decomposition (DMD) was introduced by Schmid [145] in the field of fluid mechanics to analyse the complex evolution of fluid flows. The application of DMD has been extended to other fields such as video processing [157], control systems [158], disease modelling [159], and load forecasting [160]. The increasing interest in DMD is due to the fact that it is entirely a data-driven algorithm. The DMD algorithm is analogous to the empirical approximation of the Koopman operator, an important concept in dynamical system theory [145, 161]. DMD is capable of accurately decomposing non-stationary data into spatiotemporal coherent patterns that will distinguish prominent features of the data [33]. It can extract dynamic information from non-stationary datasets such as whale vocalisations. The extracted information structurally describes the inherent physical properties that dominate the entire dataset.

The DMD algorithm basically depends on assembling the observations from a dynamic system, which are denoted as snapshots,

$$\mathbf{x}_i \qquad i = 1,\ 2,\ 3, \cdots,\ m.$$

The points within each snapshot are uniformly time-spaced. Therefore, the DMD algorithm collects the observations in a data, arranges them per snapshots in the form of:

$$\mathbf{X} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \\ | & | & \cdots & | \end{bmatrix}, \tag{3.25}$$

and organises them into tall and skinny matrices $\mathbf{X}_a$ and $\mathbf{X}_b$, also known as snapshot matrices:

$$\mathbf{X}_a = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{m-1} \\ | & | & \cdots & | \end{bmatrix} \in \mathbf{X}, \tag{3.26}$$

$$\mathbf{X}_b = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{x}_2 & \mathbf{x}_3 & \cdots & \mathbf{x}_m \\ | & | & \cdots & | \end{bmatrix} \in \mathbf{X}. \tag{3.27}$$

The columns of $\mathbf{X}_a$ and $\mathbf{X}_b$ are the snapshots of the reshaped data $\mathbf{X}$, and their column length is $m-1$. The objective of DMD is to understand the dynamic behaviour or inherent patterns in the dataset by efficiently capturing the changes between successive snapshots. The DMD achieves this objective by analysing pairs of successive snapshots within the dataset, each pair separated by a Δt time interval. Therefore, the column length of $\mathbf{X}_a$ and $\mathbf{X}_b$ is set to $m-1$. $\mathbf{X}_a$ starts from the first column to the $m-1$ column of $\mathbf{X}$, while $\mathbf{X}_b$ starts from the second column to the last column of $\mathbf{X}$.

Thereafter, DMD finds a best-fit linear operator, $\mathbf{Z}$, that advances the leading eigenvalues and eigenvectors of the data [145, 154]. The aim of DMD is to efficiently compute the leading eigendecomposition of the best-fit linear operator $\mathbf{Z}$. The local linear approximation can be written with respect to the data matrices given in Equations (3.26) and (3.27) as defined by:

$$\mathbf{X}_b \approx \mathbf{Z}\mathbf{X}_a \quad \Longrightarrow \mathbf{Z} = \mathbf{X}_b\mathbf{X}_a^{\dagger}, \tag{3.28}$$

where † is the Moore-Penrose pseudo-inverse. The eigenvectors, Ψ, and their corresponding eigenvalues, Υ, derived from the eigendecomposition of $\mathbf{Z}$, are used to compute the spatiotemporal modes, $\tilde{\boldsymbol{M}}$. Most of the time, the data matrices (Equations (3.26) and (3.27)) are high-dimensional, which makes them intractable for computing the eigendecomposition of $\mathbf{Z}$ directly.

Instead, DMD leverages the concept of a low-rank approximation to address this problem by projecting the data onto a low-dimensional subspace [145]. This subspace is meant to capture the most relevant and dominant modes of the data's behaviour. The low-rank approximation effectively captures the underlying structure of the data, allowing the extraction of dominant modes and their associated temporal behaviour. A low-dimensional operator, $\tilde{\mathbf{Z}}$, is introduced in this subspace. The $\tilde{\mathbf{Z}}$ enables the computation of the leading eigendecomposition of $\mathbf{Z}$ without explicitly computing $\mathbf{Z}$. Thus, Equation (3.28) can be represented as:

$$\mathbf{X}_b \approx \tilde{\mathbf{Z}}\mathbf{X}_a \quad \Longrightarrow \tilde{\mathbf{Z}} = \mathbf{X}_b\mathbf{X}_a^{\dagger}. \tag{3.29}$$

The DMD spatiotemporal modes, denoted as $\tilde{\boldsymbol{M}}$, are reconstructed using the eigenvectors of $\tilde{\mathbf{Z}}$ of the reduced data and the shifted snapshot matrix $\mathbf{X}_b$. Each of these modes is associated with an individual eigenvalue that gives information about the frequency of oscillation and the growth or decay rate. The projected future solution for the data can be reconstructed through the derived low-rank approximation. The approximate solution, $\bar{\mathbf{F}}(t)$, for all times in the future is derived as defined by:

$$\bar{\mathbf{F}}(t) = \tilde{\boldsymbol{M}}\exp(\bar{\Upsilon}\ t_F)\mathbf{b}, \tag{3.30}$$

where $\tilde{\boldsymbol{M}}$ is the matrix of the DMD modes, $\bar{\Upsilon}$ is a diagonal matrix of the eigenvalues, t_F is the time of future forecast, and $\mathbf{b}$ is a vector of the coefficients of the initial amplitudes of each dynamic mode. It is important to note that Ψ and Υ are complex valued; therefore, the $\tilde{\boldsymbol{M}}$ are also complex values, and that in general, $\bar{\mathbf{F}}(t)$ has non-zero imaginary components; therefore, if the raw data $\mathbf{X}$ is strictly real values, only the real component of $\bar{\mathbf{F}}(t)$ may be considered for further analysis [159]. In the

literature, the DMD algorithm is primarily used for diagnostics, state estimation and future-state prediction, and control, often centred around Equation (3.30).

The application of DMD has gained traction beyond the fluid mechanics field, where it was introduced. Its application has since been extended to other areas of application. From the extensive review of the literature, we broadly categorise DMD applications into three tasks: First, as a diagnostic technique, the DMD enables the data-driven discovery of fundamental low-rank structures in complex systems. This allows for the physical interpretation of results in terms of the spatial structures and their associated temporal reactions, therefore revealing the underlying patterns, behaviours, and faults of the system. Examples of this application include mechanical, fluid, and disease modelling [33, 145, 159, 161]. Second, as a state estimation and future-state forecast technique, where the dominant spatiotemporal structures in the data are utilised to construct dynamical models of the underlying processes that are observed, this is applied in areas such as electrical load forecast, weather forecast, and finance market analysis [159, 160, 162, 163]. Third, as a control tool, by giving insights into the underlying dynamics of systems, this is applied in system tracking and system stabilisation, among other control objectives [157, 158, 161, 164]. In this study, we explore the potential of DMD as a technique for FE. The data-driven framework of the DMD, which eliminates the need to understand or have prior knowledge of the underlying equations governing the data being modelled, is one of the motivations for the exploration of the algorithm as a feature extraction technique in our research.

3.9 Proposed DMD Feature Vectors for HMM

According to the literature, the DMD algorithm is generally used on various data types for diagnostics, state estimation and future-state prediction, and control, often centred around Equation (3.30). The algorithm has not been used exclusively as a technique for FE. In this study, we present a novel application of DMD as a technique

for FE. The development of the proposed DMD-based feature vector is outlined as follows:

Given a data matrix, $\mathbf{X}$, the snapshot matrices $\mathbf{X}_a$ and $\mathbf{X}_b$ are derived as defined in Equations (3.26) and (3.27). The *economy* SVD of $\mathbf{X}_a$ is computed as defined:

$$\mathbf{X}_a = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*, \tag{3.31}$$

where $*$ represents the transpose, $\mathbf{U} \in \mathbb{R}^{n\times p}$, $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, $\mathbf{V} \in \mathbb{R}^{m\times p}$, and p is the rank of the reduced SVD approximation to $\mathbf{X}_a$.

The matrix $\tilde{\mathbf{Z}}$ is computed using the pseudo-inverse of $\mathbf{X}_a$ obtained through the SVD as defined:

$$\begin{aligned} \mathbf{X}_b &\approx \mathbf{X}_a\tilde{\mathbf{Z}} \\ \mathbf{X}_b &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\tilde{\mathbf{Z}} \\ \tilde{\mathbf{Z}} &= \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{X}_b. \end{aligned} \tag{3.32}$$

The eigendecomposition of $\tilde{\mathbf{Z}}$ is then computed:

$$\tilde{\mathbf{Z}}\Psi = \Psi\Upsilon. \tag{3.33}$$

The DMD modes, $\tilde{\boldsymbol{M}}$ are computed using:

$$\tilde{\boldsymbol{M}} = \mathbf{X}_b\mathbf{V}\mathbf{\Sigma}^{-1}\Psi,$$

where $\tilde{\boldsymbol{M}} \in^{n\times p}$. Therefore, the DMD modes, $\tilde{\boldsymbol{M}}$ can be expressed in matrix form as:

$$\tilde{\boldsymbol{M}} = \begin{bmatrix} \tilde{M}_{1,1} & \tilde{M}_{1,2} & \dots & \tilde{M}_{1,p} \\ \tilde{M}_{2,1} & \tilde{M}_{2,2} & \dots & \tilde{M}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{M}_{n,1} & \tilde{M}_{n,2} & \dots & \tilde{M}_{n,p} \end{bmatrix}. \tag{3.34}$$

The number of samples in an observation (snapshot), $\mathcal{W}$, and the rank, p, determine the dimension of $\tilde{\boldsymbol{M}}$. The p enables the low-rank truncation of the data, specifically to aid the quick convergence of the dominant modes in the data. Therefore, the values of p will be varied per $\mathcal{W}$ to analyse the rate of convergence of the modes. The $\tilde{\boldsymbol{M}}$ derived for each $\mathbf{X}_y$ in Equation (3.4) is used to compute individual feature vector, $\mathcal{F}_y$ as:

$$\begin{aligned}
\delta_{y,1} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{M}_{i1} \\
\delta_{y,2} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{M}_{i2} \\
&\vdots \\
\delta_{y,p} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{M}_{ip} \\
\mathcal{F}_y &= \begin{bmatrix}\delta_{y,1} & \delta_{y,2} & \cdots & \delta_{y,p}\end{bmatrix},
\end{aligned} \tag{3.35}$$

where $y = 1, 2, \cdots, h$. Thus, given $\bar{\mathbf{r}}$, the feature vectors derived for each $\mathbf{X}_y$ are restructured to form the feature matrix, $\boldsymbol{F}$, for the HMM as defined by:

$$\boldsymbol{F} = \begin{bmatrix}
\delta_{1,1} & \delta_{1,2} & \dots & \delta_{1,p} \\
\delta_{2,1} & \delta_{2,2} & \dots & \delta_{2,p} \\
\vdots & \vdots & \ddots & \vdots \\
\delta_{h,1} & \delta_{h,2} & \dots & \delta_{h,p}
\end{bmatrix}. \tag{3.36}$$

Each $\boldsymbol{F}$ represents the feature vectors of a portion of the sampled datasets, $\bar{\mathbf{s}}$, which can either be vocalisation of interest or noise, as stated earlier. The value of p has a direct impact on the computational load and performance of the HMM. Thus, different values of p will be simulated per $\mathcal{W}$, during the experimentation phase to determine the performance of the model until an optimal p is achieved, where the best performance is achieved for the model. A summary of the DMD FE process described in this study is presented in Algorithm 3.

Algorithm 3: Feature Extraction with DMD

1: **Input:** $[\bar{\mathbf{s}}, \mathcal{W}, p]$
2: **Output:** $\boldsymbol{F}$
3: Decide the portion of $\bar{\mathbf{r}}$ to be taken from $\bar{\mathbf{s}}$ by annotation
4: Transform $\bar{\mathbf{r}}$ to h numbers of $\mathbf{X}$ with respect to $\mathcal{W}$ and p
5: Obtain $\mathbf{X}_a$ and $\mathbf{X}_b$ from respective $\mathbf{X}$ as described in Equation (3.26) and Equation (3.27)
6: Calculate $\tilde{\boldsymbol{M}}$ from Equations (3.26) and (3.27) as enumerated from Equations (3.31)–(3.34)
7: Compute $\mathcal{F}_y$ from $\tilde{\boldsymbol{M}}$ as described in Equation (3.35)
8: **Do** horizontal concatenation for each $\mathcal{F}_y$ to obtain $\boldsymbol{F}$ for respective $\bar{\mathbf{r}}$ as described in Equation (3.36)

The DMD is demonstrated on the data of the numerical example presented in Section 3.6. We performed different rank (p) truncations to gain insights into how the features are represented at different mode values. Figure 3.5 shows the original data, $\mathbf{X}$, and the DMD modes reconstruction at different p values. The visual representation illustrates that DMD modes capture the underlying patterns of the original data. The accuracy of the low-rank reconstruction varies per truncation value, p. This indicates that there will be a trade-off between the selection of p and the DMD's ability to capture the underlying dominant features of the data. Therefore, we will experiment with different p values with the HMM and evaluate the results.

3.10 Conclusion

In this chapter, PCA and DMD are introduced as FE techniques. The PCs computed from PCA and the modes computed from DMD are uniquely designed to extract features from PAM datasets. The respectively developed feature matrices would be fed into HMM for the detection of whale vocalisations. While PCA finds a low-dimensional linear subspace that accounts for the majority of the variation in a dataset, it may not effectively create a low-dimensional non-linear subspace if it

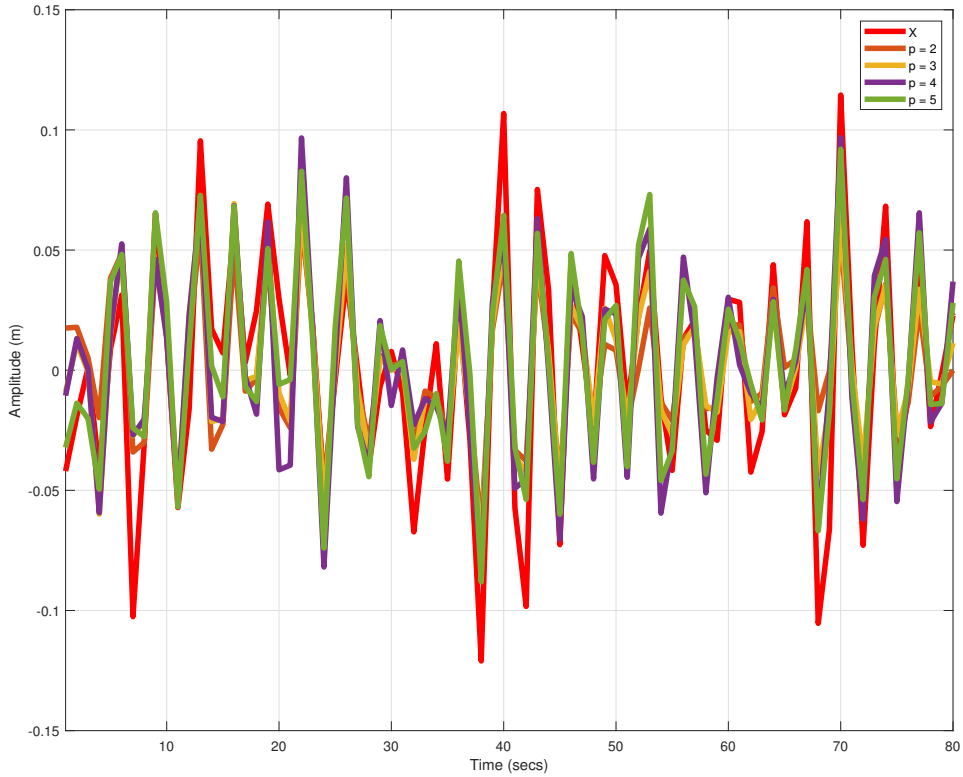


FIGURE 3.5: Comparison of the original data, $\mathbf{X}$, and the DMD modes at different p low-rank truncation

exists in a dataset. Similarly, for DMD, the presence of non-linear characteristics in the datasets may affect the optimal representation of the spatiotemporal patterns in the underlying system. However, the whale vocalisations sometimes exhibit non-linear characteristics (Section 2.2). Therefore, enhanced FE techniques are proposed to further deepen the use of PCA and DMD for FE. The enhanced FE techniques are achieved through the *kernelisation* of PCA and DMD, which is the focus of Chapter 4. The introduction of kernel PCA and kernel DMD in Chapter 4 explores the potential of enhancing these techniques to better align with the non-linear nature of certain datasets. The enhanced FE techniques seek to overcome the limitations posed by linear subspace assumptions, thus allowing for a more comprehensive extraction of features from complex datasets.

Chapter 4

Enhanced ED Feature Extraction Techniques

4.1 Introduction

This chapter presents an enhanced feature extraction (FE) process through the introduction of the kernel method to principal component analysis (PCA) and dynamic mode decomposition (DMD), which were introduced in Chapter 3 as FE techniques. The *kernelisation* of an algorithm is a process of improving the efficiency of the algorithm by reducing the computational process through the replacement of the inputs with a smaller input, called a "kernel". The incorporation of the kernel in each of the ED-based FE techniques offers an opportunity to enhance their ability to capture complex relationships within the data, thus increasing the efficacy of the FE process. The kernel process is advantageous for reasons such as capturing non-linear relationships within the data, implicit mapping of data to high-dimensional spaces for more effective separation and representation of features, and being flexible because it allows for the incorporation of prior assumptions or domain knowledge of the data by choosing the right kernel function. A general background on kernel methods is described in the next section.

4.2 Kernel Methods

The kernel methods can be described as universal function approximators. The schemes can estimate a non-linear mapping with any given accuracy based on a solid mathematical framework of reproducing the kernel Hilbert spaces (RKHS) [165]. The practical implication of applying this framework is that it allows the computation of dot products in high, possibly infinite-dimensional feature space as kernel functions in the input space. The RKHS's property enables the transformation of linear inner product-based algorithms to a high-dimensional space by just converting their inner products into kernels. This transformation serves to unveil intricate relationships and patterns that may not be readily evident within the original data, thus easing the data analysis process. When solving a kernel-based method, one is indirectly solving the linear algorithm in feature space, where the changed data is more likely to conform to a linear model [165]. Consequently, kernel methods have become an established framework for providing solutions to many ML problems, as evidenced by their experimental achievements in the field of image processing, computational biology, bioinformatics, telecommunications, and medicine [165–169].

4.2.1 Kernel Definitions

A kernel is a continuous, symmetric function $\mathrm{K} : X \times X \longrightarrow \mathbb{R}$ that operates on data in an input space X. The basic definitions of the types of kernels are [166]:

1. **Positive definite kernel**: A kernel K is said to be positive definite if, for any input data points $\{\mathbf{x}_i\}_{i=1}^{k} \in X$, it meets the following requirement:

$$\sum_{i,j=1}^{k} \beta_i \beta_j \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \qquad \forall \beta \in \mathbb{R}. \tag{4.1}$$

2. **Kernel matrix**: The $k \times k$ matrix $\mathbf{K}$ with elements $x_{ij} = \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j)$ emanating from a given set of k data points $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ is called the kernel matrix of K

with respect to the data, for $i, j = 1, \cdots, k$. A positive definite kernel is used to form a positive definite kernel matrix.

3. **Positive definite matrix**: A squared real-valued matrix $\mathbf{X}$ is said to be a positive definite matrix if it meets the requirement of (4.1). This condition is equivalent to requiring that $\beta^* \mathbf{X} \beta \geq 0, \forall \beta \in \mathbb{R}^k$. For the kernel matrix, this means that all its eigenvalues must be non-negative.

4.2.2 Reproducing the Kernel Hilbert Spaces

It can be shown that the feature space is related to a positive definite kernel in such a way that the kernel is a dot product in that feature space. When such a feature space is to be constructed, we define a feature mapping from X into the space of function $\mathbb{S}$ for a given positive definite kernel K, as [165]:

$$\Theta : X \longrightarrow \mathbb{S} \tag{4.2}$$

$$\mathbf{x} \rightarrowtail \mathrm{K}(\mathbf{x}, \cdot) \tag{4.3}$$

The function $\Theta(\mathbf{x})(\cdot)$ allocates the value $\mathrm{K}(\mathbf{x}, \mathbf{x}')$ to the input point $\mathbf{x}'$. By decoding the kernel function as a similarity function, this mapping corresponds to every input point $\mathbf{x}$ by its similarity, $\mathrm{K}(\mathbf{x}, \cdot)$, to all other points in the domain X.

For the purpose of constructing a feature space connected with Θ, the image of Θ must be turned into a vector space and provided with an inner product [166]. A likely vector space can be defined by taking linear combinations of the form:

$$f(\cdot) = \sum_{i=1}^{P} \beta_i \mathrm{K}(\mathbf{x}, \cdot), \tag{4.4}$$

where P, β_i and $\mathbf{x}_i$ are selected arbitrarily, and $i = 1, \cdots, P$. The inner product between f and another function, $g(\cdot) = \sum_{j=1}^{P'} \alpha_j \mathrm{K}(\mathbf{x'}_j, \cdot)$, in this space is defined as:

$$\langle f, g \rangle := \sum_{i=1}^{P} \sum_{j=1}^{P'} \beta_i \alpha_j \mathrm{K}(\mathbf{x}, \mathbf{x'}_j). \tag{4.5}$$

A property worth noting that arises directly from the definition of Θ is that all functions of the form of Equation (4.4) satisfy:

$$\langle \mathrm{K}(\mathbf{x}, \cdot), f \rangle = f(\mathbf{x}). \tag{4.6}$$

This means that K is representative of the evaluation of f. Specifically, the kernel K has the reproducing property [170]:

$$\langle \mathrm{K}(\mathbf{x}, \cdot), \mathrm{K}(\mathbf{x}', \cdot) \rangle = \mathrm{K}(\mathbf{x}, \mathbf{x}'). \tag{4.7}$$

Hence, positive definite kernels are also called reproducing kernels. The aforementioned narrative indicates that any positive definite kernel has an associated feature space where it can be viewed as a dot product:

$$\mathrm{K}(\mathbf{x}, \mathbf{x}') = \langle \Theta(\mathbf{x}), \Theta(\mathbf{x}') \rangle. \tag{4.8}$$

4.2.3 Kernel Trick

So far, we have shown that a feature map can be constructed from a kernel. Remarkably, the reverse also holds, where for every mapping Θ from the input X to a dot-product space, a positive definite kernel is obtained from Equation (4.8) (see [166] for the proof). This interchangeability between feature spaces and positive definite kernels gives rise to the property that is called the "kernel trick". Assuming an algorithm is expressed using a positive definite kernel K, it becomes possible to create an alternative algorithm by substituting K with another positive definite kernel, K′ [166].

Basically, the kernel trick transforms any inner-product-based algorithm into an alternative algorithm by replacing the inner products with a non-linear kernel. As per the identity (Equation (4.8)), implementing this alternative kernel-based algorithm corresponds to applying the original inner-product-based algorithm within the feature space. The feature space of the latter algorithm is of high dimensionality, which usually makes it intractable or even impossible to carry out precisely. Fortunately, the corresponding kernel-based algorithm can be applied in the input space using Equation (4.8), which essentially calculates the dot products between the images of all pairs of input data in the feature space. This saves the computational cost of computing the coordinates of the data feature space. The strength of this straightforward and elegant concept is that the obtained solution is a non-linear function of the input data, achieved by carrying out convex optimisation problems implicitly in the feature space. The advantages of transitioning to a higher-dimensional space were established by Cover's theorem [165, 166], which established that achieving linear separability becomes more feasible in such an expanded space.

4.2.4 Kernel Functions

In implementing the kernel trick, a kernel function must be selected. It should be noted that a kernel function symbolises the inner product in some feature space, but it does not need the input data to have a vector representation. The kernel function is a similarity function in some feature space that corresponds to a dot product. There are different types of kernel functions. The commonly used kernel functions are defined as follows [165, 171]:

1. **Polynomial kernel**: also known as order $\mathbf{p}$ is derived as:

$$\mathrm{K}(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + \mathbf{q})^{\mathbf{p}}, \tag{4.9}$$

   where $\mathbf{q}$ is a non-negative constant, typically 1.

2. **Linear kernel**: is a special type of polynomial kernel of order 1 defined as:

$$\mathrm{K}(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle = \mathbf{x}^* \mathbf{x}'. \tag{4.10}$$

3. **Gaussian kernel**: also known as the radial basis function (RBF) kernel, is computed as follows:

$$\mathrm{K}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2g^2}\right), \tag{4.11}$$

   where $\|\mathbf{x}-\mathbf{x}'\|^2$ represent the Euclidean distance between the two vectors and $\mathbf{g}$ is the scaling constant. Note that the factor 2 in the denominator is sometimes not included, with the assumption that it is included in the factor g^2 [165].

4. **Sigmoid kernel**: this is derived as:

$$\mathrm{K}(\mathbf{x}, \mathbf{x}') = \tanh(\mathbf{a}\langle \mathbf{x}, \mathbf{x}' \rangle + \mathbf{q}), \tag{4.12}$$

   where $\mathbf{a}$, $\mathbf{q} \in \mathbb{R}$ and are suitable constants. It should be noted that this kernel does not meet the positive-definite condition.

Each of the kernel functions comes with unique properties [131, 165]. The choice of kernel function to be used is inherently informed by a complex interplay of multiple factors, such as the data characteristics, problem understanding, and leveraging on domain knowledge about the inherent structures of the data, among others [32, 165, 171]. Generally, Gaussian kernels or polynomial kernels are often considered for data that exhibits linear or non-linear separability and other non-linear relationships [169, 171].

The kernel-based learning method was first introduced in the field of pattern analysis to enable researchers to analyse the non-linear relationships in data, with its first application in support vector machines (SVM) [166, 167, 169]. Since its introduction, it has been applied in conjunction with different ML algorithms to derive the kernelised versions of such algorithms. Examples include kernel principal component

analysis (kPCA), kernel independent component analysis (KICA), kernel linear discriminant analysis (KLDA), kernel springy discriminant analysis (KSDA), and kernel Gabor wavelets [166, 168, 171]. The kernel-based algorithms have been deployed in the literature for the analysis of diverse arrays of data types [169], in areas such as speech recognition, geostatistics, pattern recognition, communications, bioinformatics, and chemoinformatics, among others. In this study, we are exploring the versatility of kernel methods by integrating them with PCA and DMD for enhanced feature vectors for HMM for the detection of whale vocalisations.

4.3 Proposed Enhanced PC Feature Vectors for HMM

The kPCA is employed to effectively locate non-linear subspace in a dataset. In essence, it extends the PCA into a higher-dimensional feature space in order to uncover the non-linear characteristics of a dataset. It can be applied to domains where the PCA has been used for FE [32]. In this study, we leverage the potential of kPCA to enhance the PC-based feature vectors. The proposed enhanced PC feature vectors for HMM is developed as follows:

Non-linear data $\mathbf{X}$ can become linearly separable by mapping it into a higher-dimensional feature space as defined by:

$$x_i \in \mathbb{R}^m \longrightarrow \Theta(x_i) \in \mathbb{R}^d, \tag{4.13}$$

where m is the dimension of the original data, d is the projected high-dimensional space, x_i represents each data point of the original data in m-dimensional space, and $\Theta(x_i)$ represents each data point of the high-dimensional projection in d-dimensional space. The projected high-dimensional data is assumed to have zero mean, that is:

$$\frac{1}{d}\sum_{i=1}^{d}\Theta(x_i) = 0. \tag{4.14}$$

The covariance matrix, $\ddot{\mathbf{V}}$ of the d-dimensional data is defined as:

$$\ddot{\mathbf{V}} \in \mathbb{R}^{d\times d} = \frac{1}{d}\sum_{i=1}^{d} \Theta(x_i)\Theta(x_i)^*. \tag{4.15}$$

The eigendecomposition of Equation (4.15) is defined by:

$$\ddot{\mathbf{V}}\Psi_b = \Psi_b \Upsilon_b, \tag{4.16}$$

where $b = 1, 2, \cdots, d$, Ψ_b is the eigenvectors of the d-dimensional data and Υ_b are the corresponding eigenvalues of the d-dimensional data. From Equation (4.15) and Equation (4.16), we have [172]:

$$\frac{1}{d}\sum_{i=1}^{d} \Theta(x_i)[\Theta(x_i)^*\Psi_b] = \Psi_b \Upsilon_b, \tag{4.17}$$

which can be restructured as:

$$\Psi_b = \sum_{i=1}^{d} \varsigma_b \Theta(x_i). \tag{4.18}$$

Substituting Ψ_b in Equation (4.17) with Equation (4.18) yields:

$$\frac{1}{d}\sum_{i=1}^{d} \Theta(x_i)\Theta(x_i)^* \sum_{j=1}^{d} \varsigma_b \Theta(x_j) = \Upsilon_b \sum_{i=1}^{d} \varsigma_b \Theta(x_i). \tag{4.19}$$

The resulting high-dimensional data Equation (4.19) is intractable to compute. However, it can be simplified using the kernel trick. This is done with the introduction of a kernel function as defined in Equation (4.8) and pre-multiplying both sides of Equation (4.19) with $\Theta(x_i)^*$:

$$\frac{1}{d}\sum_{i=1}^{d} \mathrm{K}(x_i, x_j) \sum_{j=1}^{d} \varsigma_b \mathrm{K}(x_i, x_j) = \Upsilon_b \sum_{i=1}^{d} \varsigma_b \mathrm{K}(x_i, x_j). \tag{4.20}$$

Equation (4.20) can be written in matrix notation as:

$$\frac{1}{d}\mathbf{K}\mathbf{K}\varsigma_b = \Upsilon_b\mathbf{K}\varsigma_b,$$

$$\mathbf{K}\varsigma_b = d\Upsilon_b\varsigma_b, \tag{4.21}$$

where $\mathbf{K}_{i,j} = \mathrm{K}(x_i, x_j)$ and ς_b are the eigenvectors in d-dimensional space, and Υ_b are the corresponding eigenvalues in d-dimensional space. The resulting PCs can be computed using the kernel trick as defined by:

$$\tilde{\mathbf{Pc}} = \sum_{i=1}^{d} \varsigma_b \mathrm{K}(x_i, x_j). \tag{4.22}$$

The beauty of the kernel methods is that we do not have to explicitly compute $\Theta(x_1)$. The kernel matrix, $\mathbf{K}$

$$\mathbf{K} = \begin{bmatrix} \mathrm{K}(x_1, x_1) & \mathrm{K}(x_1, x_2) & \dots & \mathrm{K}(x_1, x_m) \\ \mathrm{K}(x_2, x_1) & \mathrm{K}(x_2, x_2) & \dots & \mathrm{K}(x_2, x_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{K}(x_m, x_1) & \mathrm{K}(x_m, x_2) & \dots & \mathrm{K}(x_m, x_m) \end{bmatrix}, \tag{4.23}$$

can be constructed from the original data, $\mathbf{X}$. In summary, the kPCA locates the non-linear subspace by projecting the dataset into high-dimensional space. The kernel trick is used to compute the new high-dimensional dataset by calculating the dot product of the input data without explicitly computing the projected dataset. A summary of the kPCA FE process described in this work is presented in Algorithm 4.

Algorithm 4: Feature Extraction with kPCA

1: **Input:** $[\bar{\mathbf{s}}, \mathcal{W}, p]$
2: **Output:** $\boldsymbol{F}$
3: Decide the portion of $\bar{\mathbf{r}}$ to be taken from $\bar{\mathbf{s}}$ by annotation
4: Transform $\bar{\mathbf{r}}$ to h number of $\mathbf{X}$ with respect to $\mathcal{W}$ and m
5: Pick a kernel function, $\mathrm{K}(x_i, x_j)$
6: Compute the kernel matrix, $\mathbf{K}$
7: Solve for eigenvectors ς_i using Equation (4.21)
8: Compute $\tilde{\mathbf{P}}\mathbf{c}$ using Equation (4.22)
9: Compute $\tilde{\mathbf{G}}$ as described in Equation (3.16)
10: Compute $\mathcal{F}_y$ from $\tilde{\mathbf{G}}$ as described in Equation (3.17)
11: **Do** horizontal concatenation for each $\mathcal{F}_i$ to obtain $\boldsymbol{F}$ for respective $\bar{\mathbf{r}}$ as described in Equation (3.18)

4.4 Proposed Enhanced DMD Feature Vectors for HMM

The key component of the DMD algorithm is SVD, as indicated in Equation (3.31). In this study, we introduce the concept of kernel into the DMD algorithm. The kernel dynamic mode decomposition (kDMD) intends to find a more efficient mathematical way of computing $\tilde{\boldsymbol{Z}}$, from which the modes are derived without computing the SVD. We achieved this by using the kernel trick described in Section 4.2.3. In using the kernel trick, the computation complexity of the model is determined by the number of snapshots, p, rather than the number of samples, n, as in the original DMD architecture.

Given that the observation data matrices $\mathbf{X}_a, \mathbf{X}_b \in \mathbb{R}^{n \times p}$, the computation of $\tilde{\boldsymbol{Z}}$ can be anticipated to the principal component space described by the SVD of the data matrix $\mathbf{X}_a$, as seen in Equation (3.31). Thus, when considering the $\tilde{\boldsymbol{Z}}$ eigenvalue problem of Equation (3.33), the eigenvector itself can be considered to be constructed

by the expansion as follows [173]:

$$\Psi = \mathbf{U}\psi. \tag{4.24}$$

Inserting Equation (4.24) in Equation (3.33),

$$\begin{aligned}
\mathbf{U}\psi\Upsilon &= \tilde{\boldsymbol{Z}}\mathbf{U}\psi && (4.25)\\
&= \mathbf{X}_b\mathbf{X}_a^{\dagger}\mathbf{U}\psi \\
&= \mathbf{X}_b\mathbf{X}_a^{\dagger}(\mathbf{X}_a\mathbf{V}\boldsymbol{\Sigma}^{\dagger})\psi \\
&= \mathbf{I}\mathbf{X}_b(\mathbf{V}\boldsymbol{\Sigma}^{\dagger})\psi && \text{where } \mathbf{I} = \mathbf{X}_a^{\dagger}\mathbf{X}_a \text{ is the identity matrix} \\
&= (\mathbf{X}_a^{*})^{\dagger}(\mathbf{X}_a^{*}\mathbf{X}_b)(\mathbf{V}\boldsymbol{\Sigma}^{\dagger})\psi && \text{as } \mathbf{I} = \mathbf{X}_a^{\dagger}\mathbf{X}_a^{*} \\
&= \mathbf{U}(\boldsymbol{\Sigma}\mathbf{V}^{*})(\mathbf{X}_a^{*}\mathbf{X}_b)(\mathbf{V}\boldsymbol{\Sigma}^{\dagger})\psi && (4.26)\\
&= \mathbf{U}\tilde{\boldsymbol{Z}}\psi. && (4.27)
\end{aligned}$$

Relating Equation (4.27) to Equation (4.26), $\tilde{\boldsymbol{Z}}$ can now be evaluated by the expression:

$$\tilde{\boldsymbol{Z}} = (\boldsymbol{\Sigma}\mathbf{V}^{*})(\mathbf{X}_a^{*}\mathbf{X}_b)(\mathbf{V}\boldsymbol{\Sigma}^{\dagger}). \tag{4.28}$$

The parameters in Equation (4.28) are of the following structure: $(\boldsymbol{\Sigma}\mathbf{V}^{*}) \in \mathbb{R}^{p\times p}$, $(\mathbf{X}_a^{*}\mathbf{X}_b) \in \mathbb{R}^{p\times p}$, and $(\mathbf{V}\boldsymbol{\Sigma}^{\dagger}) \in \mathbb{R}^{p\times p}$.

Thus, the computation of $\tilde{\boldsymbol{Z}}$ is determined by the number of snapshots taken, p, rather than the number of observables, n. However, instead of carrying out an SVD operation to determine $\boldsymbol{\Sigma}$ and $\mathbf{V}$ in Equation (4.28), the eigendecomposition of the $p \times p$ matrix, $\mathbf{X}_a^{*}\mathbf{X}_a\mathbf{V} = \boldsymbol{\Sigma}^2\mathbf{V}$, is implemented. Therefore, having noted that the kernel function represents a dot product in some feature space, the kernel function, $\mathrm{K}(\mathbf{x}, \mathbf{x}')$, can be related to the observables used to construct $\mathbf{X}_a$ and $\mathbf{X}_b$. Using a polynomial kernel:

$$\mathrm{K}(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + \mathbf{q})^2, \tag{4.29}$$

where $\mathbf{q}$ is a non-negative constant, usually 1, and $\mathbf{x}$, $\mathbf{x}'$ are data points in $\mathbb{R}^2$. Hence, the computation of the observables is reduced to a simple dot product between the

vector pairs. In practice, in place of defining observables $\mathbf{X}_a^*\mathbf{X}_b$ in Equation (4.28) with Equation (3.26) and Equation (3.27), we define the observables $\mathbf{X}_a$ using the kernel function, K($\mathbf{x}$, $\mathbf{x}$'), for producing the dot products associated with the feature space. To be precise, elements of the observable matrices are defined by:

$$\mathbf{X}_a^*\mathbf{X}_b(i,j) = \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j{}'), \tag{4.30}$$

where (i, j) represents the i-th and j-th column of the correlation matrix, and $\mathbf{x}_i$ and $\mathbf{x}_j{}'$ are the data from the i-th and j-th columns. The computation of matrices $\mathbf{V}$ and $\mathbf{\Sigma}$ required in the kDMD formulation of Equation (4.28) is defined as:

$$\mathbf{X}_a^*\mathbf{X}_a\mathbf{V} = \mathbf{\Sigma}^2\mathbf{V} \qquad \text{(recall the definition Equation (4.24))} \tag{4.31}$$

$$\mathbf{X}_a^*\mathbf{X}_a(i,j) = \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j{}'). \tag{4.32}$$

Consequently, the kernel method is used to efficiently produce the dot products after choosing a kernel type (in this study, the polynomial kernel Equation (4.29) is used). $\tilde{\boldsymbol{Z}}$ can be calculated through Equation (4.28). The eigendecomposition of $\tilde{\boldsymbol{Z}}$ derived from Equation (4.28) will give the modes $\tilde{\boldsymbol{M}}$, which are the same as the structure of Equation (3.34). Thereafter, the FE steps are the same as defined from Equations (3.34)–(3.36). A summary of the kDMD FE process described in this chapter is presented in Algorithm 5:

Algorithm 5: Feature Extraction using kDMD

1: **Input:** $[\bar{\mathbf{s}}, \mathcal{W}, p]$
2: **Output:** $\boldsymbol{F}$
3: Following steps (3) to (5) in Algorithm (3)
4: Define observable matrices $\mathbf{X}_a$, $\mathbf{X}_b$ using the kernel function as defined in Equation (4.30)
5: Compute matrices $\boldsymbol{\Sigma}$ and $\mathbf{V}$ required for kDMD formulation of Equation (4.28) using polynomial kernel
6: Compute the eigendecomposition of $\tilde{\boldsymbol{Z}}$ derived from Equation (4.28)
7: Obtain $\tilde{\boldsymbol{M}}$, which has the same structure as Equation (3.34)
8: Compute $\boldsymbol{F}$ with the $\tilde{\boldsymbol{M}}$ obtained in step (7) above by following step (7) and step (8) in Algorithm (3)

4.5 Conclusion

The existence of non-linear characteristics in whale vocalisations may present a challenge for optimal FE using the ED algorithms, namely PCA and DMD, which were introduced as FE techniques in Chapter 3. Given that the quality of feature vectors plays a crucial role in the performance of HMM, this chapter has been dedicated to the development of enhanced feature vectors through the introduction of the kernel method into these techniques. The kernel method enables the proper capture of the non-linear characteristics, as may be found in whale vocalisations. The augmentation of the ED-FE with kPCA and kDMD aims to further enhance the feature extraction process. The emerging feature vectors developed from PCA, DMD, kPCA, and kDMD will be separately fed into the HMM for the detection of whale vocalisations. The next chapter will demonstrate the practical application of the developed feature vectors as used with HMM for the detection of whale vocalisations.

Chapter 5

Data Description and Experimental Set-up

5.1 Introduction

In Chapter 3, eigendecomposition-based (ED) feature extraction (FE) techniques were developed, specifically utilising principal component analysis (PCA) and dynamic mode decomposition (DMD). Furthermore, the kernel method was proposed in Chapter 4, to enhance the developed PCA and DMD based FE techniques. In this chapter, the emerging feature vectors developed from PCA, DMD, kPCA, and kDMD will be separately fed into the hidden Markov model (HMM) for the detection of whale vocalisations. First, a detailed description of the datasets used in this study is given. Second, a summary of the workings of HMM as used in this study is presented, and lastly, the experimental setup is described.

5.2 Data Description

This section describes the datasets used for testing the developed models in this study. Two species of whales were selected: southern right whales (SRW) and humpback whales (HW). The datasets deployed were visually and acoustically analysed for manual identification and annotation of the vocalisations using *Sonic Visualiser* (v4.4) [174]. Furthermore, custom-written MATLAB codes were utilised to visualise the spectrograms of the respective datasets. The MATLAB codes enable us to zoom-in on any interesting portion of the datasets for ease of analysis. The consistency of any identified vocalisations was confirmed and documented for the purpose of annotation while listening to the audio files as they ran across the waveforms and spectrogram. The parameters used for the manual annotation of the vocalisations present in the datasets include the start-time, the end-time, the low frequency, the high frequency, and the visual acoustic 'shape' or contour, respectively, for each vocalisation. The time-frequency bounds of SRW and HW vocalisations are marked in the respective datasets. The vocalisations in the respective datasets were categorised into two classes: whale vocalisations and noise. Whale vocalisations are sounds that have been identified by an expert as signals of interest for a respective species, while any vocalisations that do not represent the sound of interest are classified as noise.

5.2.1 Southern Right Whales

Southern right whales (SRW), also known as *Eubalaena australis*, are large whales that can reach a length of 17 m and weigh up to $60,000$ kg. They have extremely large heads, which can be one-fourth to one-third of their body length [2]. The SRW are found in the oceans south of the Equator, from approximately 20°S to 60°S. They are migratory, generally moving south in the summer months to feed, and in the winter they migrate north, where mating and calving take place. SRW presence is confirmed in South Africa, Argentina, Australia, and parts of New Zealand, and they are mostly concentrated near coastlines. The SRW are one of the three recognised

species of right whales. The other two are the North Atlantic right whales (*Eubalaena glacialis*) and the North Pacific right whales (*Eubalaena japonica*). The three different species are separated (in terms of habitats) by ocean basins, as indicated by their names. They have a very similar physical appearance but quite distinct genetics and differing conservation statuses [2]. Before now, the SPCC and the NN are the available automated methods used for the detection of SRW [94]. This study introduces ML methods for the detection of SRW vocalisations. The waveform and spectrogram views of SRW vocalisations are as depicted in Figure 5.1.

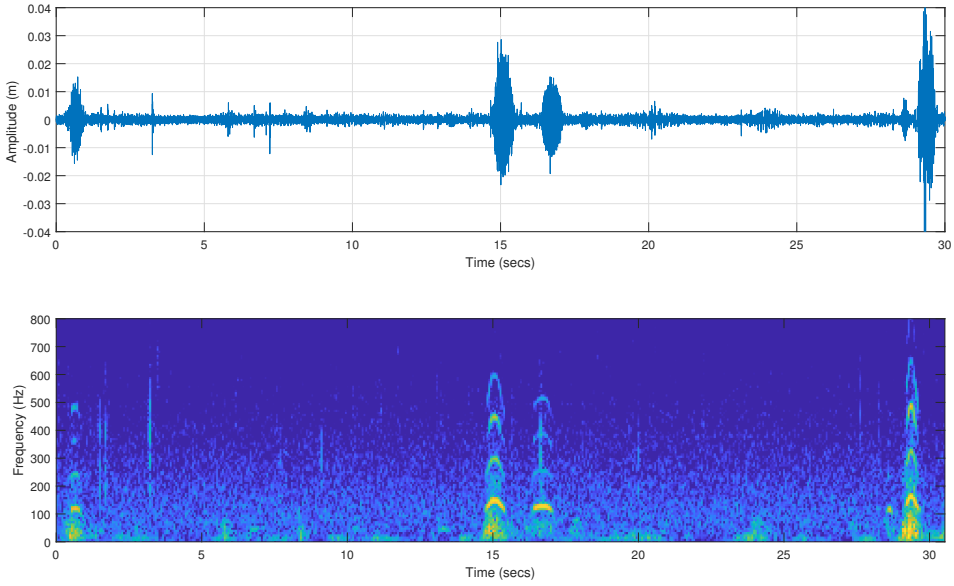


FIGURE 5.1: The waveform and spectrogram views of southern right whale (SRW) vocalisations.

SRW are known for their diverse repertoire of acoustic signals, which have been broadly divided into three groups: calls, blows, and slaps. The calls are complex and characterised by varying patterns and pitches. They serve as a major means of communication and social interaction among individual SRW. Furthermore, when SRW surface to breathe, they produce distinct exhalations known as 'blows' or 'spouts' through their blowholes. The slaps occur when SRW strike the water's surface with their flippers or tails to create fascinating visual and acoustic displays. SRW produce harmonic calls with a frequency range of 50 Hz to 2,000 Hz, which can last from 0.5 seconds to 1.5 seconds [175].

The recordings of the SRW dataset used in this research took place around the coastal waters of False Bay, Western Cape, South Africa. The recordings were done via a hydrophone carefully deployed as a drifting buoy at False Bay at the stroke of noon on December 10, 2020. Numerous SRW calls were collected during a period of 2 hours. The hydrophone (Aquarian Audio H2C) was suspended approximately 9 m below a custom-made drifting buoy. The buoy is an ingeniously constructed 110 mm-diameter PVC pipe about 500 mm long, with strong end caps. The buoy contained a 2 kg ballast weight and housed a Zoom H1n audio recorder. The audio is sampled at $96,000$ Hz and saved as .WAV files. The audio files were downsampled to $8,000$ Hz from the original $96,000$ Hz sampled frequency for ease of analysis. An elliptical filter was then used to obtain a passband between 100 Hz and $1,000$ Hz, as described in [176].

A total of 169 SRW calls were manually identified and annotated. The vocalisations of SRW have been classified according to their acoustic contours into four types: *up* calls, *down* calls, *flat* calls, and *variable* calls [177]. The *variable* call type is present in the dataset. The spectrogram of selected examples of SRW calls in the dataset is shown in Figure 5.2. The spectrogram plotting parameters were set as follows: sampling frequency $8,000$ Hz, window length 512, fast Fourier transform (FFT) length of 1024 samples, 50% overlap, Hamming window, frequency range $0 - 800$ Hz. The mean duration of identified calls is 0.65 seconds. The red boxes are the time-frequency boundaries of the manual annotation of the identified calls, while the black boxes are examples of selected noise portions.

5.2.2 Humpback Whales

Humpback whales (*Megaptera novaeangliae*) (HW), are medium-sized baleen whales, about 11 m to 15 m long and $20,000$ kg to $30,000$ kg in weight. The HW are easily identifiable from other whale species based on their different physical structure, which includes a pectoral fin, extremely long flippers, and knobbly head. HW are one of

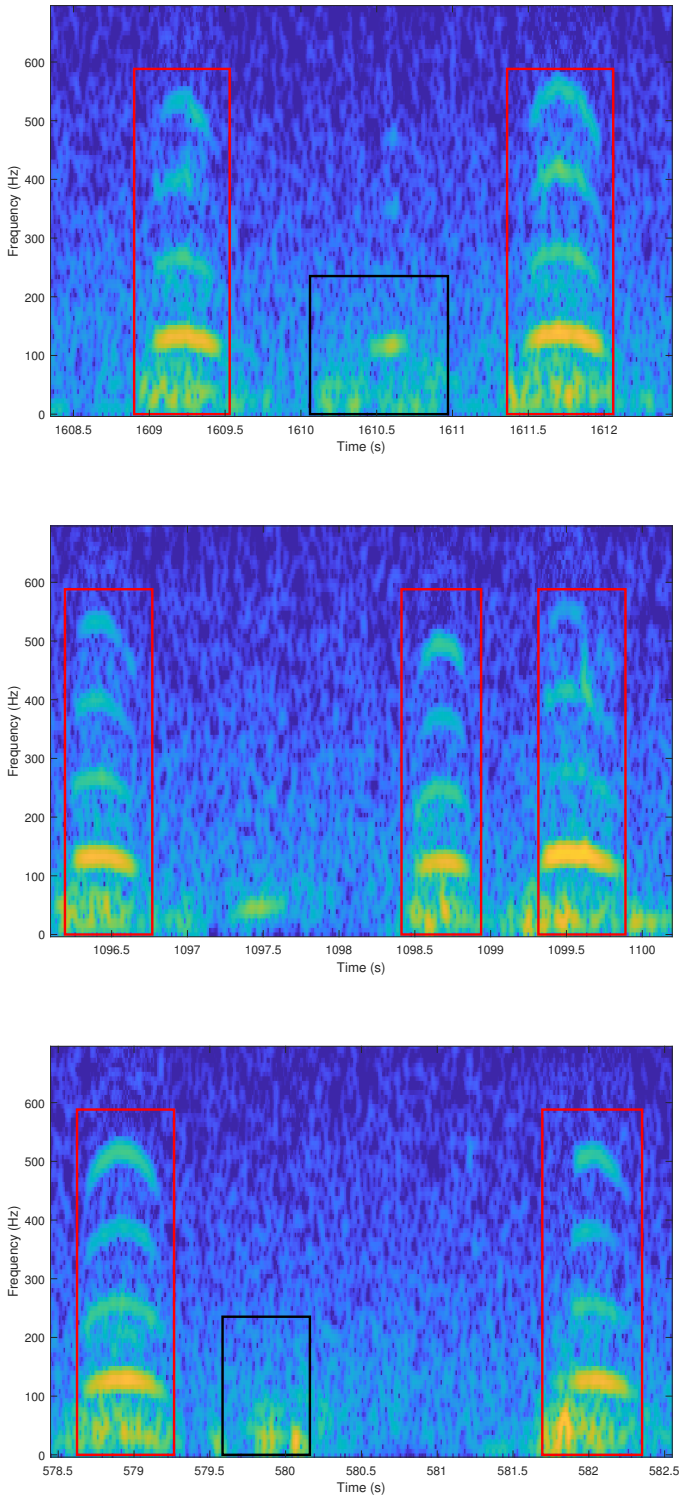


FIGURE 5.2: Spectrogram views of different selected portions of SRW vocalisations at different durations within the dataset.

the best known of all the large whales. They feed and breed in coastal waters over the continental shelves of all continents, often near centres of human population.

HW emit and apply various kinds of vocalisations, including complex periodic sequences known as songs on their breeding grounds. The song, which changes over time, includes both tonal and pulsed sounds. The song units are individual calls, which are repeated and merged to create phrases. These phrases are then repeated to create longer signature tunes [34], also known as themes. Thus, a song is made up of 4 to 12 themes. Song units can last for fractions of a second to several seconds, while a song can last from 5 minutes to 30 minutes. HW can sing for a few minutes or for up to 48 hours or more. Song units frequency range from 20 Hz to 4 kHz, and sometimes to 8 kHz [13, 34]. The HW vocalisations have been gathered in different locations over the years, and this has led to an abundant availability of datasets, thus motivating the use of several automatic detection methods to analyse HW signals. MFCC-based hidden Markov model (MFCC-HMM) and LPC-based hidden Markov model (LPC-HMM) [19, 25] are some of the previous models that have been used to detect HW vocalisations. Therefore, the use of HW vocalisations is encouraged in this study to confirm the viability and performance of the proposed models for the detection of whale vocalisations.

The HW datasets used in this work were retrieved from *MobySound*,[1] a public database library for research on the automatic recognition of cetacean species through their vocalisations. The recordings were made off the North coast of Kaui Island, Hawaii, USA. The dataset was captured using custom-made hydrophones (see [178] for more details on the data recording). The recordings were sampled at $4,000$ Hz. Figure 5.3 shows the waveform and spectrogram views of HW vocalisations.

A total of 289 HW calls were manually identified and annotated. The calls vary in duration, frequency range, and amplitude. The variation in the acoustic contours makes it complex to categorise the units. In the literature, different researchers give different names to the different call units [179]. Therefore, we based our identification

[1] http://www.mobysound.org/mobysound.html

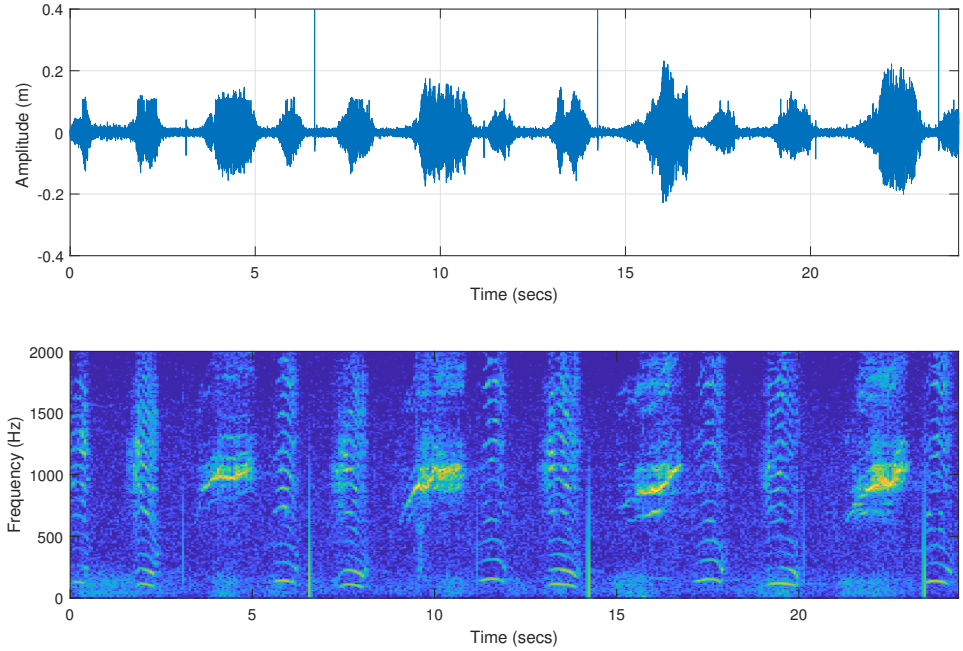


FIGURE 5.3: The waveform and spectrogram views of humpback whale (HW) vocalisations.

on the consistency of the acoustic contours as observed in the dataset. The spectrogram of selected portions of the HW calls in the dataset is shown in Figure 5.4. The spectrogram plotting parameters were set as follows: sampling frequency $4,000$ Hz, window length 256, fast Fourier transform (FFT) length of 1024 samples, 50% overlap, Hamming window, frequency range $0 - 2,000$ Hz. The mean duration of identified calls is 2.20 seconds. The time-frequency boundaries of the identified call types are in the red, magenta, and white boxes, while the signals in the black boxes are examples of noise in the dataset. It can be noted that the same call type often runs in succession. Due to the complexity of the acoustic contours of the signal emitted by whales, the model proposed is built upon the sampling points of the datasets.

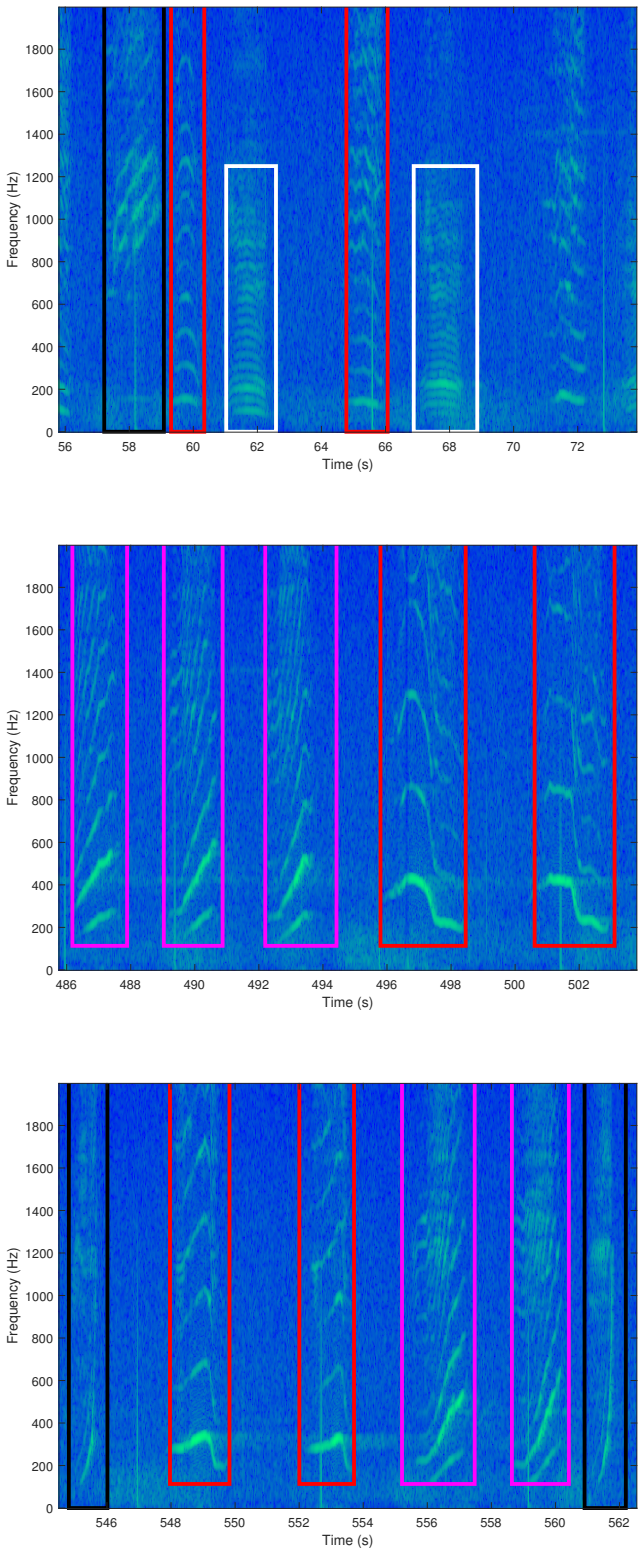


FIGURE 5.4: Spectrogram views of different selected portions of HW vocalisations at different durations within the dataset.

5.3 Summary of the HMM Process for the Detection of Whale Vocalisations

The block diagram of the HMM process for the detection of whale vocalisations is shown in Figure 5.5. The whale vocalisation data gathered from PAM are preprocessed. Thereafter, the preprocessed dataset is divided into two uneven portions: 75% and 25%. The 75% is reserved for the training stage of the HMM and is thus referred to as the "Training" data. Likewise, the 25% balance is reserved for the detection stage of the HMM process and is thus referred to as the "Testing" data.

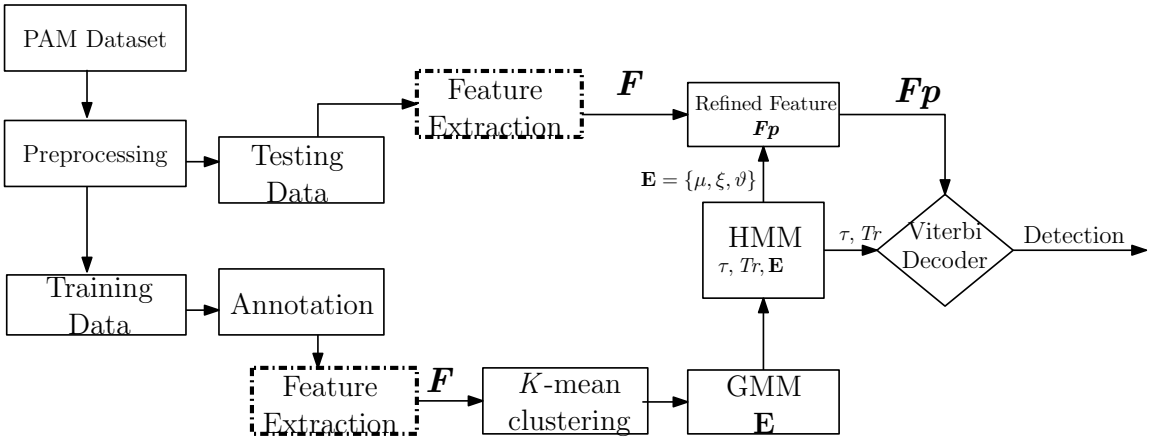


FIGURE 5.5: HMM process for the detection of whale vocalisations.

For the training stage, training data is annotated before ED-FE techniques are deployed to extract features. The extracted features are then transformed into an appropriate feature matrix, $\boldsymbol{F}$ for the HMM. The $\boldsymbol{F}$ represent the observation sequence, Q of the HMM architecture. Therefore, $\boldsymbol{F}$ are used to train the HMM parameters, $\Pi = (\tau, Tr, \mathbf{E})$. As previously noted, HMM is sensitive to random starting values for the emission distribution parameters, $\mathbf{E}$. Hence, $\boldsymbol{F}$ is passed through the K-means algorithm and GMM before being fed into the HMM block as shown in Figure 5.5. The HMM trains and outputs optimal parameters, $\Pi = (\tau, Tr, \mathbf{E})$.

Similarly, the detection stage starts with the extraction of features from the testing data. The features are extracted using ED-FE techniques. The extracted features are

then transformed into $\boldsymbol{F}$. The $\mathbf{E}$ is used to refine the $\boldsymbol{F}$ of the testing data in order to produce refined feature matrix, $\boldsymbol{Fp}$. Next, the refined feature matrix, $\boldsymbol{Fp}$, the initial probabilities, τ and the transition probabilities, Tr are passed into the Vit-alg as shown in Figure 5.5. Given the observation sequence in the form of $\boldsymbol{Fp}$, τ and Tr, the Vit-alg goes on to compute all possible hidden state paths and output the paths with the best probability. The output of Vit-alg represents the state sequence. The state sequence derived from Vit-alg are then compared with the original (testing) datasets to evaluate the performance of the models.

5.4 Experiments

This section describes the experimental set-up for the implementation of the developed models.

5.4.1 Data Preprocessing

The PAM datasets containing SRW and HW vocalisations were used for the analyses of the models developed in this study. The PAM datasets in .WAV format were discretised into time intervals through sampling. The sampled dataset, contains samples, which were measured at temporally equispaced instants in time. The audio files of the datasets were then visually and acoustically analysed using *Sonic Visualiser* (v4.4) [174]. The waveform and spectrogram views of the recognised whale vocalisations of the two species were noted in both analyses. The consistencies of the recognised whale vocalisations were confirmed and documented for the purpose of annotation while listening to the audio files as they ran across the waveforms and spectrogram.

The datasets were divided into two categories: recognised whale vocalisations and noise. The start and end points of recognised whale vocalisations were diligently annotated during the categorisation process, and any other sounds in the audio file

that were not recognised as whale vocalisations were classified as noise. Thereafter, we divided the datasets into two uneven fractions: 75% and 25% for training and testing purposes, respectively.

5.4.2 Implementation

The proposed ED-FE techniques developed were separately used with the HMM for the detection of whale vocalisations. The two portions of the annotated datasets were deployed for training and testing.

Selected fraction of the datasets, $\bar{\mathbf{r}}$, representing either a recognised whale vocalisation or noise, was selected for training as explained in Section 3.2. To model the training datasets, two separate four-state HMMs with two mixtures each were created. One for the recognised whale vocalisations and the other for the noise. In the construction of the data matrix, $\mathbf{X}$, the numbers of samples in an observation, n, were determined based on acoustic window size, $\mathcal{W}$, (as explained in Section 3.2). During the experimental phase, $\mathcal{W}$ were selected at different times as either 32, 64, or 128. Specifically, when $\mathcal{W} = 32$, $\mathbf{X}$ is constructed with 32 rows per column, each column represents a distinct observation. Consequently, each column of $\mathbf{X}$ is populated from $\bar{\mathbf{r}}$, as each column fills up with $\mathcal{W}$ samples, the next column is filled with the next $\mathcal{W}$ samples, continuing this pattern until all the samples are taken. Thus, the number of columns, m, in $\mathbf{X}$ is subject to the number of sample points in $\bar{\mathbf{r}}$. The recognised whale vocalisations and the noise were trained separately to obtain the trained parameter $\mathit{\Pi}$, as summarised in Section 5.3. Therefore, the trained parameters for the recognised whale vocalisations are represented as:

$$\mathit{\Pi}^D = (\tau^D, Tr^D, \mathbf{E}^D), \tag{5.1}$$

while the trained parameters for the noise are represented as:

$$\mathit{\Pi}^N = (\tau^N, Tr^N, \mathbf{E}^N). \tag{5.2}$$

Thereafter, the two HMMs were combined to form an eight-state four-mixture HMMs given as:

$$\Pi^{DN} = [\tau^D \mid \tau^N], \quad \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \\ S_8 \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{bmatrix} & & & \\ & Tr^D & & \\ & & & \end{bmatrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \delta & 0 & 0 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 \end{bmatrix} \\ \begin{bmatrix} & & & \\ & Tr^N & & \\ & & & \end{bmatrix} \end{array} \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \\ S_8 \end{array}, \quad [\mathbf{E}^D \mid \mathbf{E}^N],$$

where δ and γ are the switching probabilities from whale vocalisations to noise, whose values are determined by training the combined HMMs. States 1 to 4 in the newly formed eight-state HMM, represent the recognised whale vocalisations, Π^D, while states 5 to 8 represent the noise, Π^N.

For the detection stage, $\boldsymbol{F}$ of the 25% testing data was calculated. The emission probabilities from the newly formed model were used to refine the $\boldsymbol{F}$, derived from the testing data. The refined $\boldsymbol{Fp}$, the combined $\tau^{DN} = [\tau^D, \tau^N]$, and $Tr_{pr}^{DN} = [Tr^D, Tr^N]$ were passed into the Vit-alg as:

$$\text{Vit-alg} = \Big([\tau^D, \tau^N], [Tr^D, Tr^N], [\boldsymbol{Fp}^D, \boldsymbol{Fp}^N]\Big) \tag{5.3}$$

The Vit-alg predicts the path with the best probability, where each predicted path corresponds to one of the eight states within the combined model. Therefore, the predicted path falls into either whale vocalisation or noise, reflecting the states defined in the model. Subsequently, the detection output from the Vit-alg is compared with the original (test) data to validate the functionality of the proposed models and, hence, evaluate their performance. This comparison facilitates an assessment of the models' predictions against the actual (test) data, providing insights into the efficacy of the detection process.

5.5 Conclusion

In this chapter, we present a detailed description of the datasets used in this study as well as the practical implementation process of the developed ED-FE techniques with HMM. A block diagram was used to summarise the workings of HMM as used in this study. This is followed by explanations of the experiments as implemented in the study. The emerging models, namely: PCA-HMM, kPCA-HMM, DMD-HMM, and kDMD-HMM, were experimented on PAM datasets containing SRW and HW species. The results obtained from the various experiments carried out on each of the models, subject to various factors, are analysed and discussed in Chapter 6.

Chapter 6

Results and Discussion

6.1 Introduction

The developed ED-FE techniques in previous chapters were fed into HMM for the detection of whale vocalisations. In this chapter, we focus on the evaluation of the emerging models, namely: PCA-based hidden Markov model (PCA-HMM), kernel PCA-based hidden Markov model (kPCA-HMM), DMD-based hidden Markov model (DMD-HMM), and kernel DMD-based hidden Markov model (kDMD-HMM). These models were experimented on PAM datasets containing SRW and HW species. The experiments were done using a *Microsoft Windows 10 Pro* computer with an Intel Core(TM) i7-6600U CPU (@2.60 GHz) and 16 GB of RAM, running MATLAB 2021b. The results obtained from the various experiments carried out on each of the models, subject to various factors, are analysed and discussed in subsequent sections. The performances of the models are evaluated using the following metrics: true positive rate (TPR), precision (PREC), and error rate (ERR).

Parts of the experimental results analysed and discussed in this chapter have been published in peer-reviewed journals [180, 181]. The details of the manuscripts are as follows:

[1] A. M. Usman and D. J. J. Versfeld, "Principal components-based hidden Markov model for automatic detection of whale vocalisations," *Journal of Marine Systems Volume 242*, 2023. https://doi.org/10.1016/j.jmarsys.2023.103941.

[2] A. M. Usman and D. J. J. Versfeld, "Detection of baleen whale species using kernel dynamic mode decomposition-based feature extraction with a hidden Markov model," *Ecological Informatics Volume 71 (101766)*, 2022. https://doi.org/10.1016/j.ecoinf.2022.101766.

6.2 Results and Discussion: Performance Analysis of PCA-HMM and kPCA-HMM

In this section, we analyse and discuss the results obtained from the experiments conducted on PCA-HMM and kPCA-HMM. The PCA-HMM and kPCA-HMM were simulated at $\mathcal{W} = 32, 64, 128$ with various values of p as shown in Tables 6.1–6.3. The different values of window size, $\mathcal{W}$, represent the number of samples considered for various numbers of measurements, n. The models are p-dimensional. This implies that the higher the values of p, the more complex the computational load of the models. Hence, various simulations were run to experimentally confirm the effect of $\mathcal{W}$ with respect to different p.

TABLE 6.1: Simulation results for different p at $\mathcal{W} = 32$.

SRW vocalisations							HW vocalisations						
p	TPR (%)		PREC (%)		ERR (%)		p	TPR (%)		PREC (%)		ERR (%)	
	PCA	kPCA	PCA	kPCA	PCA	kPCA		PCA	kPCA	PCA	kPCA	PCA	kPCA
3	75.89	76.21	68.03	68.90	23.50	22.40	3	75.22	76.15	66.14	67.58	23.98	22.48
4	76.29	76.98	68.98	69.90	21.20	20.20	4	75.98	76.40	67.10	68.54	22.43	20.35
5	79.78	80.01	70.99	73.00	19.20	18.45	5	78.80	79.20	69.25	70.70	21.00	18.60
6	84.00	85.49	84.23	85.90	18.30	17.60	6	83.33	84.53	82.49	84.95	20.22	17.75
7	84.01	85.50	84.25	85.91	18.00	17.40	7	83.35	84.56	82.50	84.98	19.50	17.50
8	84.01	85.50	84.25	85.91	17.90	17.30	8	83.37	84.59	82.51	84.99	19.20	17.45
9	84.01	85.50	84.25	85.90	17.80	17.10	9	83.40	84.60	82.52	85.00	19.00	17.41
10	84.02	85.51	84.25	85.90	17.75	17.15	10	83.41	84.60	82.52	85.00	19.00	17.38
11	84.02	85.52	84.25	85.90	17.85	17.25	11	83.41	84.60	82.52	85.00	19.20	17.45
12	84.01	85.53	84.25	85.90	18.00	17.40	12	83.41	84.60	82.52	85.00	19.60	17.65

TABLE 6.2: Simulation results for different p at $\mathcal{W} = 64$.

SRW vocalisations							HW vocalisations						
p	TPR (%)		PREC (%)		ERR (%)		p	TPR (%)		PREC (%)		ERR (%)	
	PCA	kPCA	PCA	kPCA	PCA	kPCA		PCA	kPCA	PCA	kPCA	PCA	kPCA
3	77.01	78.15	69.20	70.85	22.80	21.75	3	77.75	78.90	69.70	71.50	22.45	21.35
4	77.80	78.95	70.21	71.92	20.50	19.20	4	78.50	79.85	70.75	72.75	20.15	18.82
5	80.82	81.70	73.90	75.25	18.75	17.40	5	81.50	82.60	74.43	75.91	18.40	17.03
6	86.65	88.90	85.00	87.16	17.87	16.70	6	87.33	89.80	85.60	87.85	17.50	16.34
7	86.75	89.00	85.12	87.25	17.05	16.20	7	87.51	89.85	85.68	87.90	17.45	16.20
8	86.85	89.00	85.15	87.35	16.90	15.90	8	87.55	89.90	85.70	87.92	17.35	16.10
9	86.85	89.10	85.15	87.35	16.80	15.80	9	87.56	89.90	85.75	87.92	17.30	16.05
10	86.85	89.00	85.15	87.35	16.81	15.70	10	87.56	89.90	85.75	87.92	17.30	16.00
11	86.86	89.00	85.16	87.35	16.90	15.75	11	87.57	89.90	85.75	87.92	17.35	16.10
12	86.86	89.00	85.18	87.40	17.50	16.20	12	87.57	89.90	85.75	87.92	17.45	16.25

TABLE 6.3: Simulation results for different p at $\mathcal{W} = 128$.

SRW vocalisations							HW vocalisations						
p	TPR (%)		PREC (%)		ERR (%)		p	TPR (%)		PREC (%)		ERR (%)	
	PCA	kPCA	PCA	kPCA	PCA	kPCA		PCA	kPCA	PCA	kPCA	PCA	kPCA
3	79.40	80.50	71.05	72.20	21.40	20.50	3	80.25	81.10	72.30	73.80	21.05	19.80
4	80.25	81.70	72.50	73.00	19.50	18.90	4	80.90	82.25	74.90	75.20	18.99	18.10
5	83.00	85.85	76.70	78.95	17.75	17.40	5	83.65	87.40	78.25	81.04	17.40	16.80
6	90.50	92.25	86.00	88.40	16.89	16.60	6	91.05	93.80	87.10	89.60	16.33	15.80
7	91.31	92.40	86.20	88.56	16.50	16.10	7	91.55	93.90	87.30	89.73	16.05	15.43
8	91.34	92.45	86.20	88.56	16.40	15.55	8	91.67	93.95	87.35	89.78	15.99	15.35
9	91.34	92.44	86.25	88.56	16.30	15.30	9	91.68	93.97	87.38	89.79	15.95	15.25
10	91.35	92.48	86.27	88.56	16.30	15.25	10	91.68	93.98	87.40	89.80	15.92	15.20
11	91.35	92.50	86.27	88.56	16.50	15.25	11	91.68	93.98	87.40	89.80	15.90	15.20
12	91.35	92.50	86.28	88.56	16.70	15.75	12	91.68	93.98	87.40	89.80	16.15	15.35

The true positive rate (TPR) and precision (PREC) performance considerably improved as p values increased from 3 to 4, and there were substantial improvements from 5 and 6 as displayed in Figures 6.1 and 6.2. But as can be observed, there are no further gains from 6 to 12; instead, the error rate (ERR) gradually rises, as shown in Figure 6.3. Nonetheless, it can be observed occasionally that the TPR and PREC values increase past $p = 6$. For instance, the difference in TPR value between $p = 6$ and $p = 7$ at $\mathcal{W} = 32$ is 0.01% (SRW) and 0.27% (HW), whereas at $\mathcal{W} = 128$ the difference is 0.81% (SRW) and 0.05% (HW). Since $p = 6$ is the value at which the majority of the results level off, it is chosen as the optimal dimension for the models. Therefore, increasing the value of p beyond 6 only adds to the computational load of the models without necessarily improving their performance. Accordingly, $p = 6$ provides the best dimension for the model as it strikes a balance between computational

load and the models' performance.

We compared the models' performance across species. It is observed that the number of samples, $\mathcal{W}$, has an effect on the performance of the models in terms of the vocalisations of different species. For example, across the three performance measuring metrics, SRW performed better than HW at $\mathcal{W} = 32$. However, HW performance improved more than the respective SRW performance with increase in $\mathcal{W}$. As shown in Tables 6.1–6.3, for $\mathcal{W} = 32$, the TPR for SRW is 85.49% whereas the TPR for HW is 84.53% at $p = 6$. On the other hand, for $\mathcal{W} = 128$, the TPR for HW is 93.80%, while the TPR for SRW is 92.25%. Besides, the ERR is observed to be higher at $\mathcal{W} = 32$ for HW than SRW. This is because the HW vocalisation is of longer duration compared to SRW. Therefore, a larger number of samples are needed to effectively train HW vocalisations. Though both species performance improved as the number of samples increased.

It is also observed that the models' performance stabilises earlier on SRW than HW. The TPR/PREC for HW is slightly better at $p = 7$ than $p = 6$ when compared to respective TPR/PREC for SRW. Across the three metrics, the marginal improvement in TPR and PREC values after $p = 6$ is observed to be significant with an increase in $\mathcal{W}$. In SRW, for instance, for $\mathcal{W} = 32$, the difference in PREC/TPR between respective $p \geq 6$ is ≤ 0.02; however, for $\mathcal{W} = 64$, the difference in PREC/TPR ≈ 0.1, and for $\mathcal{W} = 128$, the difference is ≈ 0.2. This demonstrates that training with more samples improves model performance. Based on the significance level that is often used to test statistical hypotheses, a paired t-test was conducted to compare the performance of PCA-HMM and kPCA-HMM. The level of significance

TABLE 6.4: Paired t-test results to compare the performance of PCA-HMM and kPCA-HMM on SRW and HW vocalisations at a significance level of 0.05.

	$\boldsymbol{p}$-values		
	TPR	PREC	ERR
SRW	2.4680×10^{-5}	4.1110×10^{-6}	4.6347×10^{-5}
HW	6.5887×10^{-6}	6.2600×10^{-6}	1.1048×10^{-6}

was set at 0.05. The results of the paired t-test as shown in Table 6.4, for $\mathcal{W} = 128$,

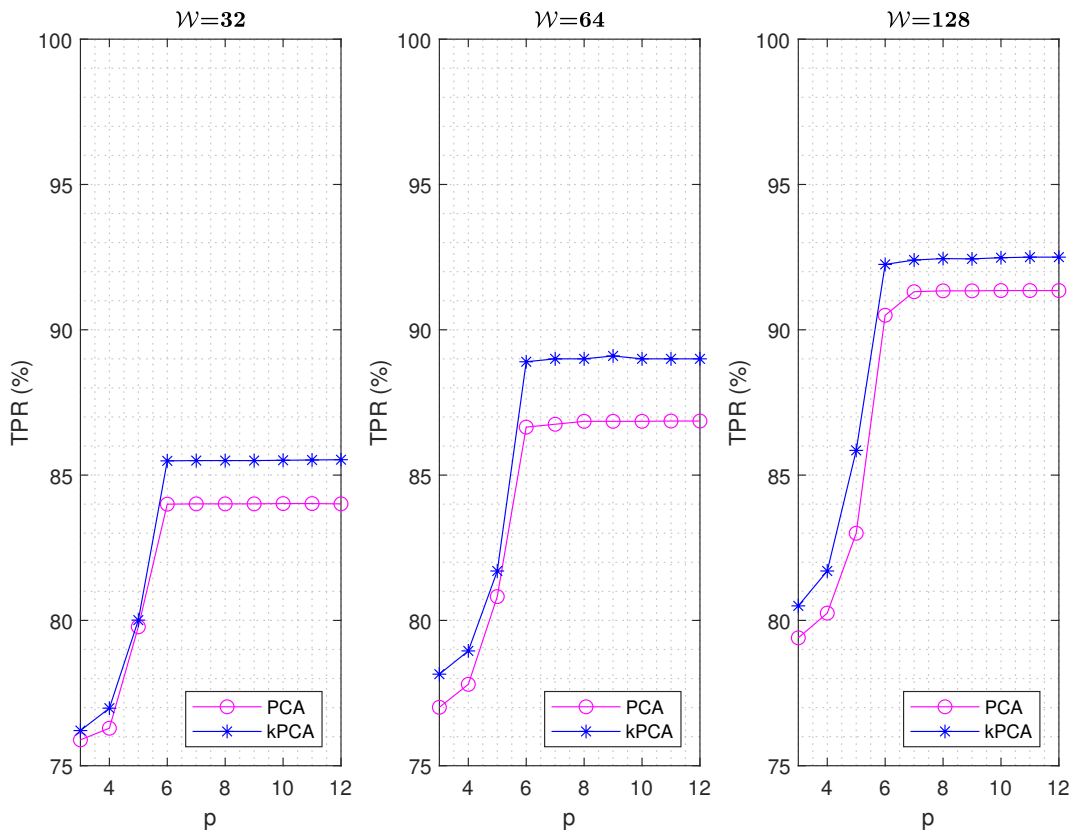
indicated that the kPCA-HMM demonstrated significantly better performance than the PCA-HMM ($\boldsymbol{p} < 0.05$). These findings confirm that the superior performance of the kPCA-HMM is indeed optimal over the PCA-HMM and not merely as a result of sampling variation.

The complexity of the FE algorithms is analysed to further assess the computational cost of the models. The worst-case time complexity analysis for each algorithm is carried out using the big-$\mathcal{O}$ notation as shown in Table 6.5. This is accomplished by calculating the number of basic operations performed in each model as a function of the size of input, n (the number of elements in the data structure or the size of the problem), and selecting the worst-case scenario. The PCA exhibits lower computational complexity than the kPCA. This can be attributed to the computational cost of the kernel operation involved in the kPCA feature extraction process.

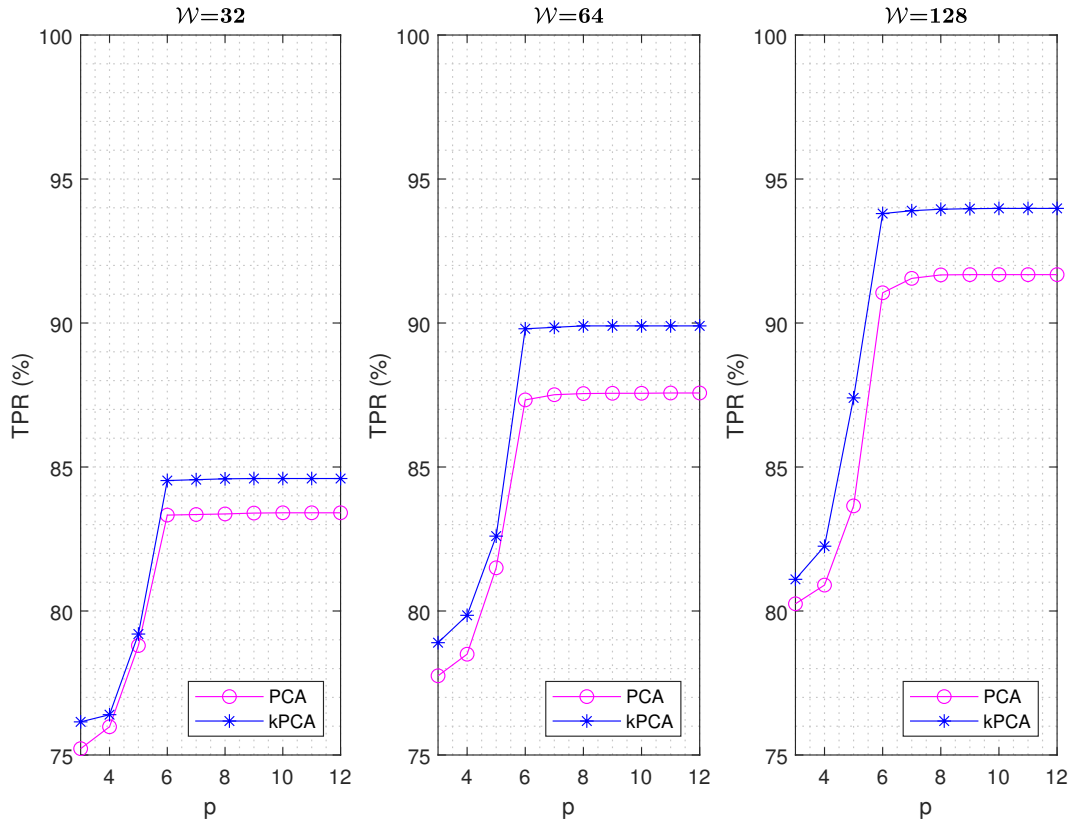
TABLE 6.5: Worst-case time complexity analysis.

Steps	Big-$\mathcal{O}$ notation	
	PCA	kPCA
$\mathbf{X}$-data formation	$\mathcal{O}(n)$	$\mathcal{O}(n)$
Kernel	-	$\mathcal{O}(n)$
Eigendecomposition	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$
SVD	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$
Feature vectors formation	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$

The kernel trick is introduced to manage the non-linear subspace that may exist in the datasets. However, the high computational cost of the kPCA-HMM led to better performance than the PCA-HMM. Therefore, there is a trade-off between computational complexity and improved performance. A general comparison of the two models, PCA-HMM and kPCA-HMM, as shown in Figures 6.1 and 6.2, confirms that the kPCA-HMM outperformed the PCA-HMM. The superior performance of kPCA-HMM can be attributed to kPCA's capacity to locate non-linear subspaces within the datasets.

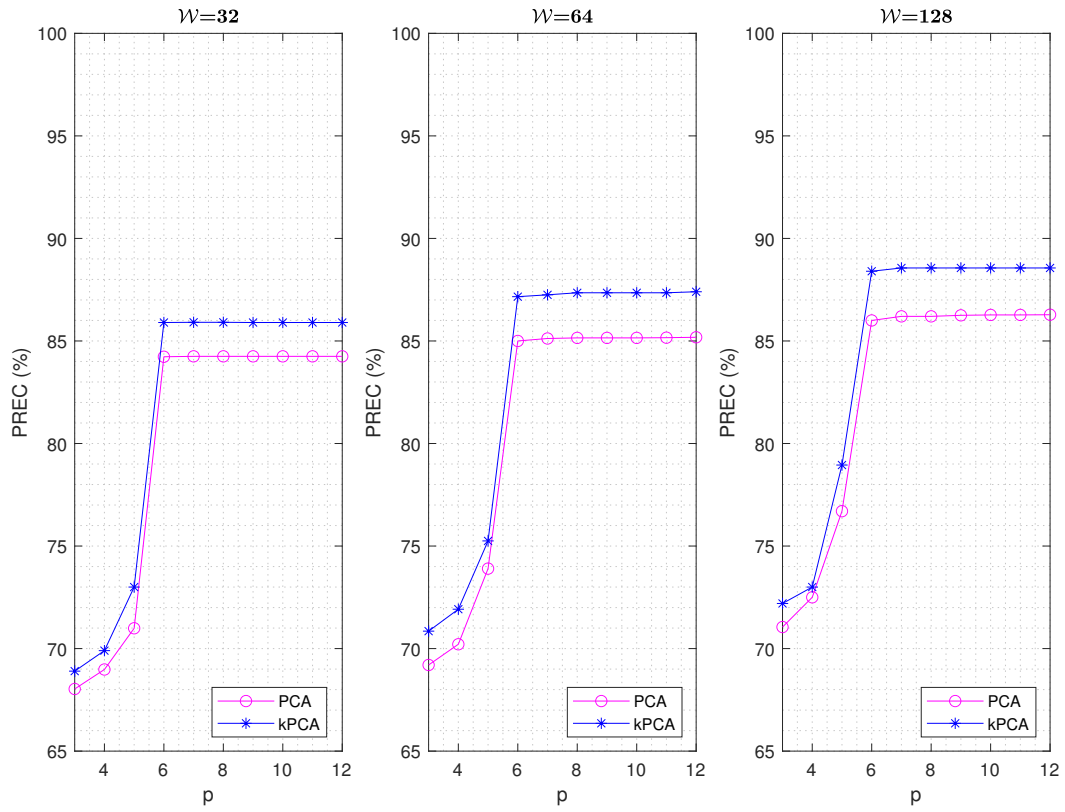


(a) Southern right whale

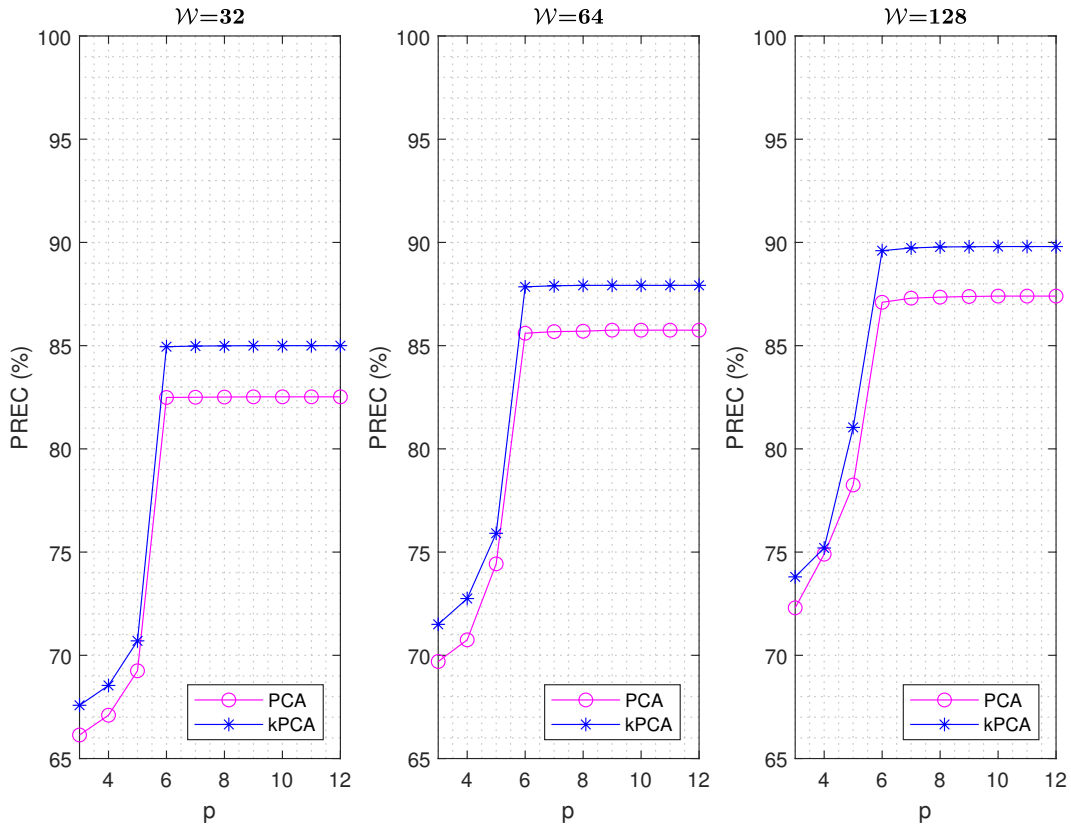


(b) Humpback whale

FIGURE 6.1: TPR performance for different p.

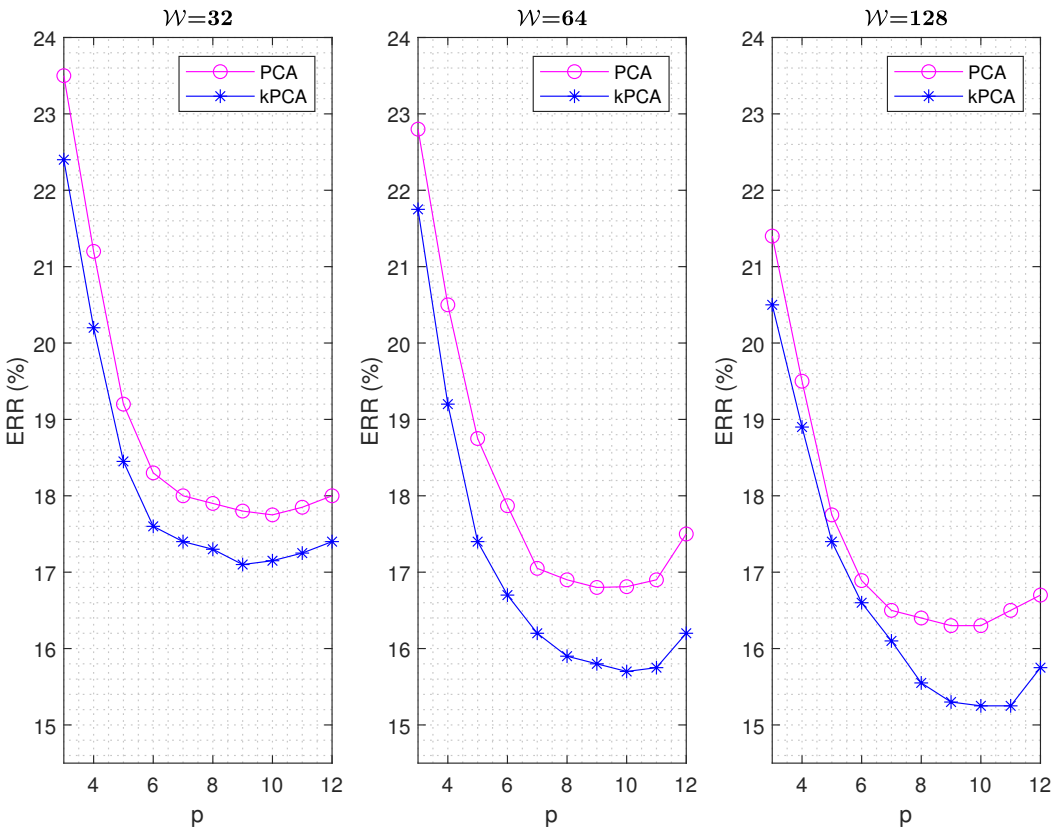


(a) Southern right whale

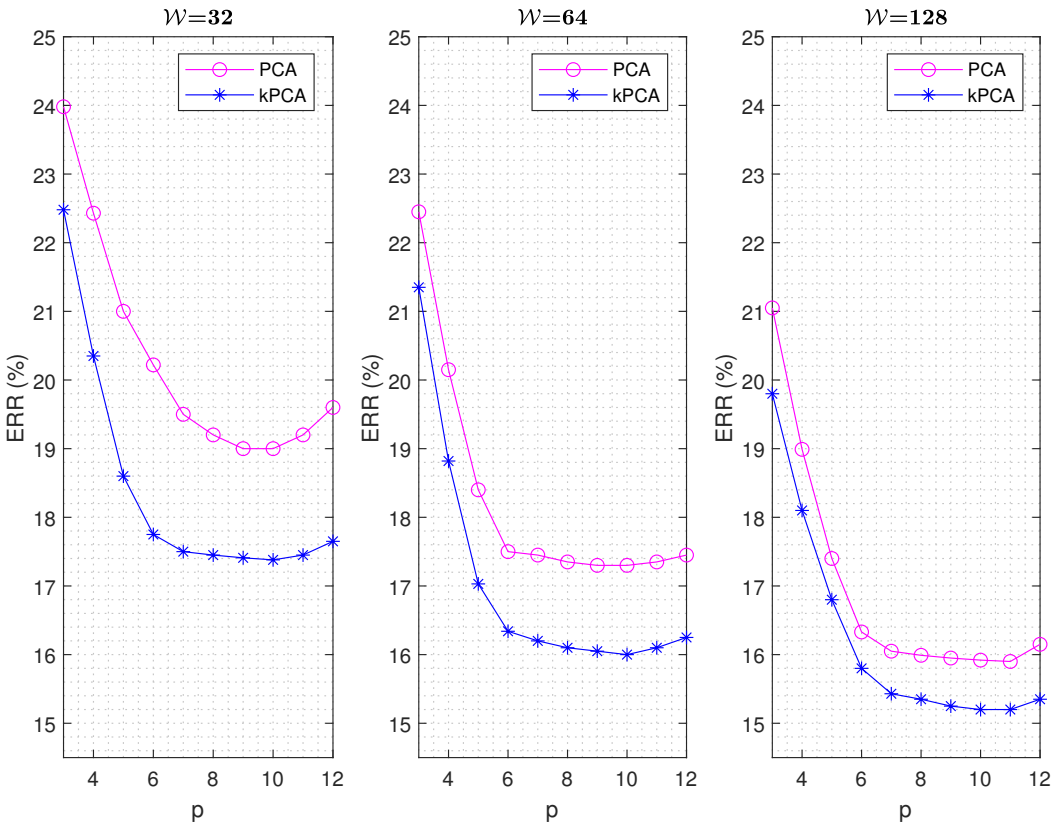


(b) Humpback whale

FIGURE 6.2: PREC performance for different p.



(a) Southern right whale



(b) Humpback whale

FIGURE 6.3: ERR performance for different p.

The following conclusions have been drawn based on the analysis of the results obtained from the experiments conducted on PCA-HMM and kPCA-HMM. The models are p-dimensional, as determined by the number of PCs chosen for each simulation. The numbers of samples, $\mathcal{W}$, were varied per window size. The performance of the models was evaluated based on TPR, PREC, and ERR. The performance of models varied depending on factors such as the dimension of the feature vectors, species, and number of samples. The worst-case time complexity analysis carried out on the respective FE techniques indicated that the kPCA exhibits slightly increased computational complexity compared to the PCA. However, the kPCA-HMM outperformed the PCA-HMM, thus justifying the slight high complexity in the kPCA. Therefore, there is a trade-off between computational complexity and improved performance. A general comparison of the two models shows that the performance of kPCA-HMM is better when compared to the respective performance of PCA-HMM. This is due to kPCA's ability to find the non-linear subspaces that exist in the datasets. In general, it was discovered that the structure of feature vectors has a significant impact on model performance.

6.3 Results and Discussion: Performance Analysis of DMD-HMM and kDMD-HMM Models

In this section, we analyse and discuss the results obtained from the experiments conducted on DMD-HMM and kDMD-HMM. Specifically, experiments were carried out for both the proposed DMD-HMM and kDMD-HMM across window sizes $\mathcal{W} = 32, 64$, and 128, with various values of p. It is important to emphasise that the values of $\mathcal{W}$ represent the number of samples in each snapshot, and the models are p-dimensional. This implies that the higher the values of p, the more complex the computational load of the models. Hence, various simulations were run by varying the values of p for selected $\mathcal{W}$ to experimentally confirm the effect of $\mathcal{W}$ with respect to different p.

The performance of the models for different values of p at $\mathcal{W} = 32, 64, 128$ is shown in Tables 6.6–6.8. The values of p represent the dimension of the feature vector for the models, which has a significant impact on the load complexity and performance of each model. The performance with respect to TPR and PREC improved gradually as the values of p increased from 3 to 4, and there were significant improvements from 6 to 8 as displayed in Figures 6.4 and 6.5. However, it can be seen that there are no more improvements from 8 to 12; rather, there is a gradual increase in ERR from $p = 11$, as shown in Figure 6.6. Thus, a further increase in the value of p from 8

TABLE 6.6: Simulation results for different p at $\mathcal{W} = 32$.

	SRW vocalisations							HW vocalisations						
p	TPR (%)		PREC (%)		ERR (%)		p	TPR (%)		PREC (%)		ERR (%)	
	DMD	kDMD	DMD	kDMD	DMD	kDMD		DMD	kDMD	DMD	kDMD	DMD	kDMD
3	75.61	75.63	67.44	67.50	24.22	23.77	3	76.70	76.72	65.91	65.94	24.95	23.77
4	75.94	75.96	68.13	68.20	21.44	19.62	4	76.92	76.99	66.85	66.89	23.44	22.90
5	76.99	77.65	70.99	73.00	20.05	18.00	5	77.55	79.35	69.34	72.01	22.65	21.85
6	83.99	85.19	81.41	82.99	19.22	17.01	6	84.99	86.87	78.31	80.24	21.22	20.04
7	87.78	89.60	87.28	89.41	18.48	16.00	7	88.78	90.89	85.44	88.67	19.45	18.69
8	92.99	94.50	91.63	93.28	17.25	14.80	8	94.25	95.99	89.06	91.88	18.20	16.99
9	92.99	94.51	91.62	93.28	17.25	14.80	9	94.23	96.00	89.05	91.88	17.65	17.02
10	93.00	94.50	91.62	93.29	17.29	15.01	10	94.25	95.98	89.06	91.88	17.65	17.03
11	92.99	94.50	91.63	93.29	17.95	15.55	11	94.25	95.99	89.06	91.87	17.85	17.10
12	92.99	94.50	91.62	93.27	18.60	15.98	12	94.23	95.99	89.05	91.88	17.95	17.40

TABLE 6.7: Simulation results for different p at $\mathcal{W} = 64$.

	SRW vocalisations							HW vocalisations						
p	TPR (%)		PREC (%)		ERR (%)		p	TPR (%)		PREC (%)		ERR (%)	
	DMD	kDMD	DMD	kDMD	DMD	kDMD		DMD	kDMD	DMD	kDMD	DMD	kDMD
3	76.34	76.36	68.99	69.00	23.40	22.30	3	78.33	78.32	67.04	67.06	23.97	21.88
4	76.99	77.00	69.10	69.19	20.45	19.50	4	78.52	78.54	68.90	68.90	21.90	20.40
5	77.90	79.00	73.80	75.05	19.65	17.80	5	80.01	81.67	72.05	74.45	20.17	19.45
6	86.05	87.92	83.55	85.09	18.75	16.85	6	86.16	87.35	79.90	83.45	19.05	18.50
7	90.35	91.65	89.90	90.41	18.00	15.45	7	90.98	91.99	88.25	89.96	17.80	17.07
8	94.77	96.40	92.89	94.35	16.99	13.99	8	95.05	96.99	91.09	92.55	17.35	16.10
9	94.77	96.40	92.89	94.35	17.00	13.99	9	95.05	96.98	91.08	92.55	17.38	16.10
10	94.77	96.41	92.89	94.35	17.01	14.01	10	95.06	96.98	91.10	92.55	17.40	16.18
11	94.77	96.42	92.89	94.36	17.55	14.45	11	95.05	96.99	91.10	92.56	17.65	16.45
12	94.77	96.44	92.89	94.35	17.95	14.95	12	95.05	96.99	91.10	92.56	17.82	16.85

only increases the computational burden of the models without necessarily improving performance. Although a reduced ERR is observed as the value of p increases, this gain is reversed from $p = 11$. This indicates that the models level off on all metrics at a certain point. Besides, it shows there is a trade-off between reducing the ERR

and improving the TPR and PREC. Consequently, the optimal dimension for the

TABLE 6.8: Simulation results for different p at $\mathcal{W} = 128$.

SRW vocalisations							HW vocalisations						
p	TPR (%)		PREC (%)		ERR (%)		p	TPR (%)		PREC (%)		ERR (%)	
	DMD	kDMD	DMD	kDMD	DMD	kDMD		DMD	kDMD	DMD	kDMD	DMD	kDMD
3	78.95	78.99	70.21	70.20	21.80	20.99	3	80.06	80.10	69.70	69.88	21.65	20.41
4	79.08	79.30	71.70	71.75	19.85	18.80	4	80.99	81.02	70.00	70.10	21.00	19.70
5	81.99	83.45	74.60	76.99	18.65	17.60	5	82.85	84.98	73.55	75.99	19.60	19.05
6	86.65	88.03	84.39	87.21	17.87	16.55	6	87.11	88.99	81.90	85.46	18.20	17.70
7	90.65	92.95	91.01	92.86	17.05	14.67	7	92.66	93.91	90.45	91.60	16.45	16.02
8	95.85	97.72	94.50	96.19	16.30	13.90	8	97.28	98.95	92.99	94.85	15.99	15.44
9	95.85	97.72	94.50	96.20	15.80	12.99	9	97.28	98.94	93.00	94.85	16.00	14.75
10	95.85	97.73	94.50	96.19	15.81	13.00	10	97.25	98.95	92.99	94.85	16.03	14.75
11	95.85	97.73	94.49	96.19	16.00	13.05	11	97.28	98.95	92.99	94.85	16.20	14.80
12	95.85	97.72	94.49	96.19	16.75	13.95	12	97.28	98.95	92.99	94.85	16.55	15.25

detectors will be at $p = 8$, because it offers a compromise between the computational complexity and the performance of the detectors. Besides, it can be seen that, at smaller p values ($p = 3, 4$), the performance of both detectors is almost the same. However, the kDMD-HMM model exhibits better performance than the DMD-HMM model as the values of p increase, particularly from $p = 5$. This can be attributed to a more accurate and faster approximation of $\mathbf{Z}$ using the kernel method.

We analysed the results further by comparing the respective metric performances of species to species. One key observation is that the models exhibit higher PREC performance for SRW in comparison to the HW PREC performance. For instance, when $p = 8$ and $\mathcal{W} = 128$, the kDMD-HMM achieved a 96.19% PREC for SRW and 94.85% for HW. Similarly, under the same condition ($p = 8$ and $\mathcal{W} = 128$), the DMD-HMM achieved a 94.50% PREC for SRW and 92.99% for HW. This difference can be attributed to the ratio of whale vocalisations to noise in the datasets. As can be noted from Figure 5.1, there are fewer samples of SRW vocalisations when compared to Figure 5.3, which has more samples of HW vocalisations within the same time duration. The higher PREC values on the SRW dataset indicate that our models can more precisely detect a larger proportion of SRW vocalisations present in the dataset. It is worth noting that the PREC metric is a measure of the proportion of correctly detected whale vocalisations to the total number of positive detections (true positives and false positives) in the datasets. A higher value PREC suggests

a greater assurance of detecting most of the vocalisations of interest by the models. Therefore, models with high PREC performance are less likely to predict noise as vocalisations of interest. This metric (PREC) is suited for evaluating the performance of models with respect to the fraction of correct detection of vocalisations of interest in a dataset with sparse numbers of data samples in a recording session, such as the SRW dataset used in this research. The lower PREC in HW in comparison with SRW is an indication that there are more false positive detections on the HW dataset. We can attribute this to the many samples of calls in the HW dataset. This implies that there are both more correctly detected HW vocalisations (true positives) and, inescapably, more mistakenly detected vocalisations (false positives) in the HW dataset.

On the other hand, the models exhibit higher TPR performance for the HW species when compared with the SRW TPR performance. For instance, when $p = 8$ and $\mathcal{W} = 128$, the kDMD-HMM achieved a 98.95% TPR for HW and 97.72% for SRW. Similarly, under the same conditions ($p = 8$ and $\mathcal{W} = 128$), the DMD-HMM achieved a 97.28% TPR for HW and 95.85% for SRW. This difference can be attributed to the presence of many samples of HW vocalisations in the datasets. The high TPR values are confirmation of the good sensitivity state of the models, which shows that the models are less likely to miss vocalisations of interest. In the case of the HW dataset, which contains many samples of vocalisations, the higher TPR is advantageous. Thus, the models with high TPR performance are well suited for detecting vocalisations of species in a dataset with many vocalisation samples, where a high priority is likely placed on avoiding missing vocalisations of interest. A paired t-test was also carried out to compare the performance of DMD-HMM and kDMD-HMM. The t-test

TABLE 6.9: Paired t-test results to compare the performance of PCA-HMM and kPCA-HMM on SRW and HW vocalisations at a significance level of 0.05.

	$\boldsymbol{p}$-values		
	TPR	PREC	ERR
SRW	1.6033×10^{-4}	3.7575×10^{-4}	4.1077×10^{-5}
HW	2.3375×10^{-4}	5.6244×10^{-4}	3.4867×10^{-5}

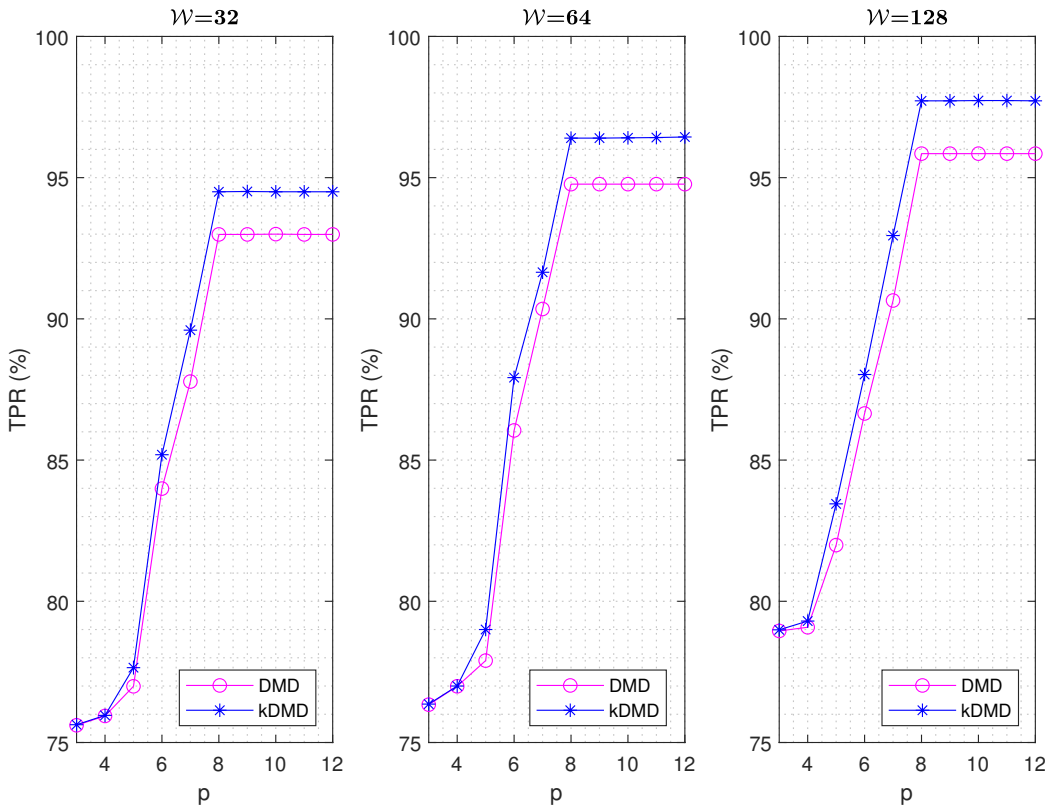
findings, as displayed in Table 6.9, for $\mathcal{W} = 128$, indicated that the kDMD-HMM performed significantly better than the DMD-HMM ($\boldsymbol{p} < 0.05$). These findings demonstrate that the superior performance of the kDMD-HMM over the DMD-HMM is not merely due to sampling variation.

The complexity of the FE algorithms is investigated to further examine the computational cost of the models. The worst-case time complexity analysis for each algorithm is carried out using the big-$\mathcal{O}$ notation as shown in Table 6.10. The kDMD exhibits lower computational complexity than the DMD since the kernel trick is deployed to compute the modes, which eliminates the need for the SVD operation. The kDMD-HMM did not only exhibit lower computational complexity; it also outperformed the DMD-HMM.

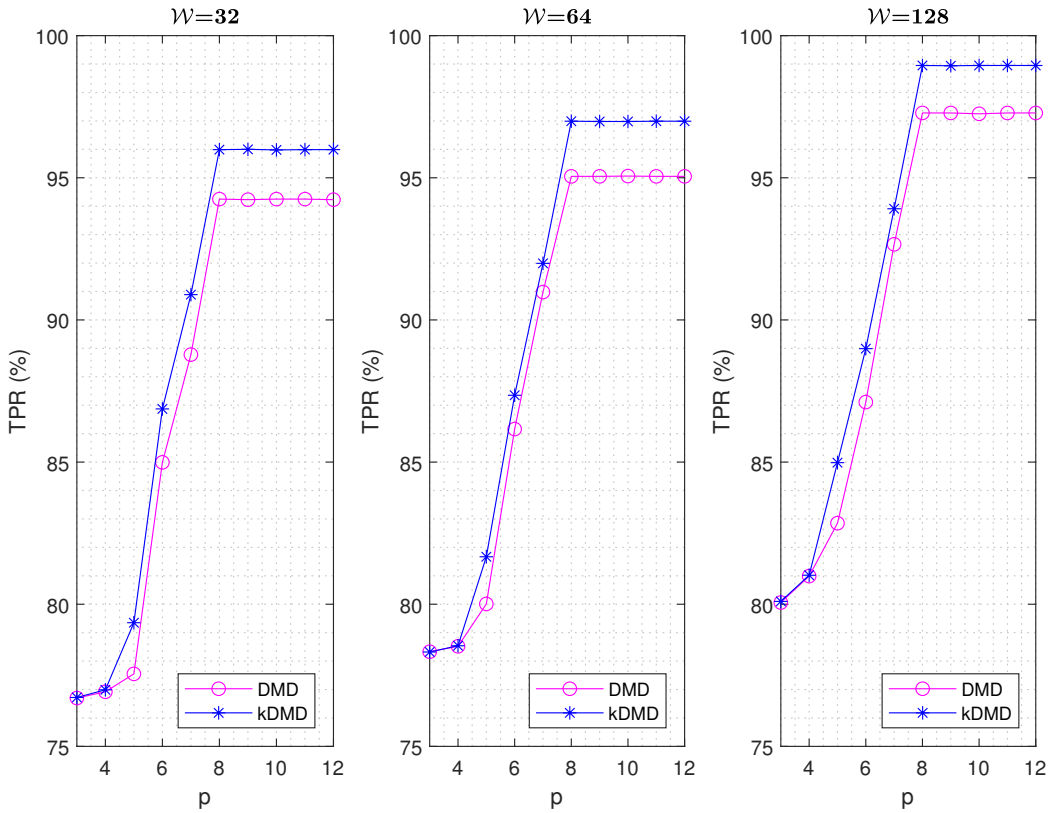
TABLE 6.10: Worst-case time complexity analysis.

Steps	Big-$\mathcal{O}$ notation	
	DMD	kDMD
$\mathbf{X}$-data formation	$\mathcal{O}(n)$	$\mathcal{O}(n)$
SVD	$\mathcal{O}(n^3)$	-
Kernel	-	$\mathcal{O}(n)$
Matrix operation	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$
$\tilde{\boldsymbol{M}}$-operation	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$
Feature vectors formation	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$

In general, a comparison of the two models shows that kDMD-HMM exhibits better performance compared to the respective DMD-HMM model outputs, as shown in Figures 6.4 and 6.5. In addition, the ERR in the DMD-HMM model is higher when compared to the kDMD-HMM model, as shown in Figure 6.6. The kernel method reduces the computational load of the process of deriving the modes from the data [173]. Nonetheless, the performance of the two models at optimal $p = 8$ falls within the acceptable level for real-time detection applications with TPR/PREC $> 90\%$. Moreover, the overall performance of $\mathcal{W} = 128$ is better than $\mathcal{W} = 32$ and 64. This shows that training with more samples means there are better intrinsic features present in the datasets. This is confirmed in the work of Pace *et al.* in [19].

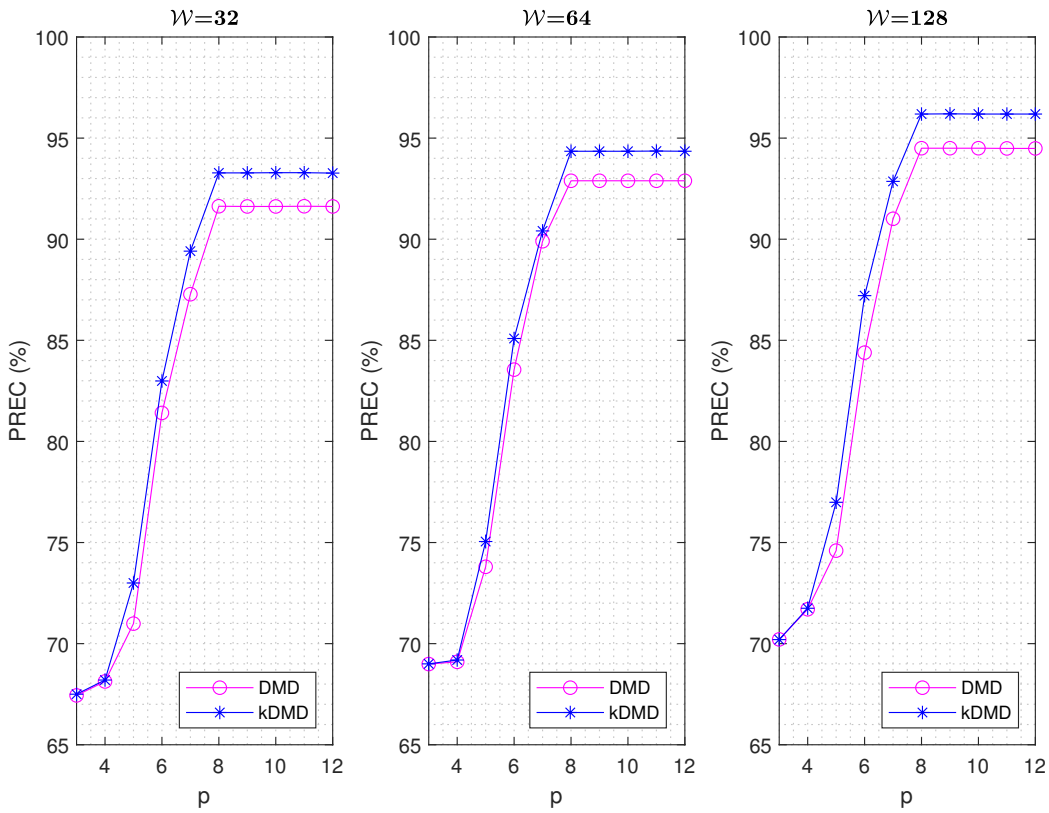


(a) Southern right whale

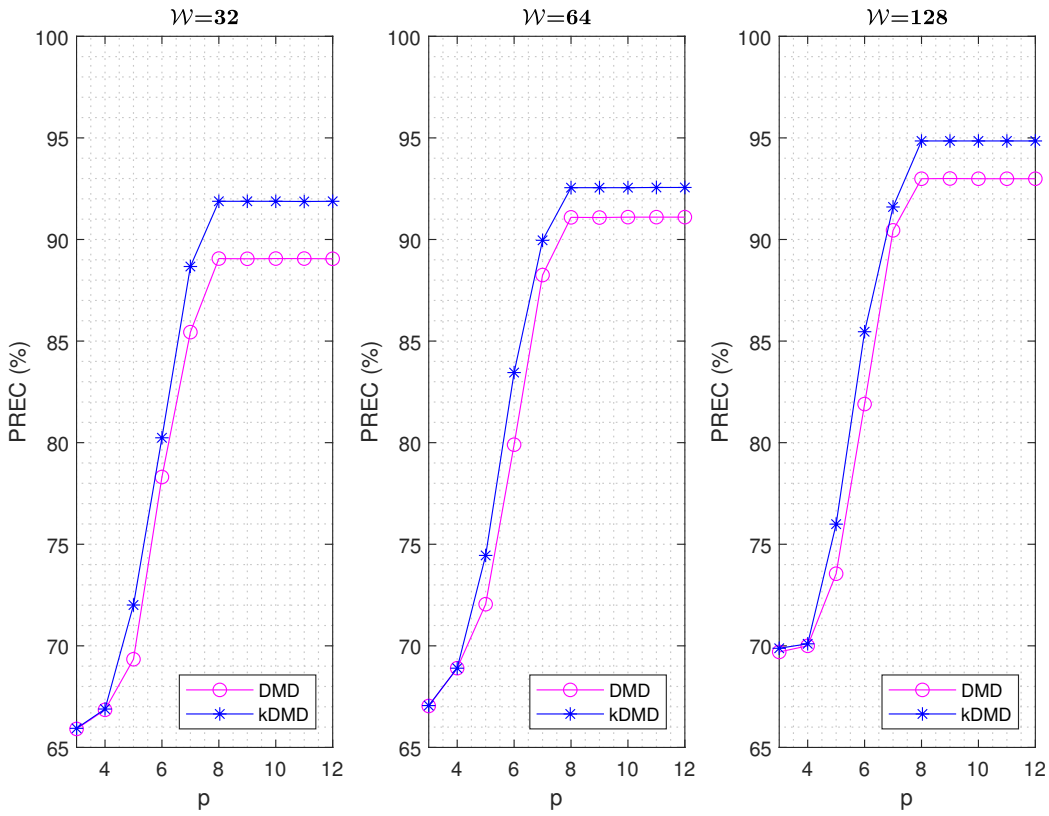


(b) Humpback whale

FIGURE 6.4: TPR performance for different p.

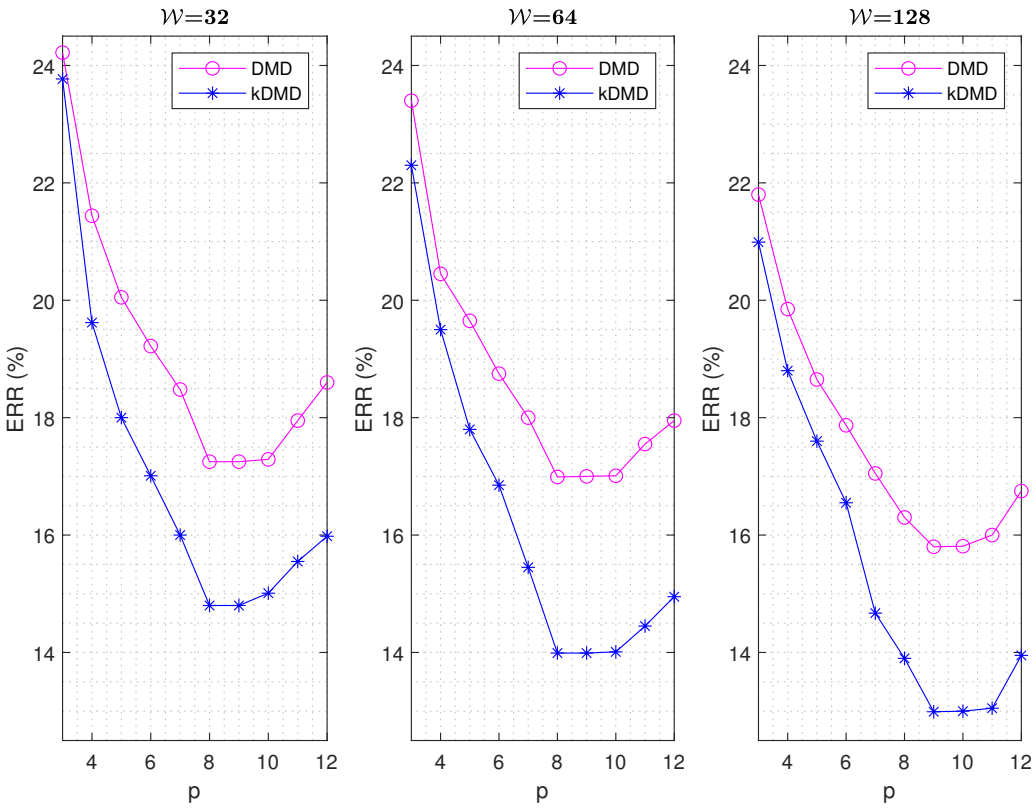


(a) Southern right whale

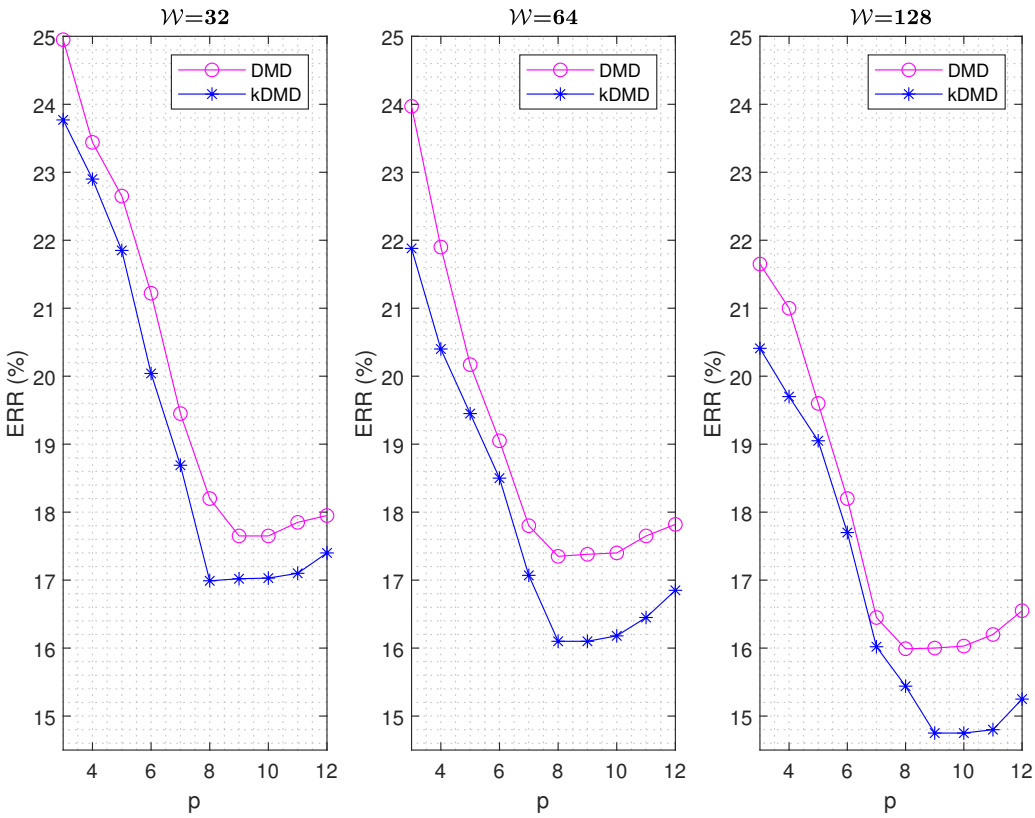


(b) Humpback whale

FIGURE 6.5: PREC performance for different p.



(a) Southern right whale



(b) Humpback whale

FIGURE 6.6: ERR performance for different p.

The transformation of modes derived from DMD into feature vectors for HMM provides more opportunity for extracting features from PAM data. The SVD is central to the mode derivation process in the DMD, which contributes substantially to the computational load of the DMD. Thus, a kernel version of DMD was introduced to find a more efficient way of computing DMD without explicitly running the SVD algorithm. This innovative approach streamlines the mode derivation process. Thereafter, the modes were transformed as feature vectors and adapted with HMM for the detection of whale vocalisations. The performance of the DMD-HMM and the kDMD-HMM models was tested on two species of whale: SRW and HW. The results from the experiments showed good PREC and TPR performance and low ERR. The kDMD-HMM model did not only outperform the DMD-HMM, it also exhibited lower computational complexity, as confirmed from the big-$\mathcal{O}$ notation analysis. Remarkably, the novel introduction of kernel method into DMD to compute the modes gives dual benefits of enhanced efficiency and effectiveness. In the subsequent sections, we carried out detailed comparison analyses of the performance of the models developed in this study. Besides, a comparison study is conducted between the developed ED-HMMs and the existing FE techniques used with HMM in the literature for the detection of whale vocalisations.

6.4 Performance Comparison of ED-HMMs

The performance of the developed eigendecomposition-based hidden Markov models (ED-HMMs): PCA-HMM, kPCA-HMM, DMD-HMM, and kDMD-HMM is compared in this section. The developed ED-HMMs are grouped according to the underlying ED algorithm deployed for the development of each model. They are grouped as either PCs-based hidden Markov models (PC-HMMs) or dynamic modes-based hidden Markov models (DM-HMMs). The analyses of the results in Sections 6.2 and 6.3 showed that the performance of models varies based on different conditions. Therefore, the remaining part of this section shows the performance comparison and analysis of the developed models. The performance metrics are given in terms of true

positive rate (TPR), precision (PREC), error rate (ERR), F_1 scores, and the error bar plots on the F_1 scores.

The simulation results of the ED-HMMs for different p values are shown in Tables 6.11–6.13. In terms of TPR and PREC results, the PC-HMMs is observed to perform better than the DM-HMMs from $p = 3$ to $p = 6$, since the PCs are hierarchically ordered according to their relevance in the data structure. This implies

TABLE 6.11: Performance comparison of ED-HMMs for different p at $\mathcal{W}$=32.

SRW Vocalisations												
p	TPR (%)				PREC (%)				ERR (%)			
	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD
3	75.89	76.21	75.61	75.63	68.03	68.90	67.44	67.50	23.50	22.40	24.22	23.77
4	76.29	76.98	75.94	75.96	68.98	69.90	68.13	68.20	21.20	20.20	21.44	19.62
5	79.78	80.01	76.99	77.65	70.99	73.00	70.99	73.00	19.20	18.45	20.05	18.00
6	84.00	85.49	83.99	85.19	84.23	85.90	81.41	82.99	18.30	17.60	19.22	17.01
7	84.01	85.50	87.78	89.60	84.25	85.91	87.28	89.41	18.00	17.40	18.48	16.00
8	84.01	85.50	92.99	94.50	84.25	85.91	91.63	93.28	17.90	17.30	17.25	14.80
9	84.01	85.50	92.99	94.51	84.25	85.90	91.62	93.28	17.80	17.10	17.25	14.80
10	84.02	85.51	93.00	94.50	84.25	85.90	91.62	93.29	17.75	17.15	17.29	15.01
11	84.02	85.52	92.99	94.50	84.25	85.90	91.63	93.29	17.85	17.25	17.95	15.55
12	84.01	85.53	92.99	94.50	84.25	85.90	91.62	93.27	18.00	17.40	18.60	15.98

HW Vocalisations												
p	TPR (%)				PREC (%)				ERR (%)			
	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD
3	75.22	76.15	76.70	76.72	66.14	67.58	65.91	65.94	23.98	22.48	24.95	23.77
4	75.98	76.40	76.92	76.99	67.10	68.54	66.85	66.89	22.43	20.35	23.44	22.90
5	78.80	79.20	77.55	79.35	69.25	70.70	69.34	72.01	21.00	18.60	22.65	21.85
6	83.33	84.53	84.99	86.87	82.49	84.95	78.31	80.24	20.22	17.75	21.22	20.04
7	83.35	84.56	88.78	90.89	82.50	84.98	85.44	88.67	19.50	17.50	19.45	18.69
8	83.37	84.59	94.25	95.99	82.51	84.99	89.06	91.88	19.20	17.45	18.20	16.99
9	83.40	84.60	94.23	96.00	82.52	85.00	89.05	91.88	19.00	17.41	17.65	17.02
10	83.41	84.60	94.25	95.98	82.52	85.00	89.06	91.88	19.00	17.38	17.65	17.03
11	83.41	84.60	94.25	95.99	82.52	85.00	89.06	91.87	19.20	17.45	17.85	17.10
12	83.41	84.60	94.23	95.99	82.52	85.00	89.05	91.88	19.60	17.65	17.95	17.40

that the first few PCs are sufficient to project the features in the original data. As a result, the PC-HMMs level off in terms of performance at $p = 6$, and therefore, further increasing the value of p only introduces more complexity to the models without really improving their performance. Thus, the PC-HMMs are 6-dimensional models.

TABLE 6.12: Performance comparison of ED-HMMs for different p at $\mathcal{W}$=64.

SRW Vocalisations												
p	TPR (%)				PREC (%)				ERR (%)			
	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD
3	77.01	78.15	76.34	76.36	69.20	70.85	68.99	69.00	22.80	21.75	23.40	22.30
4	77.80	78.95	76.99	77.00	70.21	71.92	69.10	69.19	20.50	19.20	20.45	19.50
5	80.82	81.70	77.90	79.00	73.90	75.25	73.80	75.05	18.75	17.40	19.65	17.80
6	86.65	88.90	86.05	87.92	85.00	87.16	83.55	85.09	17.87	16.70	18.75	16.85
7	86.75	89.00	90.35	91.65	85.12	87.25	89.90	90.41	17.05	16.20	18.00	15.45
8	86.85	89.00	94.77	96.40	85.15	87.35	92.89	94.35	16.90	15.90	16.99	13.99
9	86.85	89.10	94.77	96.40	85.15	87.35	92.89	94.35	16.80	15.80	17.00	13.99
10	86.85	89.00	94.77	96.41	85.15	87.35	92.89	94.35	16.81	15.70	17.01	14.01
11	86.86	89.00	94.77	96.42	85.16	87.35	92.89	94.36	16.90	15.75	17.55	14.45
12	86.86	89.00	94.77	96.44	85.18	87.40	92.89	94.35	17.50	16.20	17.95	14.95

HW Vocalisations												
p	TPR (%)				PREC (%)				ERR (%)			
	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD
3	77.75	78.90	78.33	78.32	69.70	71.50	67.04	67.06	22.45	21.35	23.97	21.88
4	78.50	79.85	78.52	78.54	70.75	72.75	68.90	68.90	20.15	18.82	21.90	20.40
5	81.50	82.60	80.01	81.67	74.43	75.91	72.05	74.45	18.40	17.03	20.17	19.45
6	87.33	89.80	86.16	87.35	85.60	87.85	79.90	83.45	17.50	16.34	19.05	18.50
7	87.51	89.85	90.98	91.99	85.68	87.90	88.25	89.96	17.45	16.20	17.80	17.07
8	87.55	89.90	95.05	96.99	85.70	87.92	91.09	92.55	17.35	16.10	17.35	16.10
9	87.56	89.90	95.05	96.98	85.75	87.92	91.08	92.55	17.30	16.05	17.38	16.10
10	87.56	89.90	95.06	96.98	85.75	87.92	91.10	92.55	17.30	16.00	17.40	16.18
11	87.57	89.90	95.05	96.99	85.75	87.92	91.10	92.56	17.35	16.10	17.65	16.45
12	87.57	89.90	95.05	96.99	85.75	87.92	91.10	92.56	17.45	16.25	17.82	16.85

On the other hand, increasing the dimension of the models from $p = 7$ to $p = 12$ indicates that the DM-HMMs outperform the PC-HMMs. This implies that more modes are selected to achieve enhanced performance at the cost of computational load. Besides, the results show that the DM-HMMs level off in terms of performance at $p = 8$, and a further increment beyond $p = 8$ only adds to the computational load of the DM-HMMs. Rather, a gradual increase in ERR is observed. Hence, the DM-HMMs are 8-dimensional models. As previously stated in Sections 6.2 and 6.3, the computational load of a model increases with every increase in p. Note that the models reach their respective points of stabilisation in terms of performance at the same p value irrespective of the window size, $\mathcal{W}$ as shown in Figures 6.7–6.9. In light of this, it is concluded that the PC-HMMs are less computationally complex than the DM-HMMs. However, the DM-HMMs exhibit better performance than the

TABLE 6.13: Performance comparison of ED-HMMs for different p at $\mathcal{W}$=128.

SRW Vocalisations												
p	TPR (%)				PREC (%)				ERR (%)			
	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD
3	79.40	80.50	78.95	78.99	71.05	72.20	70.21	70.20	21.40	20.50	21.80	20.99
4	80.25	81.70	79.08	79.30	72.50	73.00	71.70	71.75	19.50	18.90	19.85	18.80
5	83.00	85.85	81.99	83.45	76.70	78.95	74.60	76.99	17.75	17.40	18.65	17.60
6	90.50	92.25	86.65	88.03	86.00	88.40	84.39	87.21	16.89	16.60	17.87	16.55
7	91.31	92.40	90.65	92.95	86.20	88.56	91.01	92.86	16.50	16.10	17.05	14.67
8	91.34	92.45	95.85	97.72	86.20	88.56	94.50	96.19	16.40	15.55	16.30	13.90
9	91.34	92.44	95.85	97.72	86.25	88.56	94.50	96.20	16.30	15.30	15.80	12.99
10	91.35	92.48	95.85	97.73	86.27	88.56	94.50	96.19	16.30	15.25	15.81	13.00
11	91.35	92.50	95.85	97.73	86.27	88.56	94.49	96.19	16.50	15.25	16.00	13.05
12	91.35	92.50	95.85	97.72	86.28	88.56	94.49	96.19	16.70	15.75	16.75	13.95

HW Vocalisations												
p	TPR (%)				PREC (%)				ERR (%)			
	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD
3	80.25	81.10	80.06	80.10	72.30	73.80	69.70	69.88	21.05	19.80	21.65	20.41
4	80.90	82.25	80.99	81.02	74.90	75.20	70.00	70.10	18.99	18.10	21.00	19.70
5	83.65	87.40	82.85	84.98	78.25	81.04	73.55	75.99	17.40	16.80	19.60	19.05
6	91.50	93.80	87.11	88.99	87.10	89.60	81.90	85.46	16.33	15.80	18.20	17.70
7	91.55	93.90	92.66	93.91	87.30	89.73	90.45	91.60	16.05	15.43	16.45	16.02
8	91.67	93.95	97.28	98.95	87.35	89.78	92.99	94.85	15.99	15.35	15.99	15.44
9	91.68	93.97	97.28	98.94	87.38	89.79	93.00	94.85	15.95	15.25	16.00	14.75
10	91.68	93.98	97.25	98.95	87.40	89.80	92.99	94.85	15.92	15.20	16.03	14.75
11	91.68	93.98	97.28	98.95	87.40	89.80	92.99	94.85	15.90	15.20	16.20	14.80
12	91.68	93.98	97.28	98.95	87.40	89.80	92.99	94.85	16.15	15.35	16.55	15.25

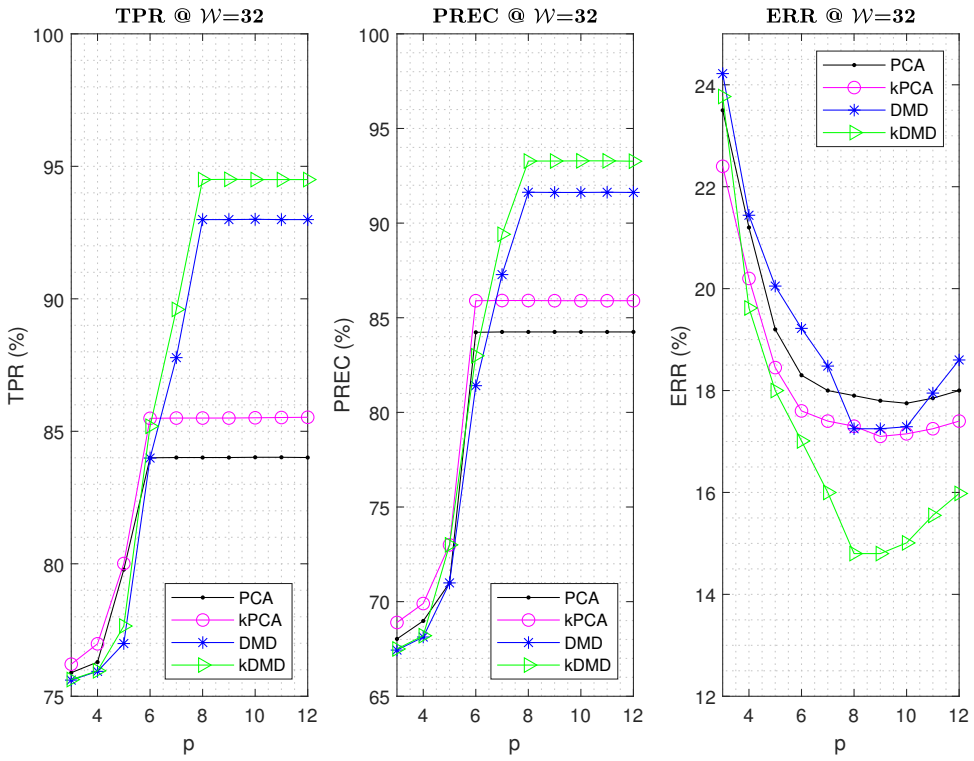
PC-HMMs, albeit at a higher computational cost. Hence, when choosing a model for detection, a trade-off must be made between the computational load and the performance outputs.

The experimental results are compared for the vocalisations of SRW species and HW species. This is because the vocalisations of the species are of different durations. The SRW species vocalises for short durations, with a mean duration of 0.65 seconds, in comparison to HW species, which vocalises for longer durations, with a mean duration of 2.20 seconds. Therefore, the performance of the models is considered at different $\mathcal{W}$, as shown in Tables 6.11–6.13. At $\mathcal{W} = 32$, the results show that SRW species exhibits superior TPR and PREC performance for PC-HMMs in comparison to the corresponding PC-HMMs' TPR and PREC performance for HW species. However, as $\mathcal{W}$ increases, the gain made by SRW species is reversed. When $\mathcal{W} = 64$ and

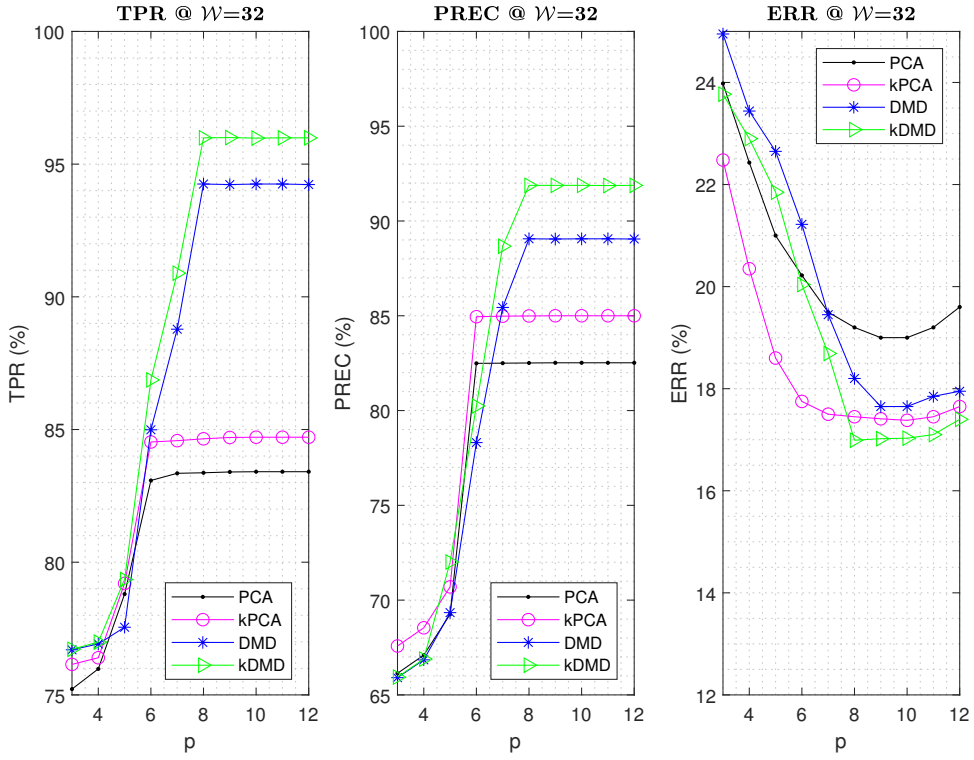
$\mathcal{W} = 128$ respectively, HW species demonstrates better TPR and PREC performance for PC-HMMs when compared to the corresponding PC-HMMs' TPR and PREC performance of SRW species. To illustrate this scenario, in Table 6.11, at $\mathcal{W} = 32$, $p = 6$, for kPCA-HMM, SRW species demonstrates 0.96% TPR and 0.95% PREC performance gain over HW species. In contrast, in Table 6.12, at $\mathcal{W} = 64$, $p = 6$, for kPCA-HMM, HW species exhibits a 0.90% TPR gain with a 0.36% reduced ERR over SRW species. Similarly, in Table 6.13, HW species demonstrates 1.65% TPR improvement with a 1.09% reduced ERR over SRW species. In contrast, for the DM-HMMs, the HW species performs better than the SRW for all $\mathcal{W}$, with respect to TPR, while the SRW species performs better than the HW species for all $\mathcal{W}$, with respect to PREC. The performance, however, continues to rise steadily as $\mathcal{W}$ increases. Moreover, a close examination of the results reveals that the differences in TPR and PREC for the PC-HMMs are small when compared to the differences in TPR and PREC for the DM-HMMs. For example, in Table 6.13, for kPCA-HMM, at $\mathcal{W} = 128$, $p = 6$, the difference between TPR and PREC is 4% while the difference between TPR and PREC for kDMD-HMM is 5.49%.

Therefore, in terms of PREC, it can be concluded that the PC-HMMs are suitable for the detection of whale species with short vocalisation duration. On the other hand, in terms of TPR, the DM-HMMs are suitable for the detection of whale species with long vocalisation duration. Besides, regardless of the value of $\mathcal{W}$, for both SRW and HW species, PC-HMMs outperform the corresponding DM-HMMs in terms of TPR and PREC from $p = 3$ to $p = 6$. Besides, PC-HMMs also exhibit low ERR when compared to corresponding DM-HMMs irrespective of $\mathcal{W}$ value. The performance gains for TPR/PREC and low ERR at small p-values ($p = 3, 4, 5, 6$) attest to PCA's ability to represent a high-dimensional dataset with a few dimensions that contain the most statistically illustrative features of the original data without losing important information. In addition, the PREC metric gives the precise whale vocalisations in the data. Therefore, PC-HMMs can be used to analyse noisy datasets that contain few whale vocalisations, such as the SRW dataset used in this study.

Furthermore, the F_1 score was computed to evaluate the overall detection rates of

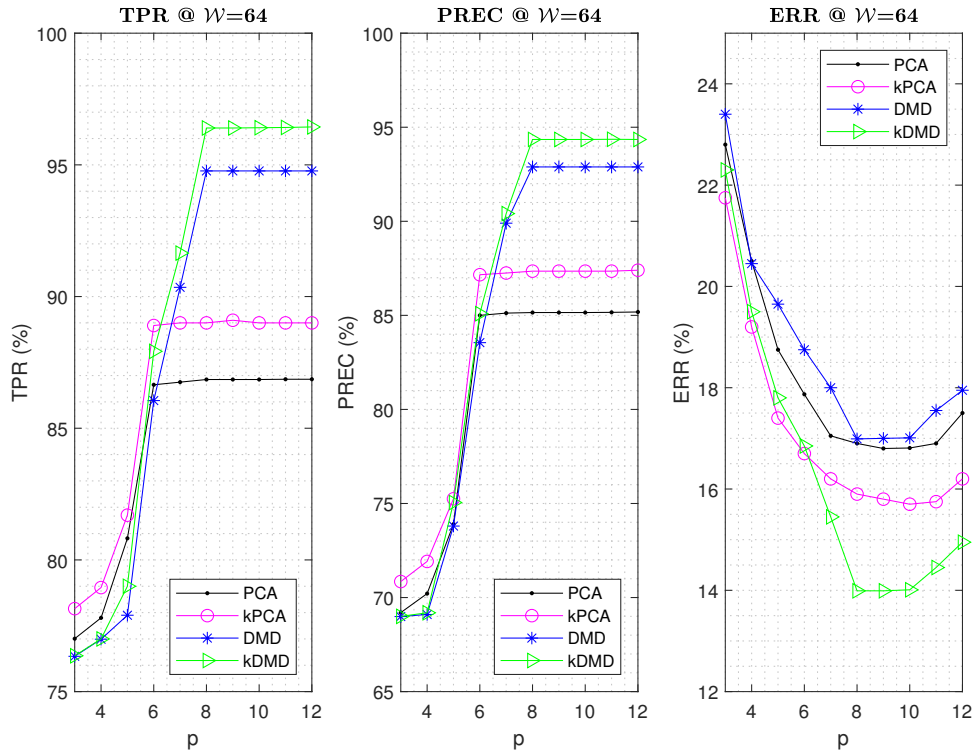


(a) Southern right whale

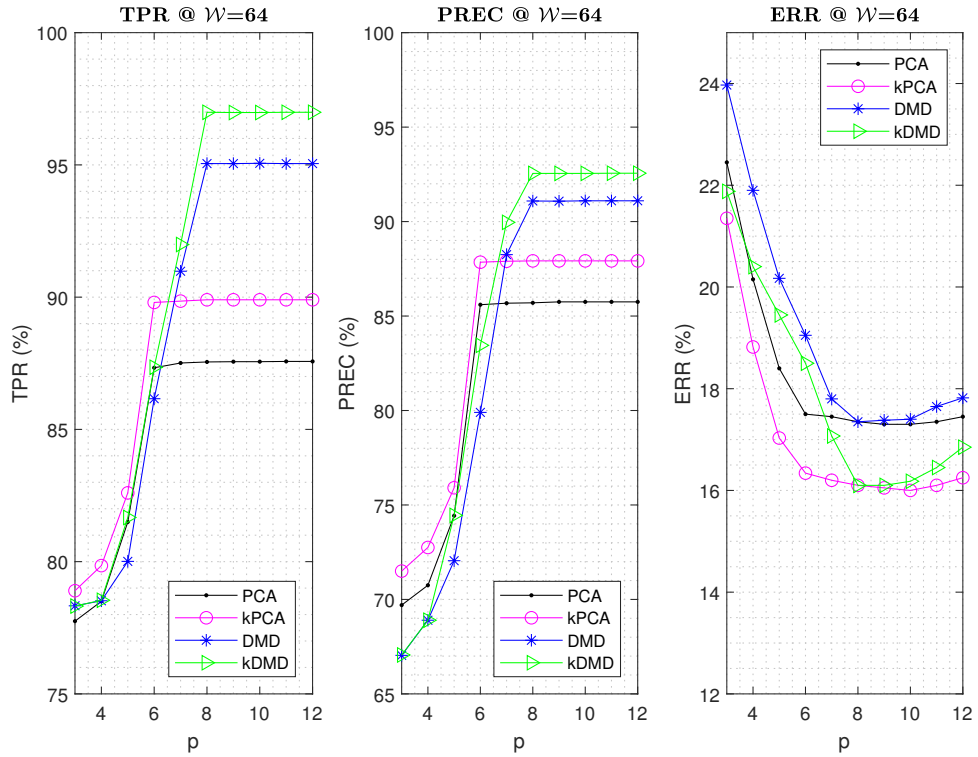


(b) Humpback whale

FIGURE 6.7: Performance comparison ED-HMMs for different p at $\mathcal{W}$=32.

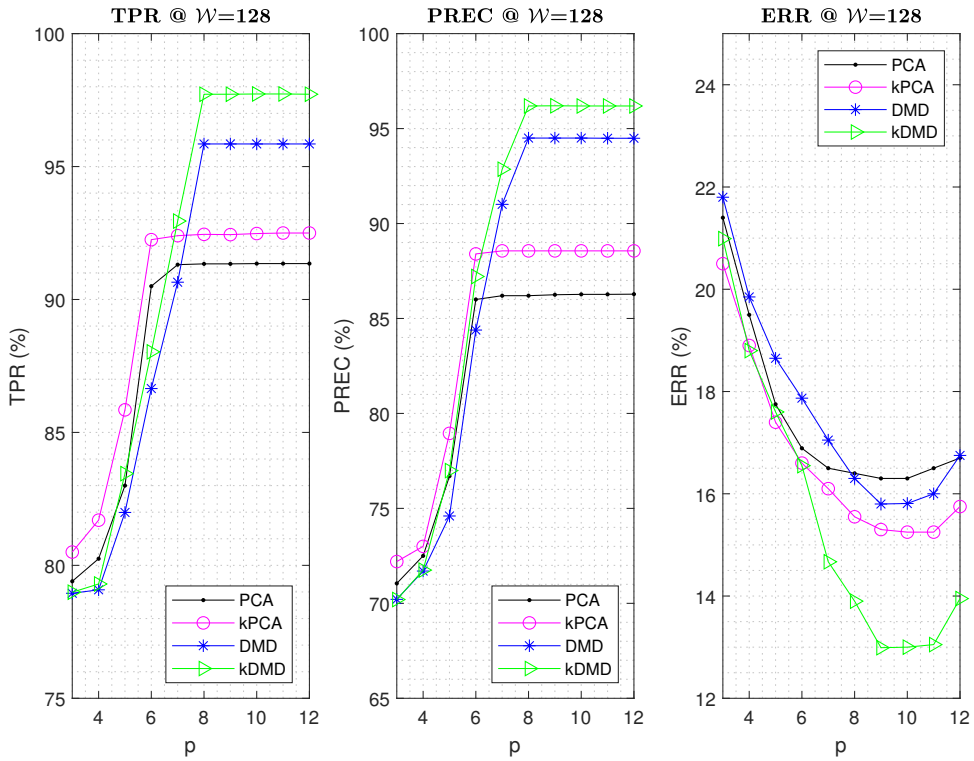


(a) Southern right whale

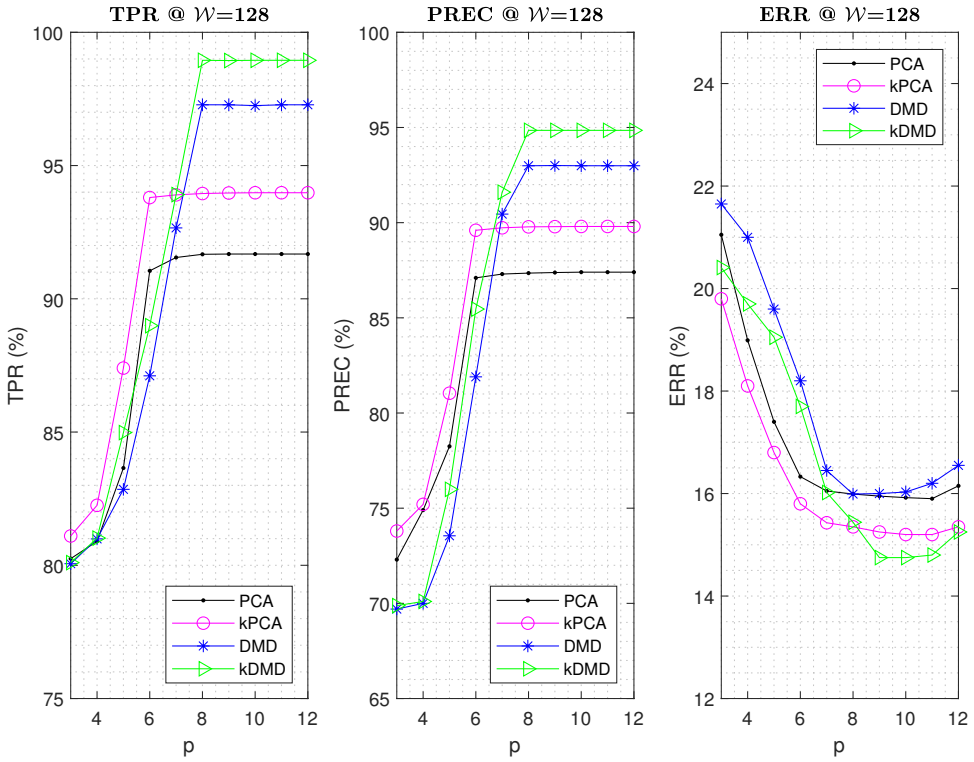


(b) Humpback whale

FIGURE 6.8: Performance comparison ED-HMMs for different p at $\mathcal{W}$=64.



(a) Southern right whale



(b) Humpback whale

FIGURE 6.9: Performance comparison ED-HMMs for different p at $\mathcal{W}$=128.

the models. The F_1 score is the harmonic mean between precision (PREC) and recall (TPR) that reveals a balanced trade-off between the positive predictions and the overall correctness of the positive predictions. The optimal F_1 scores for the respective models occurred at the same p, that is, at $p = 6$ for PC-HMMs and at $p = 8$ for DM-HMMs, as shown in Figure 6.10. The F_1 scores of kDMD-HMM are closely matched for both species, irrespective of the value of $\mathcal{W}$. However, for the rest of the models (PCA-HMM, kPCA-HMM, and DMD-HMM), there are noticeable differences in the F_1 scores, except for DMD-HMM at $\mathcal{W} = 128$, where the F_1 scores are closely matched for both SRW and HW. Specifically, from the F_1 scores at $\mathcal{W} = 128$ shown in Table 6.14, the F_1 scores of the respective DMD-HMM and kDMD-HMM for SRW and HW are almost the same. For example, considering the DMD-HMM, at $p = 8$, the F_1 score for SRW is 0.9517, and for HW, it is 0.9509. Similarly, for kDMD-HMM, at $p = 8$, the F_1 score for SRW is 0.9695 and for HW, it is 0.9698. On the contrary,

TABLE 6.14: F_1 score for SRW and HW vocalisations

F_1 Score								
p	SRW Vocalisations				HW Vocalisations			
	PCA	kPCA	DMD	kDMD	PCA	kPCA	DMD	kDMD
3	0.7499	0.7612	0.7432	0.7434	0.7607	0.7728	0.7452	0.7464
4	0.7618	0.7711	0.7521	0.7534	0.7778	0.7857	0.7510	0.7517
5	0.7973	0.8226	0.7812	0.8009	0.8086	0.8410	0.7792	0.8023
6	0.8819	0.9028	0.8551	0.8762	0.8903	0.9165	0.8442	0.8719
7	0.8868	0.9044	0.9083	0.9290	0.8937	0.9177	0.9154	0.9274
8	0.8870	0.9046	0.9517	0.9695	0.8946	0.9182	0.9509	0.9686
9	0.8872	0.9046	0.9517	0.9695	0.8948	0.9183	0.9509	0.9685
10	0.8874	0.9048	0.9517	0.9695	0.8949	0.9184	0.9507	0.9686
11	0.8874	0.9049	0.9517	0.9695	0.8949	0.9184	0.9509	0.9686
12	0.8874	0.9049	0.9517	0.9695	0.8949	0.9184	0.9509	0.9686

for PC-HMMs, the F_1 scores differ across the respective PC-HMMs for both species. For example, for PCA-HMM, at $p = 6$, the F_1 score for SRW is 0.8819, while for HW, it is 0.8903. Similarly, with kPCA-HMM at $p = 6$, the F_1 score for SRW is 0.9028, and for HW, it is 0.9165. Overall, the F_1 results show that for kDMD-HMM, there is consistency in terms of balancing PREC and TPR across the species, irrespective of the number of samples. The same assertion can be made for DMD-HMM at

$\mathcal{W} = 128$. These observations confirm the robustness of DM-HMMs' performance, particularly the kDMD-HMM, therefore indicating that it is not overly influenced by the variation in each dataset. Conversely, the differing F_1 scores for PC-HMMs suggest that these models are sensitive to the unique characteristics of each dataset.

Error bars were incorporated into the F_1 scores to gain insights into the variability associated with the F_1 scores and assess the confidence level of the results at each value of p. The plots, displayed in Figure 6.11, reveal notable variability in the F_1 scores across different runs for PC-HMMs applied to both SRW and HW vocalisations. Remarkably, the plots for ED-HMMs demonstrate consistent stability in F_1 scores for SRW vocalisations across different runs. This observation suggests that the ED-HMMs remain relatively consistent across different runs for vocalisations with shorter durations, as in the case with SRW. Consequently, they exhibit more stability and greater reliability in their performance on short-duration calls. On the other hand, the ED-HMMs seem to be sensitive for long-duration calls, as in the case of the HW, while the PC-HMMs demonstrate sensitivity to both short and long-duration calls.

6.5 Performance Comparison of the Proposed ED-HMMs with the Existing FE Techniques used with HMM

In this section, the performance of the developed ED-HMMs: PCA-HMM, kPCA-HMM, DMD-HMM, and kDMD-HMM is compared with the existing feature extraction (FE) techniques used with HMM in the literature for the detection of whale vocalisations. These FE techniques are the linear predictive coefficient-based hidden Markov model (LPC-HMM) [19], the mel-scale frequency cepstral coefficient-based hidden Markov model (MFCC-HMM) [20], the relative mean, amplitude, and power-based hidden Markov model (MAP-HMM) [25], and the empirical mode

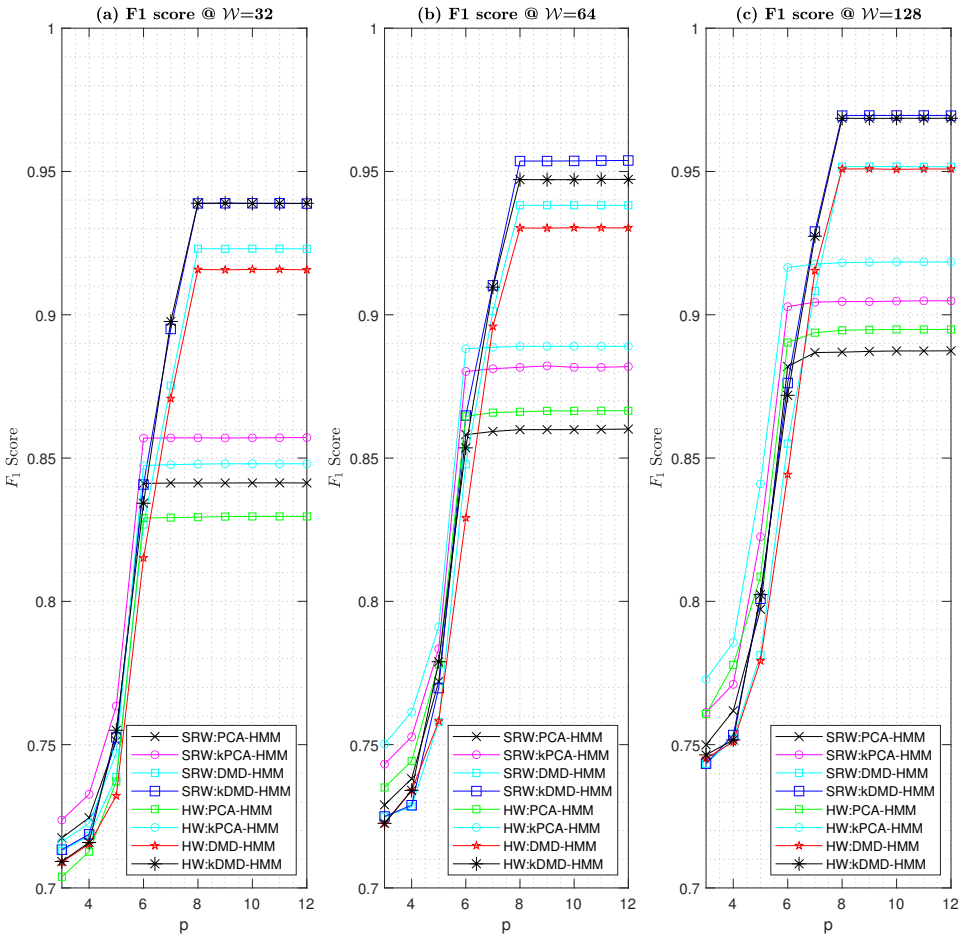


FIGURE 6.10: F_1 score for ED-HMMs for SRW and HW vocalisations.

decomposition-based hidden Markov model (EMD-HMM) [22]. The performance measurement metrics for comparison are TPR, PREC, and ERR.

From Tables 6.15–6.17, it can be observed that the models level off at different p values. Unlike the PC-HMMs and the DM-HMMs that converge at $p = 6$ and $p = 8$ respectively, the LPC-HMM and the MFCC-HMM converge at $p = 12$ while the MAP-HMM and the EMD-HMM level off at $p = 10$ and $p = 9$ respectively. Thus, the existing FE-HMMs offer high computational complexity when compared to the proposed ED-HMMs. Although the EMD-HMM is a 9-dimensional model, the model exhibits reduced performance when compared to the ED-HMMs. Therefore, different models could offer almost the same computational cost, but the performance

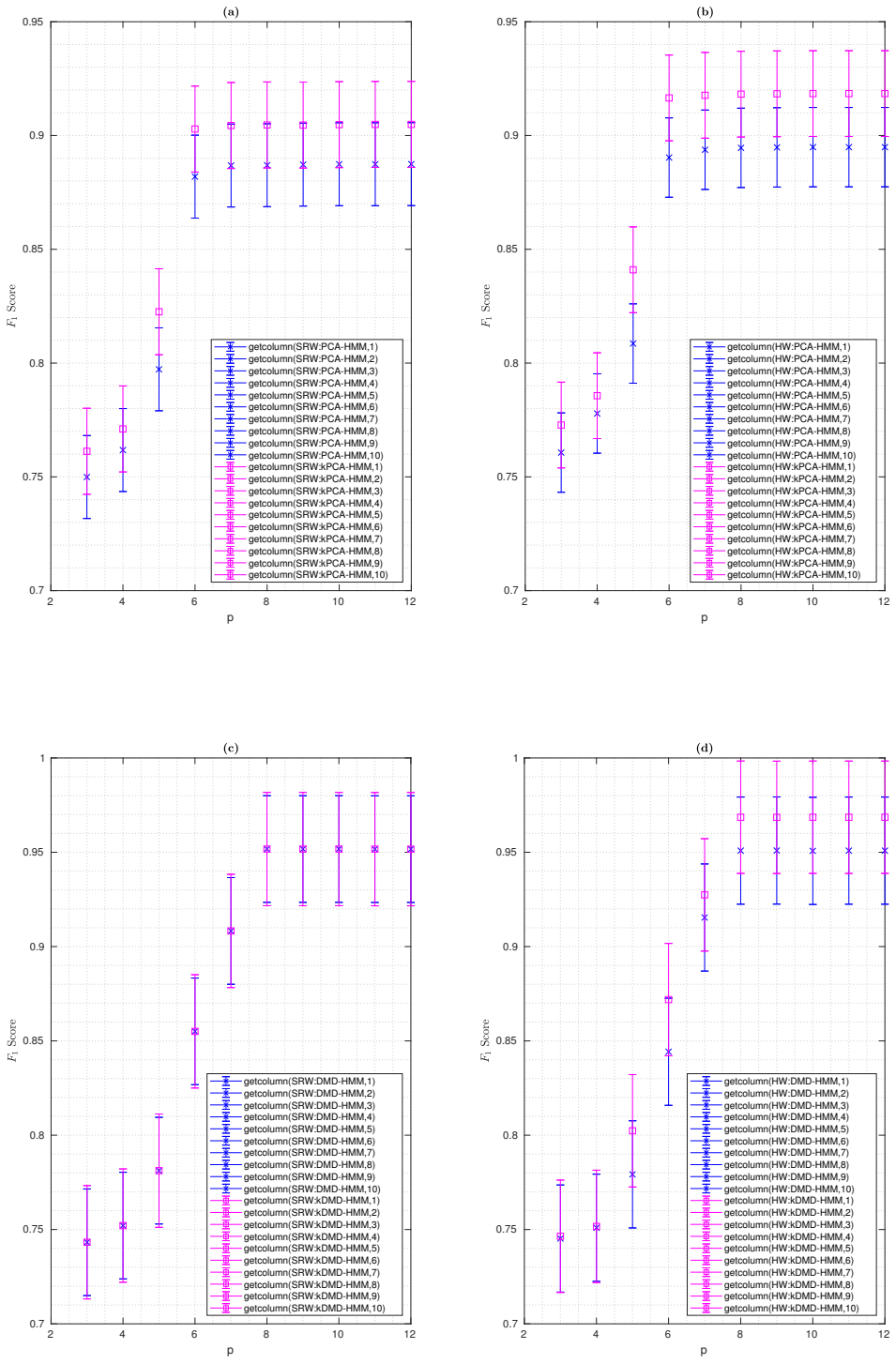


FIGURE 6.11: F_1 scores plots with error bars for ED-HMMs depict the relationship between the number of p and the corresponding F_1 scores for the respective models at $\mathcal{W} = 128$. The error bars provide a representation of the variability in F_1 scores observed across different runs. The plots are categorised as follows: (a) PC-HMMs applied to SRW vocalisations; (b) PC-HMMs applied to HW vocalisations; (c) DM-HMMs applied to SRW vocalisations; and (d) DM-HMMs applied to HW vocalisations.

TABLE 6.15: TPR performance comparison for different HMMs at $\mathcal{W}$=128.

SRW Vocalisations								
p	TPR (%)							
	LPC	MFCC	MAP	EMD	PCA	kPCA	DMD	kDMD
3	75.27	75.66	76.99	77.45	79.40	80.50	78.95	78.99
4	76.10	76.59	77.85	78.69	80.25	81.70	79.08	79.30
5	77.85	77.99	79.49	81.23	83.00	85.85	81.99	83.45
6	79.43	79.44	81.09	84.09	90.50	92.25	86.65	88.03
7	81.11	81.89	83.89	86.59	91.31	92.40	90.65	92.95
8	82.97	83.88	86.95	88.90	91.34	92.45	95.85	97.72
9	84.12	85.67	87.95	90.95	91.34	92.44	95.85	97.72
10	85.11	86.52	87.95	90.95	91.35	92.48	95.85	97.73
11	86.32	87.25	87.95	90.95	91.35	92.50	95.85	97.73
12	86.32	87.25	87.95	90.95	91.35	92.50	95.85	97.72

HW Vocalisations								
p	TPR (%)							
	LPC	MFCC	MAP	EMD	PCA	kPCA	DMD	kDMD
3	75.29	75.71	76.01	76.99	80.25	81.10	80.06	80.10
4	76.09	76.67	78.05	79.05	80.90	82.25	80.99	81.02
5	77.88	78.04	79.99	81.45	83.65	87.40	82.85	84.98
6	79.35	79.74	81.53	83.65	91.50	93.80	87.11	88.99
7	81.43	81.99	83.63	85.59	91.55	93.90	92.66	93.91
8	83.01	83.98	85.85	87.93	91.67	93.95	97.28	98.95
9	84.17	85.65	87.20	89.92	91.68	93.97	97.28	98.94
10	85.29	86.53	88.20	89.92	91.68	93.98	97.25	98.95
11	86.55	87.47	88.20	89.92	91.68	93.98	97.28	98.95
12	86.55	87.47	88.20	89.92	91.68	93.98	97.28	98.95

nonetheless may not be the same.

The ED-HMMs outperform the existing FE-HMMs in terms of TPR and PREC while offering low ERR as displayed in Tables 6.15–6.17, which is further depicted in Figure 6.12. For SRW species, EMD-HMM performance can be seen to be close to PCA-HMM's performance as regard sensitivity.

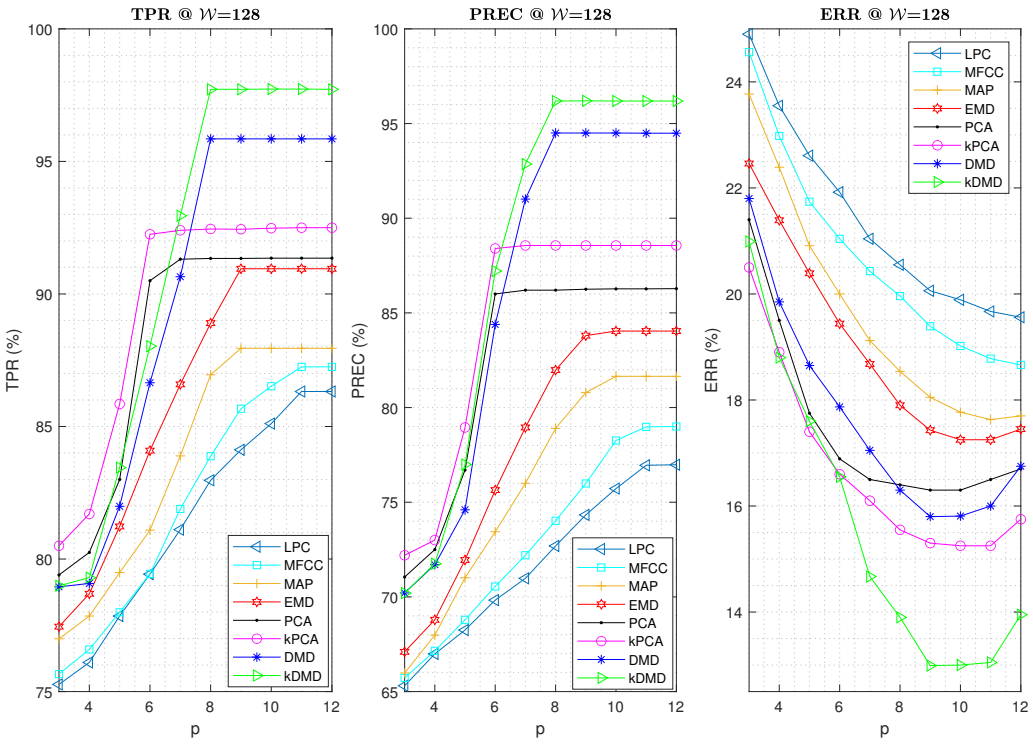
Table 6.15 shows the TPR performance of each model. The ED-HMMs outperform the existing FE-HMMs. The TPR performance of the LPC-HMM and the

TABLE 6.16: PREC performance comparison for different HMMs at $\mathcal{W}$=128.

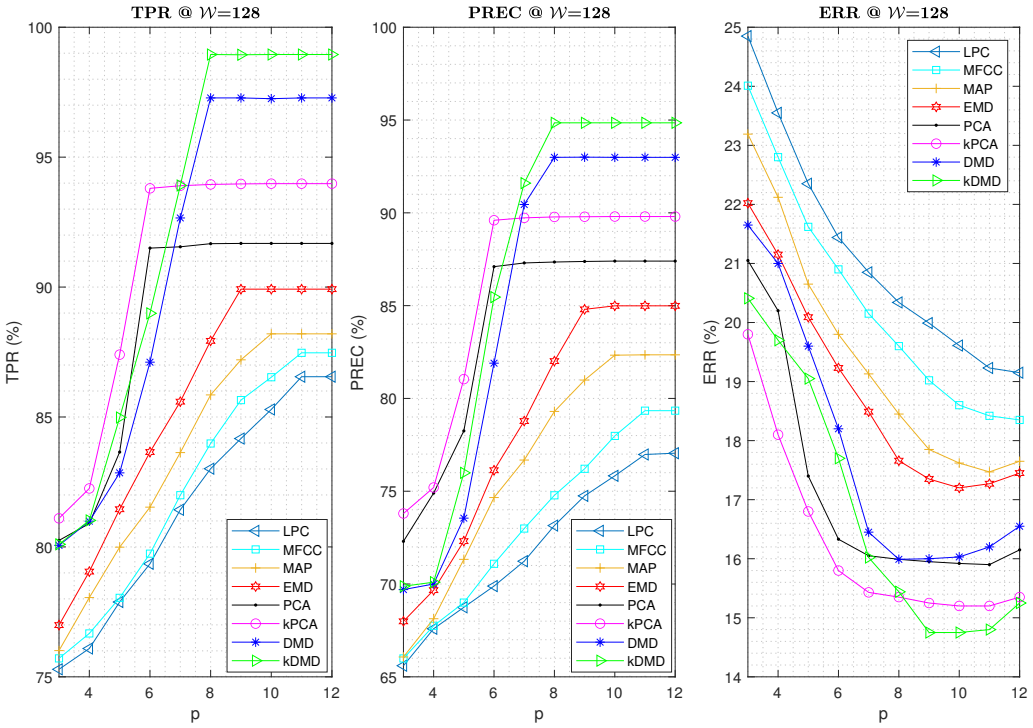
SRW Vocalisations								
p	PREC (%)							
	LPC	MFCC	MAP	EMD	PCA	kPCA	DMD	kDMD
3	65.31	65.72	65.98	67.10	71.05	72.20	70.21	70.20
4	66.99	67.15	67.98	68.79	72.50	73.00	71.70	71.75
5	68.25	68.79	71.01	71.95	76.70	78.95	74.60	76.99
6	69.83	70.55	73.44	75.64	86.00	88.40	84.39	87.21
7	70.98	72.20	76.00	78.95	86.20	88.56	91.01	92.86
8	72.69	74.02	78.90	81.98	86.20	88.56	94.50	96.19
9	74.33	75.99	80.79	83.80	86.25	88.56	94.50	96.20
10	75.72	78.26	81.65	84.04	86.27	88.56	94.50	96.19
11	76.95	78.98	81.65	84.04	86.27	88.56	94.49	96.19
12	76.98	79.00	81.65	84.04	86.28	88.56	94.49	96.19

HW Vocalisations								
p	PREC (%)							
	LPC	MFCC	MAP	EMD	PCA	kPCA	DMD	kDMD
3	65.60	65.99	66.07	67.99	72.30	73.80	69.70	69.88
4	67.59	67.75	68.13	69.67	74.90	75.20	70.00	70.10
5	68.75	69.00	71.33	72.31	78.25	81.04	73.55	75.99
6	69.89	71.09	74.66	76.13	87.10	89.60	81.90	85.46
7	71.23	72.99	76.68	78.78	87.30	89.73	90.45	91.60
8	73.15	74.78	79.30	82.01	87.35	89.78	92.99	94.85
9	74.75	76.21	80.99	84.80	87.38	89.79	93.00	94.85
10	75.83	77.98	82.33	84.99	87.40	89.80	92.99	94.85
11	76.98	79.34	82.35	84.99	87.40	89.80	92.99	94.85
12	77.05	79.34	82.35	84.99	87.40	89.80	92.99	94.85

MFCC-HMM are almost the same for both SRW species and HW species. In contrast, the TPR performance are quite different for the MAP-HMM for both SRW species and HW species. The ED-HMMs do not only offer low computational complexity but superior TPR performance over the existing FE techniques used with HMM for the detection of whale vocalisations.



(a) Southern right whale



(b) Humpback whale

FIGURE 6.12: Performance comparison for different HMMs.

TABLE 6.17: ERR performance comparison for different HMMs at $\mathcal{W}$=128.

SRW Vocalisations								
p	ERR (%)							
	LPC	MFCC	MAP	EMD	PCA	kPCA	DMD	kDMD
3	24.90	24.56	23.77	22.46	21.40	20.50	21.80	20.99
4	23.55	22.98	22.39	21.39	19.50	18.90	19.85	18.80
5	22.61	21.74	20.91	20.39	17.75	17.40	18.65	17.60
6	21.92	21.04	20.00	19.44	16.89	16.60	17.87	16.55
7	21.04	20.43	19.12	18.68	16.50	16.10	17.05	14.67
8	20.55	19.96	18.54	17.90	16.40	15.55	16.30	13.90
9	20.06	19.39	18.05	17.43	16.30	15.30	15.80	12.99
10	19.89	19.02	17.77	17.25	16.30	15.25	15.81	13.00
11	19.67	18.78	17.63	17.25	16.50	15.25	16.00	13.05
12	19.56	18.66	17.70	17.45	16.70	15.75	16.75	13.95

HW Vocalisations								
p	ERR (%)							
	LPC	MFCC	MAP	EMD	PCA	kPCA	DMD	kDMD
3	24.85	24.01	23.19	22.02	21.05	19.80	21.65	20.41
4	23.55	22.80	22.12	21.15	20.20	18.10	21.00	19.70
5	22.35	21.62	20.65	20.09	17.40	16.80	19.60	19.05
6	21.44	20.90	19.80	19.23	16.33	15.80	18.20	17.70
7	20.85	20.15	19.13	18.49	16.05	15.43	16.45	16.02
8	20.34	19.60	18.45	17.66	15.99	15.35	15.99	15.44
9	19.99	19.02	17.85	17.35	15.95	15.25	16.00	14.75
10	19.61	18.60	17.62	17.20	15.92	15.20	16.03	14.75
11	19.23	18.42	17.47	17.27	15.90	15.20	16.20	14.80
12	19.15	18.35	17.65	17.45	16.15	15.35	16.55	15.25

6.6 Worst-case Time Analysis

The worst-case time complexity analysis of the FE algorithms is carried out to investigate their impact on the performance of the models. The big-$\mathcal{O}$ notation is used to analyse the worst-case time complexity of the existing FE algorithms and the FE algorithms proposed in this study, as shown in Table 6.18. First, among the existing FE algorithms, the MAP exhibits the least time complexity, followed by the LPC, the MFCC, and the EMD, respectively. Second, for the ED-based algorithms, the

PC-HMMs display a lower time complexity than the DM-HMMs. Third, while it is generally assumed that algorithms with higher time complexity tend to perform worse than those with lower time complexity, it is observed that this is not exclusively the case, as indicated in the big-$\mathcal{O}$ notation results. For instance, the MAP-HMM performs better than the LPC-HMM. By the same token, the kDMD-HMM not only outperforms the rest of the ED-FE algorithms (PCA-HMM, kPCA-HMM, and DMD-HMM), but it also exhibits the least worst-case time complexity.

TABLE 6.18: Worst-case time complexity analysis.

Operations	Big-$\mathcal{O}$ notation							
	LPC	MFCC	MAP	EMD	PCA	kPCA	DMD	kDMD
$\mathbf{X}$-data formation	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$
Hamming window	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	-	-	-	-	-	-
Autocorrelation	$\mathcal{O}(n^2)$	-	$\mathcal{O}(n)$	-	-	-	-	-
FFT	-	$\mathcal{O}(n \log n)$	-	-	-	-	-	-
Matrix operation	$\mathcal{O}(n^2)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$
Extrema identification	-	$\mathcal{O}(n^2 \log n)$	-	$\mathcal{O}(n^2)$	-	-	-	-
SVD	-	-	-	-	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	-
$\tilde{\boldsymbol{M}}$-operation	-	-	-	-	-	-	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$
Kernel	-	-	-	-	-	$\mathcal{O}(n)$	-	$\mathcal{O}(n)$
Feature vectors formation	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$

6.7 Conclusion

In this chapter, we have analysed and discussed the results obtained from a series of experiments conducted on the developed ED-HMMs: PCA-HMM, kPCA-HMM, DMD-HMM, and kDMD-HMM. These models were tested on PAM datasets containing SRW and HW species. At various times during the course of the experiments, the experimental conditions were varied based on factors such as the dimensions of the feature vectors, species, and sample size. The performance of the models was evaluated based on TPR, PREC, and ERR. Also, a paired t-test was conducted to validate the superior performance of one model over the other. The results obtained from the paired t-test confirmed that the enhanced performance of the kernel version of the respective models, when compared to the primary version, was not merely because of sampling variation. The worst-case time complexity of the respective FE

techniques was computed to assess the computational cost of the models. While the kPCA exhibited a higher computational cost than the PCA, it performed better than the PCA. On the contrary, the kDMD offered both lower computational costs and better performance than the DMD. This is due to the novel introduction of the kernel method into the process of computing the modes, which sidesteps the conventional reliance on the SVD for mode computation.

Furthermore, we carried out a comparison study of the performance of HMM for the detection of whale vocalisations. First, the performance of the developed ED-HMMs was compared to further assess and evaluate their performance. These were done using the previous metrics (TPR, PREC, and ERR) as well as the introduction of the F_1 score. It was observed that the performance of the PC-HMMs stabilised faster than that of the DM-HMMs. Thus, the PC-HMMs are less computationally complex than the DM-HMMs in terms of model dimension. However, the DM-HMMs outperformed the PC-HMMs, albeit at a higher dimensional cost. In addition, it was observed that the performance of the models improved with an increase in $\mathcal{W}$. The DM-HMMs were sensitive for the detection of datasets with many samples of whale vocalisations, as in the case of HW species. On the other hand, the PC-HMMs were more appropriate for datasets with sparse samples of whale vocalisations, as in the case of SRW species. The F_1 score was introduced to confirm the reliability of the results. The use of F_1 score was inspired by the fact that it combines two important metrics, TPR and PREC, both of which account for the two error types in binary classification (FP and FN), to provide a more comprehensive evaluation of the model's performance. The analyses of the F_1 scores indicated that the DM-HMMs maintain a consistent trade-off between robustness and generalisations across the two species, in contrast to the PC-HMMs which may be sensitive to the characteristics of each dataset.

Second, the results of the developed ED-HMMs were compared with the existing FE techniques used with HMM in the literature for the detection of whale vocalisations. The ED-HMMs do outperform the existing HMM methods in the literature. The

big-$\mathcal{O}$ notation was used to analyse the worst-case time complexity of the FE techniques to assess their impacts on the performance of the models. The kDMD did not only offer the lowest worst-case time complexity amongst the FE techniques; it also outperformed the rest of the models. A general observation is that every model displays better performance with an increase in $\mathcal{W}$. Thus, a large sample size is encouraged for model training. The different experiment results showed that a model's performance must be evaluated on a species-to-species basis. It is also important that the training data be a subset of the datasets for testing, or at least use recordings from the same region [90]. This is to avoid bias that may arise from the variation that does exist between the vocalisations of the same species. Finally, the developed ED-HMMs can further be tested on other whale vocalisations to confirm their robustness. Besides, they can be explored by researchers working on the automatic detection of other vocalising animal species.

Chapter 7

Conclusion

The chapter provides a summary of the findings and limitations of the thesis. In addition, recommendations for future research are presented.

7.1 Research Summary

This thesis, entitled "On Eigendecomposition-based Algorithms as Feature Extraction Techniques used with Hidden Markov Model for the Detection of Whale Vocalisations," aims to improve the performance of hidden Markov model (HMM) for the detection of whale vocalisations. It was found from the reviewed literature that the performance of HMM depends on the quality of the feature vectors fed into it. This implies that the quality of feature vectors influences the performance output of HMM. Therefore, eigendecomposition (ED) algorithms were deployed to extract features from passive acoustic monitoring (PAM) datasets containing whale vocalisations. These extracted features were subsequently used with HMM for improved performance of the model towards the detection of whale vocalisations.

Specifically, principal components (PCs) computed from principal component analysis (PCA) were uniquely deployed to extract feature vectors. These feature vectors were then used with HMM for the detection of whale vocalisations. Similarly, modes

computed from dynamic mode decomposition (DMD) were uniquely transformed into feature vectors. These feature vectors were subsequently used with HMM for the detection of whale vocalisations. To further deepen the quality of the feature vectors, the kernel method was introduced into each of the feature extraction (FE) techniques. The incorporation of the kernel method in each of the ED-FE techniques offered an opportunity to enhance their ability to capture non-linear relationships within the datasets, thus increasing the efficacy of the FE, as evidenced by the experimental results of the study.

The analyses of the experimental results conducted confirm that the performance of HMM is influenced by the quality of feature vectors used with it. Also, the results showed the importance of carefully selecting the dimension of the feature vectors, as this could impact the computational load and performance of the models. Furthermore, it was observed that the sample size influenced the performance of the respective models. Overall, the kernel-based FE techniques exhibit better performance in comparison with the respective primary techniques (PCA and DMD).

7.2 Research Limitations

The datasets deployed to test the performance of the developed models were separately sourced, each with a specific focus on detecting either the southern right whale (SRW) or the humpback whale (HW). Consequently, each dataset has been tailored to address the detection of a particular species, potentially limiting the generalisability of the models to a broader range of species. Nonetheless, the proposed models in this work can be further tested on other whale species. Furthermore, only vocalisations from the designated species of interest were selected, while any other vocalisations present within the datasets were categorised as noise during both the training and testing stages of the models. While this approach enables focused analyses, it could inadvertently result in the exclusion of valuable acoustic information from other marine life that may be present in the recordings. In the SRW dataset, only the 'variable' call type is identified. Therefore, the models may not show the

same level of performance when tested on datasets containing other SRW call types. Similarly, HW are known for the complexity of their vocalisation characteristics. Consequently, variations in HW songs can occur even within the same region. This inherent variability implies that the models might not exhibit consistent performance levels when tested with datasets from diverse regions.

7.3 Future Research Directions

Whale species have continued to gain the attention of researchers due to their economic and other importance to the ecosystem. Over the years, various methods have been proposed to aid the analyses of the large volume of PAM datasets. Interestingly, the 92 whale species exhibit differences in vocalisations within and between species suborders. This diversity restricts the use of the different existing methods for the detection and classification of all whale species. So, no single method is able to detect and classify all whale species. However, there is potential for future research to explore the development of models capable of detecting a wider range of whale species while maintaining a similar level of performance. This undertaking could involve identifying common features shared among different species, which can then be harnessed for model development.

In this study, we categorised the sound signals in the PAM datasets into two categories: (1) the identified whale vocalisations and (2) other sound signals as noise. However, the term 'noise' could imply other marine mammals since the ocean environment accommodates a diverse range of marine mammals and their habitats are sometimes overlapping. Future research could look into the possibility of identifying the characteristics of other sounds such as fish, shrimp, and other aquatic life beyond the whale vocalisations in the PAM datasets and categorising them accordingly. This may reveal more information on the ecology of other animal species in the ocean. Furthermore, it is worth noting that during the annotation phase of this research, the start and end points of identified calls were selected, and these annotations were not overlapped. While this approach served our current research objectives well,

future investigations could consider implementing overlapping annotations. Such an approach may contribute to uncovering more features in the datasets, which can make the model more robust in discriminating between whale vocalisations and noise. Another area of future research will be to design an interface for incorporating the detection models with existing software such as the *Sonic Visualiser*, and *Audacity* for inspecting whale recordings, such that the results from the models would be feasible within the spectrogram or waveform views as experts scroll through the PAM recordings.

Bibliography

[1] Committee on Taxonomy. "List of Marine Mammal Species and Subspecies." *Society for Marine Mammalogy*, 2021. URL https://marinemammalscience.org/science-and-publications/list-marine-mammal-species-subspecies/. Accessed: 5th August 2021.

[2] T. A. Jefferson, M. A. Webber, and R. L. Pitman. *Marine Mammals of the World: A Comprehensive Guide to their Identification*. United Kingdom: Elsevier, 2011.

[3] N. Gales, M. Hindell, and R. Kirkwood. *Marine Mammals: Fisheries, Tourism and Management Issues: Fisheries, Tourism and Management Issues*. Australia: CSIRO Publishing, 2003.

[4] N. Gales, M. Hindell, and R. Kirkwood. *Marine Mammals: Fisheries, Tourism and Management Issues: Fisheries, Tourism and Management Issues*. Australia: CSIRO Publishing, 2003.

[5] J. S. Smith, L. K. Hill, and R. M. Gonzalez. "Whale watching and preservation of the environment in central Baja California, Mexico." *Focus on Geography*, vol. 62, no. 1, pp. 1–8, Nov. 2019. URL DOI:10.21690/foge/2019.62.3f.

[6] M. L. Dicken. "Socio-economic aspects of boat-based ecotourism during the sardine run within the Pondoland Marine Protected Area, South Africa." *African Journal of Marine Science*, vol. 32, no. 2, pp. 405–411, 2010. URL https://doi.org/10.2989/1814232X.2010.502642.

[7] E. C. M. Parsons, S. Baulch, T. Bechshoft, G. Bellazzi, P. Bouchet, A. M. Cosentino, C. A. J. Godard-Codding, F. Gulland, M. Hoffmann-Kuhnt, E. Hoyt, et al. "Key research questions of global importance for cetacean conservation." *Endangered Species Research*, vol. 27, no. 2, pp. 113–118, Feb. 2015. URL https://doi.org/10.3354/esr00655.

[8] Y. Xian. *Detection and classification of whale acoustic signals*. Ph.D. thesis, Duke University, 2016.

[9] R. Williams, A. J. Wright, E. Ashe, L. K. Blight, R. Bruintjes, R. Canessa, C. W. Clark, S. Cullis-Suzuki, D. T. Dakin, C. Erbe, et al. "Impacts of anthropogenic noise on marine life: publication patterns, new discoveries, and future directions in research and management." *Ocean & Coastal Management*, vol. 115, pp. 17–24, Oct. 2015. URL https://doi.org/10.1016/j.ocecoaman.2015.05.021.

[10] R. A. Dunlop. "The effect of vessel noise on humpback whale, Megaptera novaeangliae, communication behaviour." *Animal Behaviour*, vol. 111, pp. 13–21, Jan. 2016. URL https://doi.org/10.1016/j.anbehav.2015.10.002.

[11] L. S. Weilgart. "A brief review of known effects of noise on marine mammals." *International Journal of Comparative Psychology*, vol. 20, no. 2, pp. 159–168, Dec. 2007. URL https://doi.org/10.46867/ijcp.2007.20.02.09.

[12] H. Slabbekoorn, N. Bouton, I. van Opzeeland, A. Coers, C. ten Cate, and A. N. Popper. "A noisy spring: the impact of globally rising underwater sound levels on fish." *Trends in Ecology & Evolution*, vol. 25, no. 7, pp. 419–427, Jul. 2010. URL https://doi.org/10.1016/j.tree.2010.04.005.

[13] W. J. Richardson, C. R. Greene Jr, C. I. Malme, and D. H. Thomson. *Marine Mammals and Noise*. Academic Press, 2013.

[14] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack. "Estimating animal population density

using passive acoustics." *Biological Reviews*, vol. 88, no. 2, pp. 287–309, Nov. 2013. URL https://doi.org/10.1111/brv.12001.

[15] W. M. X. Zimmer. *Passive Acoustic Monitoring of Cetaceans*. Cambridge University Press, 2011.

[16] S. E. Nelms, J. Alfaro-Shigueto, J. P. Y. Arnould, I. C. Avila, S. B. Nash, E. Campbell, M. I. D. Carter, T. Collins, R. J. C. Currey, C. Domit, et al. "Marine mammal conservation: Over the horizon." *Endangered Species Research*, vol. 44, pp. 291–325, Mar. 2021. URL https://doi.org/10.3354/esr01115.

[17] A. M. Usman, O. O. Ogundile, and D. J. J. Versfeld. "Review of automatic detection and classification techniques for cetacean vocalization." *IEEE Access*, vol. 8, pp. 105181–105206, Jun. 2020. URL https://doi.org/10.1109/ACCESS.2020.3000477.

[18] J. Jiang, L. Bu, F. Duan, X. Wang, W. Liu, Z. Sun, and C. Li. "Whistle detection and classification for whales based on convolutional neural networks." *Applied Acoustics*, vol. 150, pp. 169–178, Feb. 2019. URL https://doi.org/10.1016/j.apacoust.2019.02.007.

[19] F. Pace, P. White, and O. Adam. "Hidden Markov modeling for humpback whale (Megaptera Novaeanglie) call classification." In *Proceedings of Meetings on Acoustics ECUA2012*, vol. 17, pp. 1–8. Acoustical Society of America, Edinburgh, Scotland: ASA, Jul. 2012. URL https://doi.org/10.1121/1.4772751.

[20] R. L. Putland, L. Ranjard, R. Constantine, and C. A. Radford. "A hidden Markov model approach to indicate Bryde's whale acoustics." *Ecological Indicators*, vol. 84, pp. 479–487, Sep. 2018. URL https://doi.org/10.1016/j.ecolind.2017.09.025.

[21] O. P. Babalola, A. M. Usman, O. O. Ogundile, and D. J. J. Versfeld. "Detection of Bryde's whale short pulse calls using time domain features with hidden

Markov models." *SAIEE Africa Research Journal*, vol. 112, no. 1, pp. 15–23, Mar. 2021. URL https://doi.org/10.23919/SAIEE.2021.9340533.

[22] O. O. Ogundile, A. M. Usman, and D. J. J. Versfeld. "An empirical mode decomposition based hidden Markov model approach for detection of Bryde's whale pulse calls." *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. EL125–EL131, Jan. 2020. URL https://doi.org/10.1121/10.0000717.

[23] O. O. Ogundile, A. M. Usman, O. P. Babalola, and D. J. J. Versfeld. "Dynamic mode decomposition: A feature extraction technique based hidden Markov model for detection of mysticetes' vocalisations." *Ecological Informatics*, vol. 63, no. 101306, pp. 1–12, Jul. 2021. URL https://doi.org/10.1016/j.ecoinf.2021.101306.

[24] S. Adams and P. A. Beling. "A survey of feature selection methods for Gaussian mixture models and hidden Markov models." *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1739–1779, Sep. 2019. URL https://doi.org/10.1007/s10462-017-9581-3.

[25] O. O. Ogundile, A. M. Usman, O. P. Babalola, and D. J. J. Versfeld. "A hidden Markov model with selective time domain feature extraction to detect inshore Bryde's whale short pulse calls." *Ecological Informatics*, vol. 57, no. 101087, pp. 1–7, Apr. 2020. URL https://doi.org/10.1016/j.ecoinf.2020.101087.

[26] S. Gupta, J. Jaafar, W. F. W. Ahmad, and A. Bansal. "Feature extraction using MFCC." *Signal & Image Processing: An International Journal (SIPIJ)*, vol. 4, no. 4, pp. 101–108, 2013.

[27] N. Desai, K. Dhameliya, and V. Desai. "Feature extraction and classification techniques for speech recognition: A review." *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 12, pp. 367–371, Dec. 2013.

[28] H. Gupta and D. Gupta. "LPC and LPCC method of feature extraction in speech recognition system." In *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, pp. 498–502. Noida, India: IEEE, Jan. 2016. URL https://doi.org/10.1109/CONFLUENCE.2016.7508171.

[29] L. Muda, M. Begam, and I. Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." *Journal of Computing*, vol. 2, no. 3, Mar. 2010. URL https://doi.org/10.48550/arXiv.1003.4083.

[30] S. L. Brunton and J. N. Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

[31] I. T. Jolliffe and J. Cadima. "Principal component analysis: a review and recent developments." *Philosophical Transactions of The Royal Society A: Mathematical, Physical And Engineering Sciences*, vol. 374, no. 2065, pp. 1–16, 2016. URL http://dx.doi.org/10.1098/rsta.2015.0202.

[32] B. Schölkopf, A. Smola, and K.-R. Müller. "Nonlinear component analysis as a kernel eigenvalue problem." *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998. URL https://ieeexplore.ieee.org/abstract/document/6790375.

[33] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. Philadelphia, USA: Society for Industrial and Applied Mathematics (SIAM), 2016.

[34] A. S. Frankel. "Sound Production." In *Encyclopedia of Marine Mammals*, pp. 1056–1071. Elsevier, 2009.

[35] M. Bittle and A. Duncan. "A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring." In *Proceedings of Acoustics*, vol. 2013, pp. 1–8. Victor Harbor, Australia: Australian Acoustical Society, Nov. 2013. URL https://www.acoustics.asn.au/conference_proceedings/AAS2013/papers/p64.pdf.

[36] O. Adam. "Segmentation of killer whale vocalizations using the Hilbert-Huang transform." *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, pp. 1–10, 2008. URL https://doi:10.1155/2008/245936.

[37] G. Qiao, M. Bilal, S. Liu, Z. Babar, and T. Ma. "Biologically inspired covert underwater acoustic communication—A review." *Physical Communication*, vol. 30, pp. 107–114, Oct. 2018. URL https://doi.org/10.1016/j.phycom.2018.07.007.

[38] J. Jiang, L. Bu, X. Wang, C. Li, Z. Sun, H. Yan, B. Hua, F. Duan, and J. Yang. "Clicks classification of sperm whale and long-finned pilot whale based on continuous wavelet transform and artificial neural network." *Applied Acoustics*, vol. 141, no. 1, pp. 26–34, Dec. 2018. URL https://doi.org/10.1016/j.apacoust.2018.06.014.

[39] P. C. Bermant, M. M. Bronstein, R. J. Wood, S. Gero, and D. F. Gruber. "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics." *Scientific Reports*, vol. 9, no. 12588, pp. 1–10, Dec. 2019. URL https://doi.org/10.1038/s41598-019-48909-4.

[40] C. Erbe. "Underwater acoustics: noise and the effects on marine mammals." *A Pocket Handbook*, vol. 164, 2011. URL http://oalib.hlsresearch.com/PocketBook3rded.pdf.

[41] D. K. Mellinger, K. M. Stafford, S. E. Moore, R. P. Dziak, and H. Matsumoto. "An overview of fixed passive acoustic observation methods for cetaceans." *Oceanography*, vol. 20, no. 4, pp. 36–45, Dec. 2007. URL http://www.jstor.org/stable/24860138.

[42] T. A. Marques, L. Thomas, J. Ward, N. DiMarzio, and P. L. Tyack. "Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales." *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 1982–1994, Apr. 2009. URL https://doi.org/10.1121/1.3089590.

[43] M. André, M. Van Der Schaar, S. Zaugg, L. Houégnigan, A. M. Sánchez, and J. V. Castell. "Listening to the deep: live monitoring of ocean noise and cetacean acoustic signals." *Marine Pollution Bulletin*, vol. 63, no. 1-4, pp. 18–26, Jun. 2011. URL https://doi.org/10.1016/j.marpolbul.2011.04.038.

[44] R. Gibb, E. Browning, P. Glover-Kapfer, and K. E. Jones. "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring." *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 169–185, Oct. 2019. URL https://doi.org/10.1111/2041-210X.13101.

[45] Y. Ren, M. Johnson, P. Clemins, M. Darre, S. S. Glaeser, T. Osiejuk, and E. Out-Nyarko. "A framework for bioacoustic vocalization analysis using hidden Markov models." *Algorithms*, vol. 2, no. 4, pp. 1410–1428, Nov. 2009. URL https://doi.org/10.3390/a2041410.

[46] L. M. Munger, D. K. Mellinger, S. M. Wiggins, S. E. Moore, and J. A. Hildebrand. "Performance of spectrogram cross-correlation in detecting right whale calls in long-term recordings from the Bering Sea." *Canadian Acoustics*, vol. 33, no. 2, pp. 25–34, Jun. 2005.

[47] B. A. Weisburn, S. G. Mitchell, C. W. Clark, and T. W. Parks. "Isolating biological acoustic transient signals." In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 269–272. Minneapolis, MN, USA: IEEE, Apr. 1993. URL https://doi.org/10.1109/ICASSP.1993.319107.

[48] O. O. Ogundile and D. J. J. Versfeld. "Analysis of template-based detection algorithms for inshore Bryde's whale short pulse calls." *IEEE Access*, vol. 8, pp. 14377–14385, Jan. 2020. URL https://doi.org/10.1109/ACCESS.2020.2966254.

[49] A. K. Ibrahim, H. Zhuang, N. Erdol, and A. M. Ali. "A new approach for north atlantic right whale upcall detection." In *2016 International Symposium on Computer, Consumer and Control (IS3C)*, pp. 260–263. Xi'an, China: IEEE, Jul. 2016. URL https://doi.org/10.1109/IS3C.2016.76.

[50] P. Peso Parada and A. Cardenal-López. "Using Gaussian mixture models to detect and classify dolphin whistles and pulses." *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3371–3380, Jun. 2014. URL https://doi.org/10.1121/1.4876439.

[51] S. Yu, K. J. Palmer, M. A. Roch, E. Fleishman, X. Liu, N. Eva-Marie, H. Tyler, C. Danielle, D. Gillespie, and K. Holger. "Deep neural networks for automated detection of marine mammal species." *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020. URL https://doi.org/10.1038/s41598-020-57549-y.

[52] P. Rickwood and A. Taylor. "Methods for automatically analyzing humpback song units." *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1763–1772, Mar. 2008. URL https://doi.org/10.1121/1.2836748.

[53] T. M. Yack, J. Barlow, S. Rankin, and D. Gillespie. "Testing and validation of automated whistle and click detectors using PAMGUARD 1.0." techreport, National Oceanic and Atmospheric Administration (NOAA), USA, May 2009. URL https://repository.library.noaa.gov/view/noaa/3665. NOAA Technical Memorandum NMFS.

[54] Q. Q. Huynh, L. N. Cooper, N. Intrator, and H. Shouval. "Classification of underwater mammals using feature extraction based on time-frequency analysis and BCM theory." *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1202–1207, May 1998. URL https://doi.org/10.1109/78.668783.

[55] A. Djebbari and F. B. Reguig. "Short-time Fourier transform analysis of the phonocardiogram signal." In *ICECS 2000. 7th IEEE International Conference on Electronics, Circuits and Systems (Cat. No. 00EX445)*, vol. 2, pp. 844–847. Jounieh, Lebanon: IEEE, Dec. 2000. URL https://doi.org/10.1109/ICECS.2000.913008.

[56] J. J. Lee, S. M. Lee, I. Y. Kim, H. K. Min, and S. H. Hong. "Comparison between short time Fourier and wavelet transform for feature extraction of heart sound." In *Proceedings of IEEE. IEEE Region 10 Conference. TENCON*

99.'Multimedia Technology for Asia-Pacific Information Infrastructure'(Cat. No. 99CH37030), vol. 2, pp. 1547–1550. Cheju Island, South Korea: IEEE, Sep. 1999. URL https://doi.org/10.1109/TENCON.1999.818731.

[57] R. X. Gao and R. Yan. "Non-stationary signal processing for bearing health monitoring." *International Journal of Manufacturing Research*, vol. 1, no. 1, pp. 18–40, Aug. 2006. URL https://doi.org/10.1504/IJMR.2006.010701.

[58] Z. Xiao, M. Zhang, L. Chen, and H. Jin. "Detection and segmentation of underwater CW-like signals in spectrum image under strong noise background." *Journal of Visual Communication and Image Representation*, vol. 60, pp. 287–294, Mar. 2019. URL https://doi.org/10.1016/j.jvcir.2019.02.036.

[59] M. P. Fargues and R. Bennett. "Comparing wavelet transforms and AR modeling as feature extraction tools for underwater signal classification." In *Conference Record of The Twenty-Ninth Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 915–919. Pacific Grove, CA, USA: IEEE, Oct. 1995. URL https://doi.org/10.1109/ACSSC.1995.540833.

[60] P. Hill, A. Achim, M. E. Al-Mualla, and D. Bull. "Contrast sensitivity of the wavelet, dual tree complex wavelet, curvelet, and steerable pyramid transforms." *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2739–2751, Jun. 2016. URL https://doi.org/10.1109/TIP.2016.2552725.

[61] M. Tausif, E. Khan, M. Hasan, and M. Reisslein. "SMFrWF: Segmented modified fractional wavelet filter: Fast low-memory discrete wavelet transform (DWT)." *IEEE Access*, vol. 7, pp. 84448–84467, 2019. URL https://doi.org/10.1109/ACCESS.2019.2924490.

[62] M. H. M. Alhabib, M. Z. N. Al-Dabagh, F. H. AL-Mukhtar, and H. I. Hussein. "Exploiting wavelet transform, principal component analysis, support vector machine, and k-nearest neighbors for partial face recognition." *Cihan University-Erbil Scientific Journal*, vol. 3, no. 2, pp. 80–84, Aug. 2019. URL https://doi.org/10.24086/cuesj.v3n2y2019.pp80-84.

[63] Y. Wu, G. Gao, and C. Cui. "Improved wavelet denoising by non-convex sparse regularization under double wavelet domains." *IEEE Access*, vol. 7, pp. 30659–30671, Mar. 2019. URL https://doi.org/10.1109/ACCESS.2019.2903125.

[64] L. Chun-Lin. "A tutorial of the wavelet transform.", Feb. 2010. URL http://disp.ee.ntu.edu.tw/tutorial/WaveletTutorial.pdf.

[65] G. G. Yen and K. Lin. "Wavelet packet feature extraction for vibration monitoring." *IEEE Transactions on Industrial Electronics*, vol. 47, no. 3, pp. 650–667, Jun. 2000. URL https://doi.org/10.1109/41.847906.

[66] O. Duzenli. *Classification of underwater signals using wavelet-based decompositions*. Master's thesis, Naval Postgraduate School Monterey California, Monterey California, Jun. 1998.

[67] M. Lopatka, O. Adam, 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, pp. 463–467, 2005.

[68] O. Adam, M. Lopatka, C. Laplanche, and J.-F. Motsch. "Sperm whale signal analysis: comparison using the autoregressive model and the Daubechies 15 wavelets transform." *International Journal of Electrical and Computer Engineering*, vol. 1, no. 4, pp. 188–195, Apr. 2007. URL https://doi.org/10.5281/zenodo.1055375.

[69] J. A. Antonino-Daviu. "Operation of the Hilbert-Huang transform: basic overview with examples." *Technical School of Industrial Engineering*, pp. 1–11, Jul. 2013. URL http://hdl.handle.net/10251/30740.

[70] 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, Nov. 2006. URL https://doi.org/10.1121/1.2354003.

[71] 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 non-linear and non-stationary time series analysis." *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, Mar. 1998. URL https://doi.org/10.1098/rspa.1998.0193.

[72] O. Adam. "The use of the Hilbert-Huang transform to analyze transient signals emitted by sperm whales." *Applied Acoustics*, vol. 67, no. 11-12, pp. 1134–1143, Aug. 2006. URL https://doi.org/10.1016/j.apacoust.2006.04.001.

[73] Z. K. Peng, W. T. Peter, and F. L. Chu. "An improved Hilbert–Huang transform and its application in vibration signal analysis." *Journal of Sound and Vibration*, vol. 286, no. 1-2, pp. 187–205, Dec. 2005. URL https://doi.org/10.1016/j.jsv.2004.10.005.

[74] J. Liu, X.-K. Li, T. Ma, S.-C. Piao, and Q.-Y. Ren. "An improved Hilbert-Huang transform and its application in underwater acoustic signal detection." In *2009 2nd International Congress on Image and Signal Processing*, pp. 1–5. Tianjin, China: IEEE, 2009. URL https://doi.org/10.1109/CISP.2009.5304603.

[75] K. D. Seger, M. H. Al-Badrawi, J. L. Miksis-Olds, N. J. Kirsch, and A. P. Lyons. "An empirical mode decomposition-based detection and classification approach for marine mammal vocal signals." *The Journal of the Acoustical Society of America*, vol. 144, no. 6, pp. 3181–3190, Dec. 2018. URL https://doi.org/10.1121/1.5067389.

[76] L. Rabiner and R. W. Schafer. "Digital speech processing." *The Froehlich/Kent Encyclopedia of Telecommunications*, vol. 6, pp. 237–258, 2011.

[77] L. Wang, Z. Chen, and F. Yin. "A novel hierarchical decomposition vector quantization method for high-order LPC parameters." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 212–221, Jan. 2014. URL https://doi.org/10.1109/TASLP.2014.2380352.

[78] B. S. Atal and S. L. Hanauer. "Speech analysis and synthesis by linear prediction of the speech wave." *The journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, Aug. 1971. URL https://doi.org/10.1121/1.1912679.

[79] S. Mazhar, T. Ura, and R. Bahl. "Effect of temporal evolution of songs on cepstrum-based voice signature in humpback whales." In *OCEANS 2008-MTS/IEEE Kobe Techno-Ocean*, pp. 1–8. Marine Technology Society, Kobe, Japan: IEEE, 2008. URL https://doi.org/10.1109/OCEANSKOBE.2008.4531057.

[80] E. Chandra et al. "Keyword spotting system for Tamil isolated words using Multidimensional MFCC and DTW algorithm." In *2015 International Conference on Communications and Signal Processing (ICCSP)*, pp. 0550–0554. IEEE, Melmaruvathur, India: IEEE, Apr. 2015. URL https://doi.org/10.1109/ICCSP.2015.7322545.

[81] S. Khalid, T. Khalil, and S. Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." In *2014 Science and Information Conference*, pp. 372–378. London, UK: IEEE, 2014. URL https://doi.org/10.1109/SAI.2014.6918213.

[82] L. Shi, I. Ahmad, Y. He, and K. Chang. "Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments." *Journal of Communications and Networks*, vol. 20, no. 5, pp. 509–518, Oct. 2018. URL https://doi.org/10.1109/JCN.2018.000075.

[83] K. M. Ravikumar and S. Ganesan. "Comparison of multidimensional MFCC feature vectors for objective assessment of stuttered disfluencies." *International Journalof Advanced Networking and Applications (IJANA)*, vol. 2, no. 05, pp. 854–860, 2011.

[84] M. R. Hasan, M. Jamil, M. G. R. M. S. Rahman, et al. "Speaker identification using mel frequency cepstral coefficients." *Variations*, vol. 1, no. 4, 2004.

[85] C. P. Dalmiya, V. S. Dharun, and K. P. Rajesh. "An efficient method for Tamil speech recognition using MFCC and DTW for mobile applications." In *2013 IEEE Conference on Information & Communication Technologies*, pp. 1263–1268. Thuckalay, India: IEEE, 2013. URL https://doi.org/10.1109/CICT.2013.6558295.

[86] P. Somervuo, A. Harma, and S. Fagerlund. "Parametric representations of bird sounds for automatic species recognition." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006. URL https://doi.org/10.1109/TASL.2006.872624.

[87] M. A. Roch, M. S. Soldevilla, R. Hoenigman, S. M. Wiggins, and J. A. Hildebrand. "Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes." *Canadian Acoustics*, vol. 36, no. 1, pp. 41–47, 2008. URL http://www.caa-aca.ca/old_site/E/journal_frames.htm.

[88] V. Kandia and Y. Stylianou. "Detection of sperm whale clicks based on the Teager–Kaiser energy operator." *Applied Acoustics*, vol. 67, no. 11–12, pp. 1144–1163, Jul. 2006. URL https://doi.org/10.1016/j.apacoust.2006.05.007.

[89] Y. Xian, A. Thompson, Q. Qiu, L. Nolte, D. Nowacek, J. Lu, and R. Calderbank. "Classification of whale vocalizations using the Weyl transform." In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 773–777. South Brisbane, QLD, Australia: IEEE, Apr. 2015. URL https://doi.org/10.1109/ICASSP.2015.7178074.

[90] D. Gillespie, M. Caillat, J. Gordon, and P. White. "Automatic detection and classification of odontocete whistles." *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2427–2437, Sep. 2013. URL https://doi.org/10.1121/1.4816555.

[91] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Pearson, Jan. 1975.

[92] D. K. Mellinger and C. W. Clark. "Methods for automatic detection of mysticete sounds." *Marine & Freshwater Behaviour & Physiology*, vol. 29, no. 1-4, pp. 163–181, 1997. URL https://doi.org/10.1080/10236249709379005.

[93] R. A. Altes. "Detection, estimation, and classification with spectrograms." *The Journal of the Acoustical Society of America*, vol. 67, no. 4, pp. 1232–1246, Apr. 1980. URL https://doi.org/10.1121/1.384165.

[94] D. K. Mellinger. "A comparison of methods for detecting right whale calls." *Canadian Acoustics*, vol. 32, no. 2, pp. 55–65, Jun. 2004. URL https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1588.

[95] D. K. Mellinger, S. L. Nieukirk, H. Matsumoto, S. L. Heimlich, R. P. Dziak, J. Haxel, M. Fowler, C. Meinig, and H. V. Miller. "Seasonal occurrence of North Atlantic right whale Eubalaena glacialis vocalizations at two sites on the Scotian Shelf." *Marine Mammal Science*, vol. 23, no. 4, pp. 856–867, Oct. 2007. URL https://doi.org/10.1111/j.1748-7692.2007.00144.x.

[96] 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, Jun. 2000. URL https://doi.org/10.1121/1.429434.

[97] Q. Xue, Y. H. Hu, and W. J. Tompkins. "Neural-network-based adaptive matched filtering for QRS detection." *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 4, pp. 317–329, Apr. 1992. URL https://doi.org/10.1109/10.126604.

[98] J. C. Bancroft. "Introduction to matched filters." *CREWES Research*, vol. 14, pp. 1–8, 2002. URL https://www.crewes.org/Documents/ResearchReports/2002/2002-46.pdf.

[99] G. Turin. "An introduction to matched filters." *IRE Transactions on Information theory*, vol. 6, no. 3, pp. 311–329, Jun. 1960. URL https://doi.org/10.1109/TIT.1960.1057571.

[100] P. Courmontagne. *The stochastic matched filter and its applications to detection and de-noising*, chap. Chapter 15, pp. 272–298. IntechOpen, Aug. 2010.

[101] L. Bouffaut, R. Dreo, V. Labat, A. Boudraa, and G. Barruol. "Antarctic blue whale calls detection based on an improved version of the stochastic matched filter." In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2319–2323. Kos, Greece: IEEE, 2017. URL https://doi.org/10.23919/EUSIPCO.2017.8081624.

[102] S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980. URL https://doi.org/10.1109/TASSP.1980.1163420.

[103] P. Senin. "Dynamic time warping algorithm review." techreport 855, Information and Computer Science Department University of Hawaii, Manoa Honolulu, USA, Dec. 2008. URL https://csdl.ics.hawaii.edu/techreports/2008/08-04/08-04.pdf.

[104] G. Cleuziou and J. G. Moreno. "Kernel methods for point symmetry-based clustering." *Pattern Recognition*, vol. 48, no. 9, pp. 2812–2830, Sep. 2015. URL https://doi.org/10.1016/j.patcog.2015.03.013.

[105] J. C. Brown, A. Hodgins-Davis, and P. J. O. Miller. "Classification of vocalizations of killer whales using dynamic time warping." *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. EL34–EL40, Feb. 2006. URL https://doi.org/10.1121/1.2166949.

[106] J. C. Brown and P. J. O. Miller. "Automatic classification of killer whale vocalizations using dynamic time warping." *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1201–1207, Aug. 2007. URL https://doi.org/10.1121/1.2747198.

[107] X. Yin, Y. Hou, J. Yin, and C. Li. "A novel SVM parameter tuning method based on advanced whale optimization algorithm." In *Journal of*

Physics: Conference Series, vol. 1237, p. 022140. IOP Publishing, 2019. URL 10.1088/1742-6596/1237/2/022140.

[108] V. Jakkula. *Tutorial on support vector machine (SVM)*. Washington State University, Washington, 2006.

[109] C. Junli and J. Licheng. "Classification mechanism of support vector machines." In *WCC 2000-ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000*, vol. 3, pp. 1556–1559. Beijing, China: IEEE, 2000. URL https://doi.org/10.1109/ICOSP.2000.893396.

[110] S. Jarvis, N. DiMarzio, R. Morrissey, and D. Moretti. "A novel multi-class support vector machine classifier for automated classification of beaked whales and other small odontocetes." *Canadian Acoustics*, vol. 36, no. 1, pp. 34–40, 2008. URL https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1988.

[111] H. Mohebbi-Kalkhoran, C. Zhu, M. Schinault, and P. Ratilal. "Classifying humpback whale calls to song and non-song vocalizations using bag of words descriptor on acoustic data." In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 865–870. Boca Raton, FL, USA: IEEE, Dec. 2019. URL https://doi.org/10.1109/ICMLA.2019.00150.

[112] A. K. Jain, J. Mao, and K. M. Mohiuddin. "Artificial neural networks: A tutorial." *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996. URL https://doi.org/10.1109/2.485891.

[113] V. Cheung and K. Cannons. "An introduction to neural networks." *Signal & Data Compression Laboratory, University of Manitoba, Winnipeg, Manitoba, Canada*, May 2002.

[114] W. S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity." *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943. URL https://doi.org/10.1007/BF02478259.

[115] S. O. Murray, E. Mercado, and H. L. Roitblat. "The neural network classification of false killer whale (Pseudorca crassidens) vocalizations." *The Journal of the Acoustical Society of America*, vol. 104, no. 6, pp. 3626–3633, Dec. 1998. URL https://doi.org/10.1121/1.423945.

[116] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer. "Efficient processing of deep neural networks: A tutorial and survey." *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, Nov. 2017. URL https://doi.org/10.1109/JPROC.2017.2761740.

[117] V. B. Deecke and V. M. Janik. "Automated categorization of bioacoustic signals: avoiding perceptual pitfalls." *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 645–653, 2006. URL https://doi.org/10.1121/1.2139067.

[118] L. Zhang, D. Wang, C. Bao, Y. Wang, and K. Xu. "Large-scale whale-call classification by transfer learning on multi-scale waveforms and time-frequency features." *Applied Sciences*, vol. 9, no. 5, pp. 1–11, Mar. 2019. URL https://doi.org/10.3390/app9051020.

[119] R. Sridharan. "Gaussian mixture models and the EM algorithm.", 2014. URL Avilablein:http://people.csail.mit.edu/rameshvs/content/gmm-em.pdf.

[120] J. A. Bilmes. "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models." *International Computer Science Institute*, vol. 4, no. 510, p. 126, Apr. 1998. URL http://www.leap.ee.iisc.ac.in/sriram/teaching/MLSP_18/refs/GMM_Bilmes.pdf.

[121] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[122] 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, Mar. 2007. URL https://doi.org/10.1121/1.2400663.

[123] L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989. URL https://doi.org/10.1109/5.18626.

[124] D. Jurafsky. *Speech & Language Processing*. Pearson Education India, 2000.

[125] X. Liu, K. Shi, Z. Wang, and J. Chen. "Exploit camera raw data for video super-resolution via hidden Markov model inference." *IEEE Transactions on Image Processing*, vol. 30, pp. 2127–2140, 2021. URL https://doi.org/10.1109/TIP.2021.3049974.

[126] E.-M. Nel, J. A. Du Preez, and B. M. Herbst. "Estimating the pen trajectories of static signatures using hidden Markov models." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1733–1746, Nov. 2005. URL https://doi.org/10.1109/TPAMI.2005.221.

[127] Z. Ghahramani. "An introduction to hidden Markov models and Bayesian networks." *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, no. 01, pp. 9–42, 2001. URL https://doi.org/10.1142/S0218001401000836.

[128] H. A. Engelbrecht and J. A. du Preez. "Efficient backward decoding of high-order hidden Markov models." *Pattern Recognition*, vol. 43, no. 1, pp. 99–112, Jun. 2010. URL https://doi.org/10.1016/j.patcog.2009.06.004.

[129] J. A. du Preez. "Efficient training of high-order hidden Markov models using first-order representations." *Computer Speech & Language*, vol. 12, no. 1, pp. 23–39, Jan. 1998. URL https://doi.org/10.1006/csla.1997.0037.

[130] J. P. Coelho, T. M. Pinho, and J. Boaventura-Cunha. *Hidden Markov Models: Theory and Implementation using Matlab®*. CRC Press, 2019.

[131] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*, vol. 4. New York: Springer, 2006.

[132] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains." *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, Feb. 1970. URL https://www.jstor.org/stable/2239727.

[133] T. K. Moon. "The expectation-maximization algorithm." *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, Nov. 1996. URL https://doi.org/10.1109/79.543975.

[134] P. Larue, P. Jallon, and B. Rivet. "Modified k-mean clustering method of HMM states for initialization of Baum-Welch training algorithm." In *2011 19th European Signal Processing Conference*, pp. 951–955. Barcelona, Spain: IEEE, 2011.

[135] J. MacQueen. "Classification and analysis of multivariate observations." In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297. 1967.

[136] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa. "Application of K–means clustering algorithm for prediction of students academic performance." *arXiv preprint arXiv:1002.2425*, 2010.

[137] H.-L. Lou. "Implementing the Viterbi algorithm." *IEEE Signal Processing Magazine*, vol. 12, no. 5, pp. 42–52, Sep. 1995. URL http://10.1109/79.410439.

[138] G. D. Forney. "The Viterbi algorithm." *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973. URL https://doi.org/10.1109/PROC.1973.9030.

[139] M. Vieira, P. J. Fonseca, M. C. P. Amorim, and C. J. C. Teixeira. "Call recognition and individual identification of fish vocalizations based on automatic speech recognition: An example with the Lusitanian toadfish." *The Journal of*

the Acoustical Society of America, vol. 138, no. 6, pp. 3941–3950, Dec. 2015. URL https://doi.org/10.1121/1.4936858.

[140] P. M. Scheifele, M. T. Johnson, M. Fry, B. Hamel, and K. Laclede. "Vocal classification of vocalizations of a pair of Asian Small-Clawed otters to determine stress." *The Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. EL105–EL109, Jul. 2015. URL https://doi.org/10.1121/1.4922768.

[141] J. C. Brown and P. Smaragdis. "Hidden Markov and Gaussian mixture models for automatic call classification." *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. EL221–EL224, May 2009. URL https://doi.org/10.1121/1.3124659.

[142] S. Datta and C. Sturtivant. "Dolphin whistle classification for determining group identities." *Signal Processing*, vol. 82, no. 2, pp. 251–258, May 2002. URL https://doi.org/10.1016/S0165-1684(01)00184-0.

[143] J. A. Hildebrand, K. E. Frasier, T. A. Helble, and M. A. Roch. "Performance metrics for marine mammal signal detection and classification." *The Journal of the Acoustical Society of America*, vol. 151, no. 1, pp. 414–427, 2022. URL https://doi.org/10.1121/10.0009270.

[144] H. A. Garcia, T. Couture, A. Galor, J. M. Topple, W. Huang, D. Tiwari, and P. Ratilal. "Comparing performances of five distinct automatic classifiers for fin whale vocalizations in beamformed spectrograms of coherent hydrophone array." *Remote Sensing*, vol. 12, no. 2, p. 326, Jan. 2020. URL https://doi.org/10.3390/rs12020326.

[145] P. J. Schmid. "Dynamic mode decomposition of numerical and experimental data." *Journal of Fluid Mechanics*, vol. 656, pp. 5–28, Aug. 2010. URL https://doi.org/10.1017/S0022112010001217.

[146] S. J. Buchan, R. Mahú, J. Wuth, N. Balcazar-Cabrera, L. Gutierrez, S. Neira, and N. B. Yoma. "An unsupervised hidden Markov model-based system for

the detection and classification of blue whale vocalizations off Chile." *Bioacoustics*, vol. 29, no. 2, pp. 140–167, Jan. 2020. URL https://doi.org/10.1080/09524622.2018.1563758.

[147] 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, Jan. 2011. URL https://doi.org/10.1121/1.3514383.

[148] S. Mazhar, T. Ura, and R. Bahl. "Vocalization based individual classification of humpback whales using support vector machine." In *OCEANS 2007*, pp. 1–9. Vancouver, BC, Canada: IEEE, Sep. 2007. URL https://doi.org/10.1109/OCEANS.2007.4449356.

[149] J. Luan, R. Bahl, T. Ura, T. Akamatsu, M. Yamaguchi, T. Sakamaki, and K. Mori. "Model-based recognition of individual humpback whales from their vocalization features." In *Ocean Engineering Symposium*. 2003. URL https://doi.org/10.1121/1.3514383.

[150] W. Luo, W. Yang, and Y. Zhang. "Convolutional neural network for detecting odontocete echolocation clicks." *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. EL7–EL12, Jan. 2019. URL https://doi.org/10.1121/1.5085647.

[151] A. Cuevas, A. Veragua, S. Español-Jiménez, G. Chiang, and F. Tobar. "Unsupervised blue whale call detection using multiple time-frequency features." In *2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, pp. 1–6. Pucon, Chile, Oct. 2017.

[152] S. Nisar, O. U. Khan, M. Tariq, et al. "An efficient adaptive window size selection method for improving spectrogram visualization." *Computational Intelligence and Neuroscience*, vol. 2016, pp. 1–13, 2016. URL https://doi.org/10.1155/2016/6172453.

[153] P. Holmes, J. L. Lumley, G. Berkooz, and C. W. Rowley. *Turbulence, Coherent Structures, Dynamical Systems And Symmetry*. Cambridge University Press, 2012. URL https://doi.org/10.1017/CBO9780511919701.

[154] J. N. Kutz. *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data*. Oxford University Press, 2013.

[155] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*, vol. 50. Society for Industrial and Applied Mathematics (SIAM), 1997.

[156] K. Pearson. "LIII. 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. URL https://doi.org/10.1080/14786440109462720.

[157] J. Grosek and J. N. Kutz. "Dynamic mode decomposition for real-time background/foreground separation in video." *arXiv preprint arXiv:1404.7592*, 2014. URL https://doi.org/10.48550/arXiv.1404.7592.

[158] J. L. Proctor, S. L. Brunton, and J. N. Kutz. "Dynamic mode decomposition with control." *SIAM Journal on Applied Dynamical Systems*, vol. 15, no. 1, pp. 142–161, 2016. URL https://doi.org/10.1137/15M1013857.

[159] B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz. "Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition." *Journal of Neuroscience Methods*, vol. 258, pp. 1–15, Jan. 2016. URL https://doi.org/10.1016/j.jneumeth.2015.10.010.

[160] N. Mohan, K. P. Soman, and S. S. Kumar. "A data-driven strategy for short-term electric load forecasting using dynamic mode decomposition model." *Applied Energy*, vol. 232, pp. 229–244, 2018. URL https://doi.org/10.1016/j.apenergy.2018.09.190.

[161] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. "On dynamic mode decomposition: theory and applications." *Journal of Computational Dynamics*, vol. 1, no. 2, pp. 391–421, Dec. 2014. URL https://cwrowley.princeton.edu/papers/Tu-DMD.pdf.

[162] J. Mann and J. N. Kutz. "Dynamic mode decomposition for financial trading strategies." *Quantitative Finance*, vol. 16, no. 11, pp. 1643–1655, Apr. 2016. URL https://doi.org/10.1080/14697688.2016.1170194.

[163] H. Wang and N. Noguchi. "Real-time states estimation of a farm tractor using dynamic mode decomposition." *GPS Solutions*, vol. 25, no. 1, pp. 1–12, 2021. URL https://doi.org/10.1007/s10291-020-01051-5.

[164] A. Narasingam and J. S.-I. Kwon. "Development of local dynamic mode decomposition with control: Application to model predictive control of hydraulic fracturing." *Computers & Chemical Engineering*, vol. 106, pp. 501–511, Nov. 2017. URL https://doi.org/10.1016/j.compchemeng.2017.07.002.

[165] S. V. Vaerenbergh. *Kernel methods for nonlinear identification, equalization and separation of signals*. Ph.D. thesis, Universidad de Cantabria, 2010. URL http://hdl.handle.net/10902/1397.

[166] B. Schölkopf, A. J. Smola, F. Bach, et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, England: The MIT Press, 2002.

[167] J. Shawe-Taylor, N. Cristianini, et al. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[168] A. Kocsor and L. Tóth. "Kernel-based feature extraction with a speech technology application." *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2250–2263, 2004. URL https://doi.org/10.1109/TSP.2004.830995.

[169] S. Y. Kung. *Kernel methods and machine learning*. Cambridge University Press, 2014.

[170] N. Aronszajn. "Theory of reproducing kernels." *Transactions of the American mathematical society*, vol. 68, no. 3, pp. 337–404, 1950.

[171] K. E. Pilario, M. Shafiee, Y. Cao, L. Lao, and S.-H. Yang. "A review of kernel methods for feature extraction in nonlinear process monitoring." *Processes*, vol. 8, no. 24, pp. 1–47, Dec. 2019.

[172] Q. Wang. "Kernel principal component analysis and its applications in face recognition and active shape models." *Website: http://arxiv. org/pdf/1207.3538, diakses tanggal*, vol. 1, 2011.

[173] M. O. Williams, C. W. Rowley, and I. G. Kevrekidis. "A kernel-based method for data-driven koopman spectral analysis." *Journal of Computational Dynamics*, vol. 2, no. 2, pp. 247–265, May 2016. URL https://doi.org/10.3934/jcd.2015005.

[174] C. Cannam, C. Landone, and M. Sandler. "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files." In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1467–1468. Special Interest Group on Multimedia Systems (SIGMM), Firenze, Italy: Association for Computing Machinery, New York, United States, Oct. 2010. URL https://doi.org/10.1145/1873951.1874248.

[175] L. H. Hofmeyr-Juritz and P. B. Best. "Acoustic behaviour of southern right whales in relation to numbers of whales present in Walker Bay, South Africa." *African Journal of Marine Science*, vol. 33, no. 3, pp. 415–427, 2011. URL https://doi.org/10.2989/1814232X.2011.637616.

[176] J. van Wyk, J. A. du Preez, and D. J. J. Versfeld. "Temporal separation of whale vocalizations from background oceanic noise using a power calculation." *Ecological Informatics*, vol. 69, no. 101627, pp. 1–10, Jul. 2022. URL https://doi.org/10.1016/j.ecoinf.2022.101627.

[177] L. H. Hofmeyr-Juritz. *The nature and rate of vocalisation by southern right whales (Eubalaena australis), and the evidence for individually distinctive calls*.

Ph.D. thesis, University of Pretoria, 2010. URL http://hdl.handle.net/2263/25299.

[178] A. S. Frankel, C. W. Clark, L. M. Herman, and C. M. Gabriele. "Spatial distribution, habitat utilization, and social interactions of humpback whales, Megaptera novaeangliae, off Hawai'i, determined using acoustic and visual techniques." *Canadian Journal of Zoology*, vol. 73, no. 6, pp. 1134–1146, Jun. 1995. URL https://doi.org/10.1139/z95-135.

[179] A. K. Stimpert, W. W. L. Au, S. E. Parks, T. Hurst, and D. N. Wiley. "Common humpback whale (Megaptera novaeangliae) sound types for passive acoustic monitoring." *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 476–482, Feb. 2011. URL https://doi.org/10.1121/1.3504708.

[180] A. M. Usman and D. J. J. Versfeld. "Detection of baleen whale species using kernel dynamic mode decomposition-based feature extraction with a hidden Markov model." *Ecological Informatics*, vol. 71, no. 101766, pp. 1–16, Nov. 2022. URL https://doi.org/10.1016/j.ecoinf.2022.101766.

[181] A. M. Usman and D. J. J. Versfeld. "Principal components-based hidden Markov model for automatic detection of whale vocalisations." *Journal of Marine Systems*, vol. 242, no. 103941, pp. 1–23, Feb. 2024. URL https://doi.org/10.1016/j.jmarsys.2023.103941.