IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 18, 2020, accepted June 3, 2020, date of publication June 8, 2020, date of current version June 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000477

Review of Automatic Detection and Classification Techniques for Cetacean Vocalization

AYINDE M. USMAN[1], (Member, IEEE), OLAYINKA O. OGUNDILE[1,2], (Member, IEEE), AND DANIEL J. J. VERSFELD[1], (Member, IEEE)

[1]Department of Electrical and Electronic Engineering, Stellenbosch University, Stellenbosch 7600, South Africa
[2]Department of Computer Science, Tai Solarin University of Education, Ijebu Ode 2118, Nigeria

Corresponding author: Ayinde M. Usman (23172908@sun.ac.za)

This work was supported by the National Research Foundation (NRF) South Africa under Grant 116036.

ABSTRACT Cetaceans have elicited the attention of researchers in recent decades due to their importance to the ecosystem and their economic values. They use sound for communication, echolocation and other social activities. Their sounds are highly non-stationary, transitory and range from short to long sounds. Passive acoustic monitoring (PAM) is a popular method used for monitoring cetaceans in their ecosystems. The volumes of data accumulated using PAM are usually big, so they are difficult to analyze using manual inspection. Therefore different techniques with mixed outcomes have been developed for the automatic detection and classification of signals of different cetacean species. So far, no single technique developed is perfect to detect and classify the vocalizations of over 82 known species due to variability in time-frequency, difference in the amplitude among species and within species' vocal repertoire, physical environment, among others. The accuracy of any detector or classifier depends on the technique adopted as well as the nature of the signal to be analyzed. In this article, we review the existing techniques for the automatic detection and classification of cetacean vocalizations. We categorize the surveyed techniques, while emphasizing the advantages and disadvantages of these techniques. The article suggests possible research directions that can improve existing detection and classification techniques. In addition, the article recommends other suitable techniques that can be used to analyze non-linear and non-stationary signals such as the cetaceans' signals. Several research have been dedicated to this topic, however, there is no review of these past results that gives a quick overview in the area of cetacean detection and classification. This review will help researchers and practitioners in the field to make insightful decisions based on their requirements.

INDEX TERMS Cetacean, classification, detection, feature extraction, passive acoustic monitoring (PAM), vocalization.

I. INTRODUCTION

The increasing human anthropogenic activities have significantly changed the soundscape in oceans. This has continued to threaten the existence of ocean mammals because they utilize sound for navigation, communication, avoidance of predators, recognition of prey for survival and to function properly within their ecosystem [1]–[8]. Anthropogenic activities, which have been detrimental to marine fauna include shipping, offshore exploration, geophysical seismic surveys and naval sonar operations [1]–[4], [7]. The potential negative effects of these activities include (a) physical injury, (b) physiological dysfunction: permanent or temporary loss of hearing sensitivity, (c) behavioral modification: decrease in exploration efficiency, or inefficient use of environment, separation of mother-calf pairs, (d) masking- difficulty in recognizing crucial sounds as a result of increase in background noise, (e) avoidance and displacement from critical feeding and breeding grounds, (f) decrease reproduction rate [7]–[10]. Reactions of marine mammals to these negative impacts varies due to factors such as species, age, gender, previous noise experience, location or body of water and behavioral state [10]. There have been growing research on the consequence of these human activities on marine mammals in recent years [4], [8], [9], [11], [12]. Ecosystem managers are particularly interested in conserving these mammals by searching for ways to mitigate the effects of human activities within the marine ecosystem [1] in order to support their

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

conservation and protection [6], [13]. However, they are faced with the challenge of inadequate knowledge on the ecosystems of these mammals [1], [13].

There are different groups of marine mammals. They are grouped them into: cetaceans (whales, dolphins and porpoises), carnivora (pinnipeds, seals, sea lions, walruses, sea otters and the polar bear) and sirenians (manatee, dugongs and sea dogs) [9], [14], [15]. The cetaceans and sirenians order live their entire life in water [13], [14], thus making it difficult for scientists to know the exact estimate of their population [1], [13]. Two suborders exist in the cetacean taxonomy: *odontocete* or toothed whales with about 72 known living species (examples include the Beluga whale, bottlenose dolphins, Cuiver's beaked whale, Killer whale, among others) and *mysticete* or baleen whales with about 14 known living species (examples include Humpback whale, Bryde's whale, Bowhead whale, among others) [9], [14]. The cetacean species are present throughout the world oceans (with the *odontocete* suborder present in some freshwater lakes and rivers) [9].

Cetaceans have elicited attention of policy makers and researchers due to their economic importance. There are continuous increase in public demands for ecotourism business which reportedly involve over 87 nations and territories [16]. Commercial tourism on free ranging cetaceans gives tourists the opportunities to observe, touch, swim. The whale-watching business enterprise generates over US$2 billion yearly [17]; thus, employing thousands of people and contributing to governments revenue. Cetaceans are also of great importance in maintenance of state of health of ecosystem and serving as sentry species for the state of marine ecosystem [18]. Some cetacean species are also used for security purposes by the US navy [19]. Besides, anthropogenic noise effects on marine mammals, in particular seismic activities,seismic activities, military operations, and the oil and gas industries, is one of the main policy concerns to government [20], [21].

Traditionally, cetaceans were visually surveyed to assess how they utilize a specific area in order to have insight into their ecology. But, their populations are often underestimated because they spend their entire life in water which makes visual observation insufficient for accurate estimation of their population [13], [22], [23]. The visual observation is also hindered by environmental conditions such as remote topography, time of the day, the short time cetaceans spend on the surface, and high mobility rate [13], [22], [24].

Acoustic monitoring on the other hand, serves as an important avenue to monitor marine mammals at great distance (as far as 100km in some instances for low frequency calls) compared to visual methods, because sound can propagate much further in oceans than light [6], [13], [25]. Acoustic monitoring is also not affected by conditions under which visual method cannot perform. Acoustic monitoring can either be active or passive. In active acoustic monitoring (AAM), sound energy is transmitted and the returning signals are analyzed. This approach is not popular because it can upset the animals behaviour due to it intrusive mode of operation [24]. For passive acoustic monitoring (PAM) however, marine mammals sounds are captured from the surrounding environment in a non-invasive manner through the use of underwater microphones (hydrophones), hence it is widely used for marine mammals observation [24], [26]. Besides being significant for the survey and census of cetacean distribution, PAM is also an important component in lessening the negative consequences of human actions on cetaceans [6], [13]. It requires the cetaceans to produce signal which makes it perform better for high vocal species. Also, it can be tricky to estimate the number of species present based on the number of detected calls because in order to obtain an accurate and precise population density, the cue rate must be properly calculated [23]. The success of PAM is however dependent on the quality of the techniques use to isolate the signal of interest from the rest of signals present in a dataset, particularly for remote sources and low signal to noise ratio (SNR) [6].

Cetaceans produce a variety of distinctive sounds for communication, echolocation and other social functions. These sounds which are non-stationary, vary in physical properties, with many species producing different or combinations of sounds [9], [13]. These vocalizations occupy a very wide frequency band and show different characteristics which include: variation within and between species [22], [27], they can be species-specific [28], temporal and geographical difference [22]; thus, making them complex to analyze [29]. The cetacean sounds are generally categorized as clicks, whistles, songs and (burst) pulsed calls [30]–[32]. The click sounds which are produced by all cetaceans are used for echolocation and foraging [30], [33]. They are impulsive short pulses of substantial strength which last for micro seconds and extremely directional. Cetaceans produce multiple clicks at a time with diverse cue rate ranging from 0.5-2 clicks per seconds. The whistle and pulsed calls are used for social activities. The whistles which are produced by all cetaceans but at different frequency [30] are relatively tonal or pulsed, narrow band, frequency-modulated signals with frequency range from 2 to 30kHz [34] depending on species. Songs are the most complicated of signals produced by the *Mysticete* suborder [13], [34]. A song is a series of individual calls organized into a hierarchical structure that can go on for several minutes in some instances, even hours [30]. Just a few cetacean generate songs, examples include humpback, bowhead, blue and fin whales [13], [34], [35]. The known frequency ranges of cetacean sounds including clicks, whistles and FM calls is from about 10Hz to 100kHz [24], [30].

Different techniques, with mixed outcomes have been developed for the detection and classification of cetacean vocalization. The difference in the outputs of these techniques are as results of time-frequency variability, difference in the amplitude among species and within species' vocal repertoire, physical environment where the sounds were recorded, ability to cope with noise, ability to perform in real time, computational requirements [1], [6], [22]. No single technique is perfect to detect and classify all species [22]. These

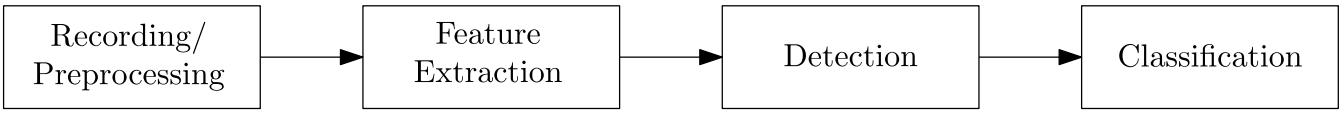


FIGURE 1. Block diagram of cetacean detection and classification stages.

techniques are centered on signal processing, pattern recognition and machine learning concepts. The vocalizations of a targeted species can be manually detected, when a specialist listens to sound or view spectrograms to find the vocalization [24], [36]. However, the large acoustic data collected during the recordings are difficult to analyze by the human operators, hence the need to have an algorithm that can automatically detect and classify these large volume of recorded sounds. This helps in processing the large acoustic datasets relatively faster and with consistency. Besides, human bias are eliminated and sounds that are beyond the human hearing limit can be detected, thus reducing the error rate [1], [36], [37].

Application of a robust automatic detection and classification model to cetacean vocalizations can give understanding into their repertoire variation, individual vocal changeability, social setting relationships and some other crucial cetacean behavioral questions [38]. In the last few years, machine learning approaches and other automatic techniques have been used for the detection and classification of cetacean sounds [33], [39]–[46] with varying performance outputs. The basic block diagram of the detection and classification stages is shown in Fig. 1. The cetacean vocalizations are recorded through the use of hydrophones and preprocessed. The preprocessing include extraction of features which is followed by the detection and classification stages. The detection stage is the process of identifying the presence of the targeted cetacean signal in the dataset from other unwanted signals that may be present. These unwanted signals may include background noises (non bioacoustics signals), presence of signals of non-targeted species in the recording area, and so on. The classification stage is the process of assigning the detected signals to a predefined category (species-specific) [31], [47] having been guided by a previous knowledge of what is expected, usually by a human expert. The variability within and between the emitted signals make the classification stage very crucial. It helps in categorizing individual species. Therefore, having different techniques that automatically detect and categorize different species of cetacean will aid comparisons both between and within species. Each of the stages in Fig. 1 are explicitly discussed in subsequent sections.

In this paper, a review of existing techniques for the detection and classification of cetacean vocalizations has been carried out. Many research work had been dedicated to developing techniques for automatic detection and classification of cetacean signals. However, there has not been a detail review to give a comprehensive overview of the existing techniques for detection and classification except in [22] which gave an overview without addressing details of each technique. A block diagram of the implementation steps for designing automatic detectors and classifiers is shown in Fig. 1. Each of the four steps involved are reviewed with more detail on popular techniques used for feature extraction as well as for design of detectors and classifiers. The strengths and weaknesses of every method are elaborated in order to guide researchers on best technique to deploy, depending on their goals. The relevance of this paper is to assist anyone engaging in related research to have a quick overview on the procedures for the detection and classification of cetacean signals.

The remainder of this paper is arranged as follows. Section II describes the process of collecting the data and preprocessing norms. Section III and Section IV reviews the existing techniques for feature extractions, detection and classification respectively. Section V discusses the style of reporting outputs in this area of research. The findings, challenges and future research prospects in this subject area are discussed in Section VI while the paper is concluded in Section VII.

II. DATA RECORDING AND PREPROCESSING

Sounds are recorded via the use of hydrophones (underwater microphones). The recording is done via two methods; mobile survey (mobile PAM) and stationary (fixed) survey (fixed PAM). In mobile survey, hydrophones can be attached at the rear of a ship or ocean glider to sample a large area. The hydrophones are towed behind a ship or ocean glider on a cable of tens to thousands of meters longs. On the other hand, hydrophones are left in a location for a long period to record sounds either continuously or with an on-off sampling plan in stationary (fixed) survey [48]. Advantages of the mobile method include; coverage of large area and ease of combining acoustics survey with a visual survey. Advantages of the fixed method include; observations usually traverse a longer time period thus giving opportunity to have a large dataset to work with and it is often less expensive than the mobile PAM [6], [24]. It is however faced with challenges such as recovering of instruments in deep water, difficulty in telling whether high vocalization rate represent many cetaceans vocalizing occasionally or a few cetaceans vocalizing a lot [49]. Thus, certain conditions such as frequency range, total deployment period, and difficulty of recovery of recording gadgets are considered when choosing a recording method [48].

The hydrophones set-up for recordings can either be an array of multiple-hydrophones or single hydrophone. In a multiple array set-up, at least four hydrophones are

set appropriately, are essential to determine the three-dimensional location of a calling animal by accessing the travel time differences of the incoming sound. This method is usually ambiguous and expensive to implement. A single hydrophone method for taking recording provides a simple means to record. It is cheaper and performs well under certain conditions [50]. Cetaceans usually move in clusters with many individuals vocalizing simultaneously as they move [51]. The recorded sound is thus complex to analyze as a result of this simultaneous vocalization and the presence of anthropogenic noise [25], [51]. Estimating the abundance of species or types from the recorded sounds is often a problem because the sound could be from an individual vocalizing continuously or multiple individual vocalizing simultaneously.

The preprocessing stage focuses on recovering of frequency data and producing a time-frequency-amplitude representation of the recorded signal to form a dataset [25], [26]. This process include the denoising done to clean and enhance the quality of the whale sound [32]. This is followed by annotation of part of the recording, where the applicable sound must be located in time which they occur within the recording by a human expert who is assumed to know the exact sound of the target. The beginning and end point of particular sound class is 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 recognized labels are then used as samples for training of the technique to be used for detection and classification. This is usually done via visual inspection of spectrogram. However, a number of software are available to carry out the annotation seamlessly. Examples of existing software for annotation are *Sonic Visualiser*, *RAVEN* © and *Audacity* ™ [29]. Fig. 2 and Fig. 3 are examples of spectrogram and time series representation of Bryde's whale and Blue whale calls respectively.

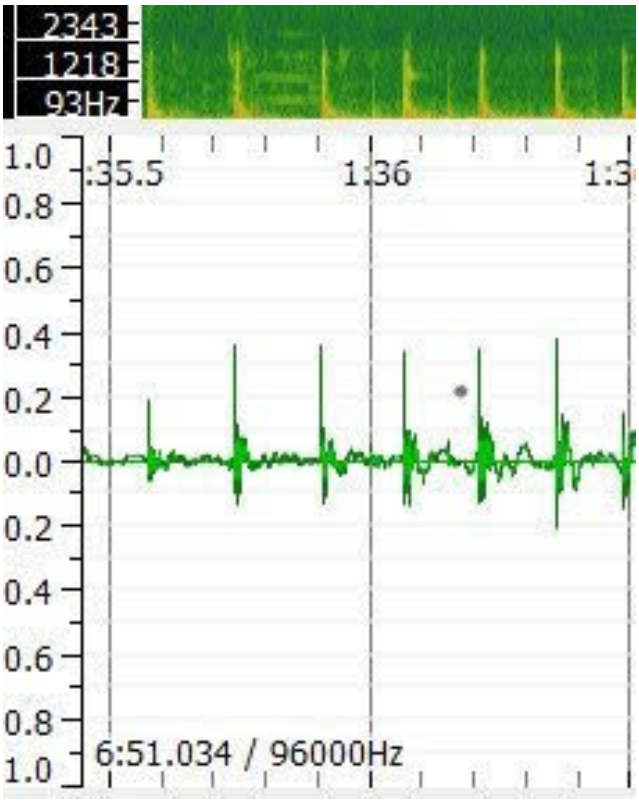


FIGURE 2. An example of spectrogram and time-series representation of Bryde's whale pulse call. The sound was recorded in Gordon's bay harbour, False bay, South-West, South Africa [52].

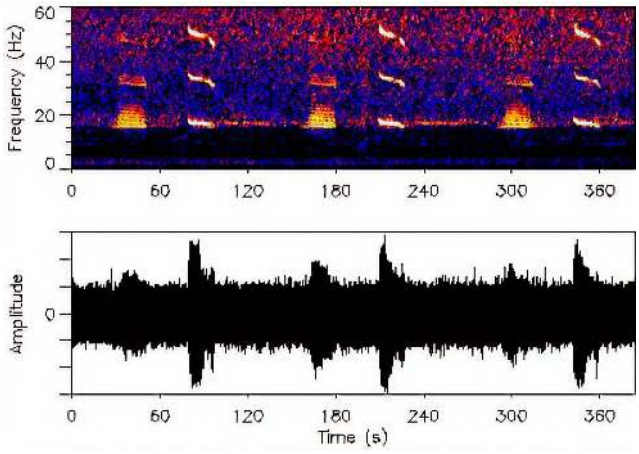


FIGURE 3. An example of spectrogram and time-series representation of Northeast Pacific Blue whale call. The sound was recorded from offshore Vancouver Island B.C., Washington and Oregon, California and off Mexico in the Gulf of California, USA. [53].

III. FEATURE EXTRACTION METHODS

Underwater signals are highly non stationary due to the presence of environmental sounds (rain, cracking of ice, estuaries, and so on), human anthropogenic sounds (shipping, offshore exploration, geophysical seismic survey, and so on) and the biological sounds (marine mammals sound) [54]. Many of these sounds are not relevant to the detection and classification of interest, thus they tend to reduce performance of detectors and classifiers. Hence, the need to extract useful information from raw sound recordings. Feature extraction (FE) is the process of extracting relevant information from the data so as to enhance the accuracy of the detector and classifier by removing redundant data. The features presented to a detector or classifier is central to its performance. Raw data usually include other unwanted information. The FE process is basically for extracting the needed features or parameters of the signal of interest from the raw dataset. The reduction can be via feature extraction and feature selection. While feature selection is the selection of subsets of relevant feature variables from the raw datasets, feature extraction on the other hand is the elimination of irrelevant variables from the raw datasets by building new datasets without losing any relevant variables of interest [55], [56]. The main difference between these two methodologies is that feature selection reduces the number of features to that are best at differentiating between classes but feature extraction will create entirely new smaller

datasets by measuring certain features or properties that can be used to differentiate between data of different classes. The required features can be selected from original dataset G with variables $\{g_1, g_2, g_3, \ldots.., g_N\}$ and we have a new subset G' with variables $\{g'_1, g'_2, g'_3, \ldots.., g'_N\}$. In feature extraction, original dataset Y with variables $\{y_1, y_2, \ldots.., y_N\}$ is transformed to a new dataset Z with variables $\{z_1, z_2, \ldots\ldots, z_M\}$, where $M < N$. The recorded sounds are translated to a set of feature vectors that interprets the prominent attributes of the recordings. In order to build an effective detector or classifier, the features to be extracted must satisfy the specific problem to be addressed. It should be noted that these methods are not exclusively used for cetacean signal analysis. In fact, they have enjoyed wide application in different fields of studies. Therefore, our explanation on each is tailored towards it application to cetacean signal analysis.

A. SHORT TIME FOURIER TRANSFORM (STFT)

The short-time Fourier transform (STFT) method also known as windowed Fourier transform is used for analyzing non stationary signals. It strives to solve the problem of loss of time information in the Fourier transform (FT) by introducing a sliding window $w(t)$ passing through the whole signal $x(t)$. It presents the time-localized details of the non-stationary signal $x(t)$ by disclosing the changes of the frequency content as time progresses. It has been used as feature extracting tool on Phonocardiogram (PCG) signals [57], [58], vibration signals measured from rolling bearings and other machine components [59]. The STFT is obtained by multiplying the time signal $x(t)$ by a suitable sliding time window function $w(t-\tau)$ which is constructed to excerpt 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 [58], [59]:

$$X(\tau, \omega) = \int_{-\infty}^{\infty} x(t) w(t - \tau) \exp^{-j\omega t} dt, \qquad (1)$$

where ω is the frequency. The STFT employs a sliding window to obtain a spectrogram which provides information of both 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 acquire the feature vectors, which is energy intensity to be used as the sample space binary clustering of Gaussian mixture model [60]. In [25], the STFT was used to create whistle contour in a denoised spectrogram of whistles of long-finned pilot whales and killer whales. A raw sound data that have been denoised were sequentially sliced into sound frames. The STFT coefficients for every single sound frame was computed to create visible contours of the whistles in the spectrogram. The sound frames containing the whistles were manually marked and labeled in preparation for their usage for training and testing of deep convolutional neural networks (CNNs) model meant to detect and classify the whistles accordingly.

B. WAVELET TRANSFORM (WT)

The wavelet transform technique (WT) is perfect in 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 [61]–[65]. The WT decomposes 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 centered on small wavelets with limited duration [66] developed as alternative to solving time-frequency resolution problems associated with short time Fourier transform (STFT) [67]. 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 provides improved time-frequency resolution of components of the signal under review. It has three advantages; (1) it is more efficient for short-lived features extraction as related with 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 $\psi_{s,u}(t)$ is defined as

$$\psi_{s,u}(t) = \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-u}{s}\right), \qquad s > 0,\ u \in \Re, \qquad (2)$$

where $\psi(t)$ is the mother wavelet function, s is the scaling parameter which allows $\psi_{s,u}(t)$ to expand or contract and u is the translating parameter (translation in time allowing time shifting of $\psi_{s,u}(t)$). By convention, the wavelet function is configured to attain a balance between time domain (limited distance) and frequency domain (limited bandwidth). As the mother wavelet dilates and time-shift (translate), a small scaling parameter s leads to high frequency wavelet function $\psi_{s,u}(t)$ which gives good time resolution with poor frequency output while a large scaling parameter s leads to low frequency wavelet function $\psi_{s,u}(t)$ which gives poor time resolution with good frequency output [67], [68].

There are two kinds of WT: the continuous wavelet transform (CWT) and the discrete wavelet transform (DWT). Both can be applied in cetacean signal feature extraction. CWT uses a continuous wavelet function to determine the complete wavelet coefficient $\psi_{s,u}$ of a continuous signal by analyzing the low-frequency content of a signal with a broad translation function, while it analyzes high-frequency content of a signal with a short-duration function [66], [67]. The CWT transform is defined as shown in equation (3) [66]

$$W_x(s, u) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) dt, \qquad (3)$$

where $\psi(t)$ is the mother wavelet function, $x(t)$ is the input signal, s and u are as defined above. The input signal $x(t)$ which is hitherto a one-dimensional signal has been transformed to a two-dimensional coefficients $W_x(s, u)$. The two variables; s and u can execute the time-frequency analysis. A specific frequency (parameter s) can be found at a definite time instant (parameter u). Intuitively, equation (3) simply

mean $W_x(s, u)$ is the energy of x of scale s at $t = u$. CWT is associated with two problems; it implementation is difficult and finding the scaling function is hard.

The DWT of a time domain signal $x(t)$ is defined as:

$$W_x(s, u) = \sum_t \frac{1}{\sqrt{s}} x(t)\psi^* \left(\frac{t-u}{s}\right) dt \tag{4}$$

The scaling parameter s and the translating parameter u only take discrete values in the DWT. DWT can be implemented efficiently using a pair of low pass and high pass filter as proposed in [69]. The low and high-frequency parts of a signal are separated through the use of filters. The signal is passed through low pass and high pass filters and down sampled by a factor of two. Details of DWT procedures are explained in [69].

The DWT is easy to implement and yields faster computational time because it uses filters. The wavelet transform is however confronted with the problem of poor discrimination between signals with close high-frequency components which is as a result of poor frequency resolution in the high-frequency region [67]. This leads to complexity in differentiating high frequency transients. The wavelet packet transform (WPT) provides an alternative that overcomes this limitation. It decomposes both low and high frequency components at each level; thus giving better resolution. Procedure for signal decomposition using WPT is explained in details in [59].

There are different wavelet transform families such as Haar wavelet, Morlet wavelet, Packet wavelet, Daubechies wavelet, symlet wavelet, Coiflet wavelet. They differ with respect to length of support of the mother wavelet, speed of decaying coefficients, symmetry and orthogonality and bi-orthogonality of the resulting functions [70]. Certain criteria are to be used to select a particular mother wavelet family. In cetacean vocalization detection and classification process, WT method is used to extract a set of features that may be used in classifying signals. The wavelet family basis are selected according to the signal to be analyzed for the formation of feature vectors. Due to difference in the species signal (between and within individual species), different wavelet families may be used to analyze the signal of a species, evident in the analysis of sperm whale signals, where the authors use different wavelet families with each stating the advantages of the wavelet family used. Morlet mother wavelet was used in [61]; Daubechies mother wavelet present better result in [71] while wavelet packet transform was preferred in [72].

The CWT has been used as feature extraction tools for sperm whale clicks by different authors employing different wavelet families [61], [71], [73]. The sound emitted by sperm whales are distinctive, short-time and have a broadband spectrum [73]. A new feature extraction method based on CWT approach was used in [32] 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 wavelet coefficient matrix was proposed, centering on the energy distribution and duration variance between the two whale clicks. The feature vector was 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. Fargues and Bennett in [61] compared classification rate achieved using DWT and AutoRegressive (AR) modeling to extract features from recordings containing killer whale, pilot whale, sperm whale, gray whale, humpback whale, and underwater earthquake signals while a back-propagation neural network is used for classification.

C. HILBERT HUANG TRANSFORM (HHT)

The Hilbert Huang transform (HHT) is an alternative method for characterizing bioacoustics signals. It enjoys wide area of applications in bioacoustics signal characterization, fault diagnosis in nuclear reactors, biomedical diagnosis, electrical machines condition monitoring, seismic studies, financial application, among others [74]. It gives an improved result than conventional time-frequency analysis methods such as STFT and WT [51]. The method is entirely empirical and it is implemented in two phases.

First phase is the decomposition of the signals into some monocomponent signals called intrinsic mode function (IMF) using Empirical mode decomposition (EMD). Each IMF represents a definite frequency range and it indicates the time evolution of the components included within that band. The EMD operates in time domain, it is adaptive and highly effective. The second phase is Hilbert spectral analysis which is the application of the Hilbert transform and a 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 analyzed dataset [51], [74], [75]. The signal $x(t)$ given can be decomposed as:

$$\hat{x}(t) = \sum_{i=1}^{n} c_i(t) + r_n(t), \tag{5}$$

where c_i is the ith IMF of the signal, r_n is the residue and $\hat{x}(t)$ is an approximate of the original signal $x(t)$. The Hilbert transform is used to derive the time-frequency representation from the modes as proposed in [75]. 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 has been used in the analysis of sperm whale clicks and killer whale clicks in [76] and [29] respectively. The original HHT technique is however faced with three limitations. One, generation of undesirable IMFs at low-frequency region which may lead to misinterpretation of the result. Two, the first obtained IMF may cover too wide frequency range such that monocomponent attribute cannot be achieved, this limitation however is subject to the analyzed signal. Three, signals with low energy

component cannot be separated via the EMD operation [77]. An improved HHT technique proposed in [77] established criteria for selection of IMF through the use of Wavelet Packet Transform (WPT) as preprocessor for decomposition of signal into a set of narrow band signals. Thus, frequency components with low-energy are easily identified at 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 EMD operation would have avoided the other two identified limitations. This improved HHT technique was shown to detect underwater acoustic signals more effectively in [78].

D. EMPIRICAL MODE DECOMPOSITION (EMD)

In some recent works [39], [45], the EMD method was used to extracted features without the HHT applied. In [39], the IMF generated from the EMD were used to obtain feature vectors. The sound sources detected were uniquely labeled and verified before manually grouped into different categories. The unique labels were used to classify the detected sound sources. This new approach is a modern way to carry out unsupervised detection and classification in the time-domain depending entirely on EMD-type processing, eliminating the necessity to apply the Hilbert transform and manual labeling 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, beluga whale whistles).Also, in [45], the generated IMFs from EMD process were used to form feature vectors which were fed into a hidden Markov model (HMM) to detect Bryde's whale pulsed calls.

E. LINEAR PREDICTION COEFFICIENTS (LPCs)

The Linear prediction coefficients (LPCs) is a signal analysis method used in speech coding, speech synthesis, speech recognition, speaker recognition and verification and for speech storage [79]. The LPCs expeditiously represent speech signals as short-time spectral information [80], [81]. The main concept of this method is that it predicts the value of the current sample signal $Y(n)$ by a linear combination of past samples and then approximates the difference between the actual value and the predicted value as shown:

$$\hat{Y}(n) = \sum_{k=1}^{m} a_k Y(n-k), \tag{6}$$

$$e[n] = Y(n) - \hat{Y}(n), \tag{7}$$

where $Y(n)$ is current sample signal, $\hat{Y}(n)$ is the predicted value, a_k are m^{th} order linear predictor coefficients, $e[n]$ is the prediction error. The LPC coefficients a_k are determined by minimizing the sum squared errors over a given interval that is, the actual speech samples and the linearly predicted ones as shown in Equation (8) [79];

$$\sum_n e^2 = \sum_n \left(Y(n) - \sum_{k=1}^{m} a_k Y(n-k) \right)^2 \tag{8}$$

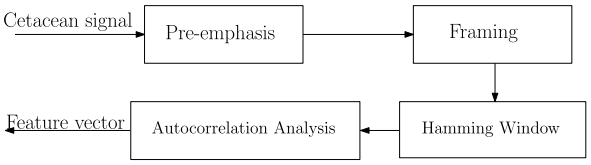


FIGURE 4. Steps for feature extraction in LPC technique.

differentiating Equation (8) with respect to a_k, $k = 1, 2, \ldots, m$ yields Equation (9)

$$\frac{\partial E}{\partial a_k} = \sum_{n=n_0}^{n_1} Y(n)Y(n-k) - \sum_{j=1}^{m} a_j \sum_{n=n_0}^{n_1} Y(n)Y(n-k)Y(n)Y(n-j) = 0, \tag{9}$$

leading to a set of m linear equations as shown in equation (10) with m unknown quantities $a_1, a_2, \ldots, a_m$ which can be solved efficiently for a_k as explained in [82]

$$\sum_{k=1}^{m} a_j c_{jk} = c_{ok}, \quad k = 1, 2, \ldots, m \quad , \tag{10}$$

where

$$c_{jk} = c_{kj} = \sum_{n=n_0}^{n_1} Y(n-j)Y(n-k), \tag{11}$$

Extracting features using LPC can be achieved in four steps as depicted in fig. 4: pre-emphasis, framing, windowing and computation of the LPC as explained in [83]. LPC was used as one of the feature extraction tools for recognition of individual humpback whale base on their vocalization data [84]. The extracted coefficients were tested on different classifier models in each of the works. The generated features were useful for the classifier; however quantization, stability and interpolation are some of the drawback of LPC.

F. MEL-SCALE FREQUENCY CEPSTRAL COEFFICIENTS (MFCCs)

The MFCC is a widely used feature extraction 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, cetacean vocalization detection and classification [56], [85]–[88]. Features are extracted in the cepstral domain to build a feature vector set for every type of signal. The wide spread use of MFCC is due to it low computational complexity, it is however sensitive to noise due to its dependence on 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 1kHz and a logarithm spacing above 1kHz [89], [90]. Generally, feature extractions using MFCC involve the following steps; Pre-emphasis, framing, hamming windowing, Fast Fourier Transform (FFT), the Mel-scale Filter bank, Logarithm operation

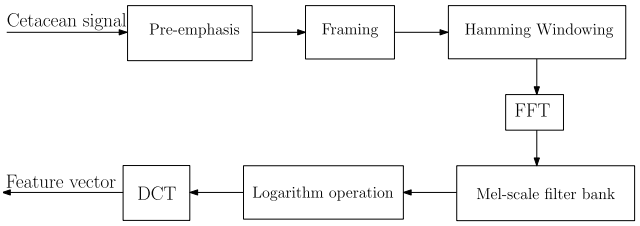


FIGURE 5. Steps for feature extraction in MFCC technique.

and Discrete Cosine Transform (DCT) as explicitly explained in [87], [90], [91]. A block diagram of these steps for extracting features using MFCC technique is shown in Fig. 5

The corresponding value for frequency f is expressed in Hz and the i_{th} mel-ceptral coefficient is shown in Equations (12) and (13) respectively, where K is the total number of cepstral coefficients, X_k is the logarithmic energy of the kth mel-spectrum band.

$$Mel_{(f)} = 1127ln\left(1 + \frac{f}{700}\right)Hz, \tag{12}$$

$$MFCC_i = \sum_{k=1}^{K} X_k \cos\left(\frac{i(k-0.5)\pi}{K}\right), \quad i = 1, 2, \ldots, K, \tag{13}$$

Typically the first twelve coefficients are utilized to compose MFCC. The set of coefficient 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.

MFCCs has been used for feature extractions in several cetacean vocalization detection and classification processes. In most cases, it produced better performance than other feature extraction techniques due to its simplicity [92]. Three feature extraction techniques (LPC, cepstrum and MFCC) were applied to extract essential features on individual Humpback whale vocalization in [93]. The MFCC was shown to outperform the other two methods.

G. OTHER FEATURE EXTRACTION METHODS

There are other feature extraction methods developed by some researchers who applied them in their work for extraction of features from cetacean signals. These isolated 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 Bienenstock, Cooper, and Munro (BCM) theory used in [54] as a feature extracting 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 an improved classification accuracy of the signals. Recently, a simple but robust feature extraction method was proposed in [42] 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 adapted 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 Teager energy operator (TEO) [44], [94], Weyl transform (WyT) [95].

A summary of surveyed feature extraction (FE) techniques in this work is given in Table 1. The examples of sound types each technique has been used to analyzed are indicated. It is observed that some FE are used directly for detection, EMD and HHT have been directly used for the detection and classification of cetacean signals as observed in [29], [39], [76]. The advantages and disadvantages of each technique are highlighted as well as general remark on the characteristics of the techniques.

IV. DETECTION AND CLASSIFICATION TECHNIQUES

Different techniques 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, amount of variation in the species' sound, feature vectors, and whether a template parameter exist for the targeted signal are considered when choosing a technique for analysis of marine sound [48]. Due to variations in acoustic repertoire with respect to region or population, it is important that the classifier is trained with calls from the intended region so as to enhance the accuracy of the classification [11]. Among the widely used existing detection and classification techniques include Gaussian mixture models (GMM), Hidden markov models (HMM), neural networks (NN), support vector machines (SVM), spectrogram cross-correlation (SPCC), matched filtering (MF) and dynamic time warping (DTW). Others that are not so popular but are potential areas where further research could be carried out to determine their viability for global application to cetacean signals include Short-Time Windowed Energy (STWE) [71]. According to the paper surveyed, we categorized these techniques into two broad classes, namely: statistical-based, and threshold-based techniques as shown in Fig. 6. We arrived at this categorization as a result of similarity in the implementation procedures shared by the techniques. The performance of each technique is dependent on the species, physical environment where the data are recorded, size of datasets available, and more importantly the feature vectors extracted from the recordings. However, some techniques do not require feature vector 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..

A. STATISTICAL-BASED TECHNIQUES

The statistical-based detection and classification techniques are techniques whose procedures use statistical inference for

TABLE 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 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 high 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 it multi-species detection and classification ability. •Faster computational time. •Does not require manually analyzed 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 characterizing low and mid frequency vocalization.	•Features generated experience issues of quantization, stability and interpolation. •Low performance in the presence of noise.	•Less complex to MFCC and easy to calculate.
MFCC		✓	✓	•Low computational complexity •Suitable for characterizing 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.

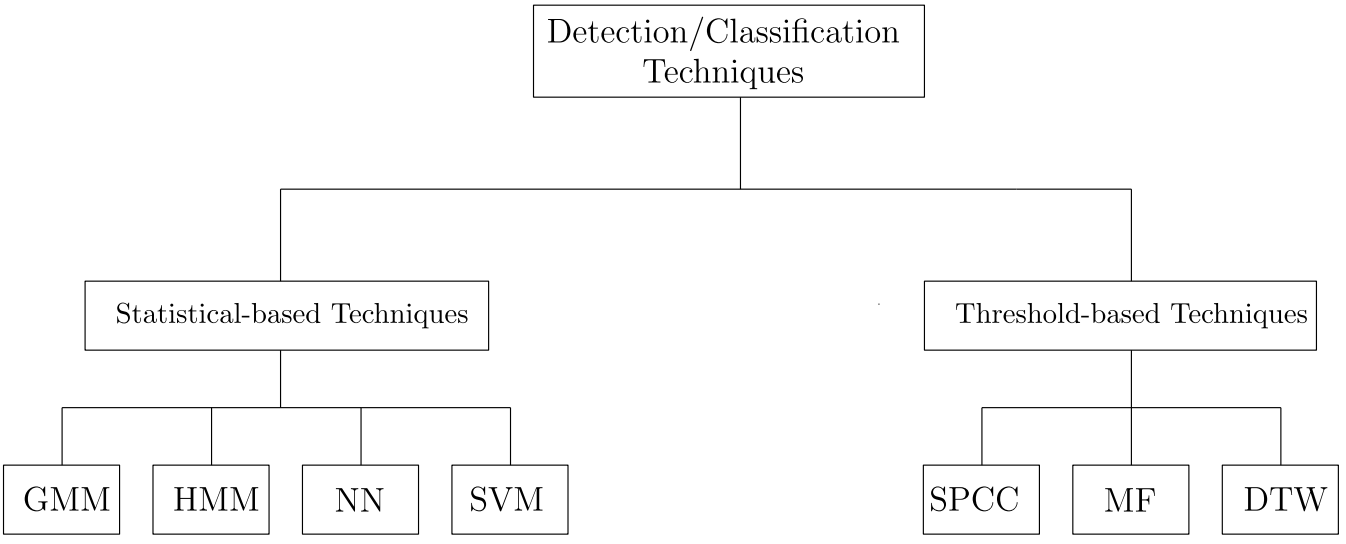


FIGURE 6. Categorization of existing detection and classification techniques.

the discovery of the best pattern that matches the features trained with the model. The modeling of these techniques are centered around statistical analysis. Examples of these techniques include GMM, HMM, NN, and SVM. They usually require large amount of datasets for the training and the testing stages. They are widely used for cetacean signal analysis and usually require a large amount of dataset for training and testing. Most of the techniques in this categorization are feature-based, i.e. they require features to be extracted from the signal to be analyzed before application of the detection and classification algorithms [96]. In this categorization, the model developed decides the outputs from the training dataset. In other words, the dataset is divided into two parts. The first part is used for training the model while the second part, which the testing stage, is used to validate the performance of the model. The model is trained

by learning to recognize salient patterns from the dataset. These patterns must be carefully defined within the context of what is expected. The training stage essentially enables the model to be able to accurately predict the patterns of the dataset. During the model testing, output will be exclusively dependent on the observed patterns from the training stage. It is important to note that the data used for testing must have never been used for training. In some of these techniques, the size of the training and the testing data have impact on the performance of the detector or classifier.

1) GAUSSIAN MIXTURE MODELS (GMM)

The Gaussian mixture model (GMM) is a classifier which uses the estimate of probability density function (PDF) to model densities of different kind of signals [97]. 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 N Gaussian distributions of feature vectors $x = (x^1, x^2, \ldots, x^D)^T$ is represented as:

$$pr\left(x\middle|\mu, \sum\right) = \sum_{i=1}^{N} k_i \frac{1}{(2\pi)^{\frac{d}{2}} |\sum_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^T \sum^{-1}(x-\mu_i)}, \tag{14}$$

where μ_i is the mean, $\sum$ is the covariance matrix of the Gaussian, $|\sum|$ is the determinant of $\sum$, $\sum^-1$ is the inverse of $\sum$, d is the dimension of the vector, k_i is the mixture coefficients (weight of the i_{th} Gaussian) and with T denoting the transpose operator. The integral sum across the total feature space is 1:

$$\sum_{i=1}^{N} k_i = 1, \quad 0 \leq k_i \leq 1$$

The two parameters that directly decide the Gaussian distribution are the μ and $\sum$ as shown in equation (16). The Expectation Maximization (EM) algorithm [98], [99] is used to obtain the best Gaussian mixture parameters for a given set of feature vectors. The EM algorithm is an iterative algorithm which is based on maximizing the log-likelihood of the training data with respect to some parameters such as the means, covariances of the matrix and the mixture coefficient [60]. The maximization is attained when the algorithm converges after iterative update of Θ where $\Theta = (\mu_i, \sum_i, k_i)$ [100]. For an assumed feature vector x parameter Θ, the likelihood of the parameters given x is calculated using:

$$pr(x|\Theta) = \sum_{i=1}^{N} k_i pr\left(x\middle|\mu_i, \sum i\right). \tag{15}$$

Therefore, the log-likelihood of Θ given x for all the feature vectors $x = (x^1, x^2, \ldots, x^D)$ generated by the same GMM is [98]:

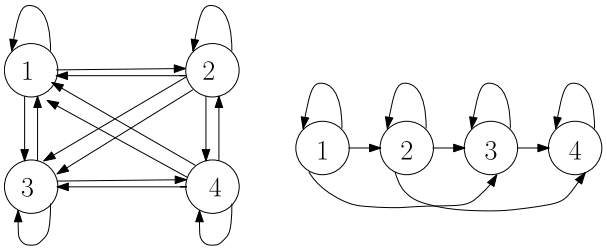
$$\log pr(x|\Theta) = \sum_{d=1}^{D} \log \sum_{i=1}^{N} k_i pr\left(x^{(d)}\middle|\mu_i, \sum i\right), \tag{16}$$

The EM algorithm is now used to find these parameters [98]. Equation (16) is for instance where samples are independent and identically distributed (i.i.d). GMMs have been applied substantially in the analysis of cetacean vocalizations. The feature vectors extracted are treated as probability distribution. 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 N mixture components is empirically chosen; depending on what is aimed to be achieved.

In several research works, GMMs have been used to develop techniques for detecting and classifying the sounds of different species of cetaceans. Peso Parada and Cardenal-Lôpez [12] used cepstral coefficients extracted as feature vectors on a GMM classifier to detect and classify sound recordings of underwater signals into one of four types: pulses, whistles, background noise, combined whistles and pulses. The detection rate achieved was 87.5% for a 23.6% classification error rate (CER). Multiple signal classification (MUSIC) algorithm and an unpredictability measure were introduced as feature vector extractor to address the problem of modeling 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 reduces the CER to 18.1%.

Roch *et al.* [101] developed a GMM classifier to classify free-ranging delphinid vocalizations of four different species of odontocete (long-beaked, short-beaked common dolphins, bottlenose and Pacific white-sided dolphins) from sounds recorded at Southern California bight. GMMs are trained with different mixtures which varies from 64 to 512. The classifier 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 has a positive effect on the accuracy of the classifier. Different output was recorded when the mixture was retain at 256 mixture with 20s test segment while the training data size is varied. The overall precision were impacted with reduction in the amount of training data. For common dolphins, it was noted that the recognition rate was higher with shorter amount of training data.

Detection and segmentation of continuous-wave (CW) underwater signals in the presence of strong background noise was carried out using a spectral feature analysis centered on soft clustering of GMM in [60]. The nature of the targeted signals are continuous, broad spectrum and the frequency is close to the frequency of background noise. These make it difficult to detect the CW signal under the impact of ocean noise reverberation. The energy intensity of the data set was acquired through some windowed FFT process; which is then used as the sample space for the binary clustering of the GMM. The GMM parameter were obtained using the EM algorithm. Their simulation result shows that the detector can produce an accurate detection and segmentation of CW underwater signals even in the presence of strong background noise.



(a) A 4-state ergodic HMM type (b) A 4-state left-to-right (Bakis) HMM type

FIGURE 7. Hidden markov model types.

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 [44]. The experiments were structured in two parts; (1) to detect the specific clicks produced by species of target X and (2) classify the set of clicks produced by particular species. Different GMM mixtures models; 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 detection error tradeoff (DET) curve. DET curve is said to be efficient plots for highlighting divergence between similar systems. DET of the three species were 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) HIDDEN MARKOV MODEL (HMM)

The Hidden Markov Model (HMM) is a popular model that is applied in a wide range of practical applications like speech recognition, speech analysis, data compression, computational molecular biology and pattern recognition, due to its ability to model non-stationary random processes [12], [38], [102]. Over the years, extensive research have been dedicated to its study and this has consequently led to availability of large tool set. HMM can be defined as a probabilistic ranking classifier that assigns a tag or class to every unit in a sequence of observations, therefore calculating the probability distribution over sequence of observations and picks the best observation sequence [103], [104]. Fig. 7 shows an example of two types of 4-state HMM; (1) ergodic (random) HMM, and (2) left-to-right (Bakis) HMM. In the ergodic HMM type, any state can be reached in a single step from any other state in the model, therefore the transition probabilities are non-zero. However, in the Bakis HMM network, the states go from left to right, thus state transitions are constrained to only from lower-numbered state to higher-numbered state [102]. Conventionally, Bakis HMM type is used to model time-varying sound signals [47], nevertheless, both types are still applicable to model cetacean signals.

The following five (5) components are used to describe a HMM [102]:

- Number of states N in a model $R = r_1 r_2 \ldots . r_N$;
- The transition probability matrix $T_p = \{t_{ij}\}$, each $\{t_{ij}\}$ typifying probability of moving from one state t_i to another state t_j;
- A sequence of $X = \{x_1, x_2, \ldots, x_T\}$ observation that is equivalent to the output of the system being modeled;
- The probability of an observation likelihood called the emission probabilities E. It indicates the probability of an observation x_t (physical output) state being generated from a (hidden) state i;
- The initial state distribution $\tau = \tau_1, \tau_2, \ldots, \tau_N$.

Thus, the HMM parameters are represented as:

$$\beta = (T_p, E, \tau) \tag{17}$$

Generally, when a signal is to be modeled with a HMM, three (3) problems are addressed in order to determine the most prospective HMM or HMM sequence given an unknown dataset (cetacean signals) [102]:

1) Estimation of the observation probability- (Likelihood calculation of a certain observation sequence); given an observation sequence $X = \{x_1, x_2, \ldots, x_T\}$ and a model $\beta = (T_p, E, \tau)$;
2) Finding the optimal sequence- (Decoding) from a related state sequence $Q = q_1\, q_2 \ldots . q_T$ of the hidden state from a given observation sequence X and a model β;
3) Choosing the model parameter- (Learning/Training the parameters T_p, E) that maximizes the probability of a specific observation in a given state $P(X|\beta)$.

The above 3 problems are solved using the following algorithms; forward-backward algorithm (Baum-Welch) and Viterbi algorithm. The detailed steps of the implementation of these algorithms with respect to HMM have been treated in detail in [102], [104]. Though, Baum-Welch has been the conventional algorithm used for learning HMM, a new learning algorithm was proposed in [105] centered on non-negative matrix factorization (NMF) which is said to effectively compact data into a statistical matrix. HMM can work on any frame-based feature extraction method such as MFCC, LPC.

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 sound types detected. In GMM, the whole sound is considered as an entity with exclusive properties that characterizes each class.

HMM has been used to analyze cetacean signals of different species due to its ability to manage duration variability through non-linear time alignment. It can also easily manage silence or delay within vocalizations [1], [38], [47]. It has also been used to analyze other bioacoustics signals like fish [106], bird [92] and mammals [107]. It offers extremely robust performances when applied to various bioacoustics signals across a diverse range of species and classification

tasks [38]. Putland *et al.* [1] used HMM technique to detect Bryde's whale vocalizations and it proved to be effective despite the duration difference in Bryde's whale vocalizations and directly overlapping vessel sounds. Classification of individual calls in Humpback whales songs was done using HMM [93]. Different training size was applied to the entire data; 50% 25%, 10%. The classification performance of each training data size differs. Their results show 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 to train, the overall classification output were 90% and 78% respectively. This implies that the size of 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 size give better classification, this can be attributed to the call types; that is, different sound type may require different training amount. However, it has been shown in [93] that minimizing 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 exist opportunity to choose between these two alternatives subject to ones requirement; considering the give-and-take between time and human efforts needed when training and the performance result.

HMM techniques have been compared with other techniques used for cetacean vocalization detection and classification. In most cases, the outcome of most of this comparison showed that HMM outperformed the other techniques; HMM and MF [108], HMM and GMM [37], HMM, SPCC and MF [109]. Though this is also subject to the feature extraction technique deployed and how it is implemented with the HMM [38]. HMM and MF techniques were applied on a set of 189 recorded underwater acoustics signals and white Gaussian noise in [108] 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 (noise that were wrongly classified as bowhead notes); HMM method 51% and matched filter method 49%. These noises are in same frequency band as the bowhead songs and sometimes resemble a portion of a note. Datta and Sturtivant in [110] applied HMM for classification of common dolphin signature whistles. The HMM was trained to represent members of the whistles class after feature extraction using MFCC. However, there is need to check suitability of HMM models for analysis of signature whistles. The dataset used in this work is small compare the data size used for analysis involving HMM.

3) NEURAL NETWORKS (NNs)

Neural networks (NNs) are one of the popularly used machine learning algorithms deployed in different fields to execute tasks such as classification, detection, pattern recognition, prediction and forecasting, optimization problems, among others [111], [112]. The NNs work like the biological neurons of the human brain by transforming inputs to outputs. Similar to the human brain, the NNs is motivated by an activation function. The mathematical neuron calculates a weighted sum of its n inputs of signals x_i, where $i = 1, 2, \ldots, n$ 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

$$g = \theta\left(\sum_{i=1}^{n} w_i x_i - t\right), \tag{18}$$

where θ is a unit step function at 0, w_i is the synapse weight associated with the i_{th} input. Equation (18) is a threshold activation function, also known as the McCulloch-Pitts model [113]. However, in machine learning, instead of the threshold, sigmoid σ is used as activation function:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{19}$$

A large postive value of x gives an output of the sigmoid function that is near 1. The output will be near to zero when x is much smaller than 0. The mathematical concept of how NNs operates can be found in [111], [112]. Many neurons are present in the NNs, each has many weights. The NNs learn through trials and the weights can always be fine-tuned for the NNs to learn (training stage).

The NNs can be categorized into two: (a) feed-forward networks and (b) recurrent (or feedback) networks. Details of this categorization of NNs architectures can be found in [111]. There can also be different connection patterns of NNs and the connection patterns influence the behaviour of the networks. Each of the connection pattern has its specific advantage. The NNs architectures are trained using suitable algorithm to learn about the dataset [111]. The ability of NNs to learn inherent rules from the given collection of dataset is what make them useful to perform various tasks in various fields; which thus makes them appealing. The training of the networks is done with the aim to have an output that is 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-organizing 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 [111], [114]. In unsupervised NNs 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 [114]. Hybrid learning merges both supervised and unsupervised learning. Comprehensive tutorials on NNs have been explained in [111], [112], [115]. Their ability to look for characteristic correlations in the dataset

and form classes base on these correlations make them good candidate for classification of cetacean vocalization [114]. NNs have been used for classification in various applications.

Two types of unsupervised, self-organizing NNs: a competitive network and a Kohonen feature map were used for classification of killer whale vocalizations in [114]. Both networks are trained with a combination of duty-cycle and peak-frequency input values. The outputs of both networks were complementary, not withstanding, each of the network 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 relative distribution of the feature space and topological correlations among categories.

A method was developed for automatic categorization of bioacoustic signals into biologically relevant categories in [116] using a combination of DTW and Adaptive resonance theory (ART) NNs dubbed ARTwarp algorithm. This method modified ART2 with the adaption 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 from 4 individuals bottlenose dolphin stereotyped whistles and field recordings of transient killer whales calls were randomly selected for testing of the method. The problem addressed here is the categorization 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 known biological behavioral patterns of these dolphins.

Jiang *et. al* in [25] put forward a novel technique based on deep Convolutional Neural Network (CNN) for the detection and classification of whistles of long-finned pilot whales and killer whales. CNNs is a class of deep feed-forward neural networks that uses variation 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 which have been earlier achieved during the preprocessing stage. No specific time-frequency features were extracted directly. The data inputs to the models were the time-frequency spectrograms that qualify 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 admit the frame spectrograms of the unknown sounds as inputs and process 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 whales species. Their proposed method was able to attain a 97% detection rate and 95% classification rate. The method is said to be adaptable for other whales or dolphins species that produce whistles or other sounds.

Back propagation (BP) NN is used for classification of clicks of long-finned pilot whales and sperm whales in [32]. BP is a class of multilayer feed-forward network and one of the commonly used NNs models [112]. The designed BP network classifier is a one hidden layer containing four nodes whose output gave an improved classification performance of the feature vectors formed from CWT. The method is also said to give acceptable performance with small training dataset size or a small number of features. The method is applicable to other species of whales or dolphins that emit clicks.

Bermant *et.al* [33] achieved a 99.5% detection accuracy of sperm whales echolocation clicks from annotated 650 spectrogram images of the click data using a CNN based approach to build an echolocation click detector. The technique design the detector to label annotated spectrogram image as 'click' or 'no click'. No feature extraction technique is explored here but the CNN is designed to understand the inherent features needed for detection. The datasets used came from long-term field studies of sperm whales from two different locations: Off Island of Dominica in the eastern Caribbean and Galapagos Islands. They further applied a Recurrent neural network (RNN) approach to carry out three classification types centered on high-quality manually annotated datasets- (1) coda types classification, where 97.5% and 93.6% classification accuracy were achieved for categorizing 23 coda types from Dominica and 43 coda types from Galapagos datasets, (2) vocal clan classification, where 95.3% and 93.1% classification accuracy was obtained for two clan classes from Dominica and four clan classes from Galapagos datasets, (3) individual whale identification, where 99.4% classification accuracy achieved using two Dominica sperm whales. The robust result achieved further emphasis the efficacy of machine learning approach to analysis of cetacean signals.

Deep neural networks-based detector was developed in [41] to detect the vocalization of North Atlantic right calls. Two types of deep neural networks architectures: 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 lower magnitude of false positive rate while exhibiting significantly increasing true positive rate. Deep learning approach to developing algorithms for detection and classification of cetacean signals is gaining more prominence as seen in recent works [25], [33], [41], [96]. Deep learning being an illustrative learning method where machine automatically learns the representations that are needed from the input raw data makes it a promising area to improving on existing techniques for detection and classification of cetacean species. This will help to pass up the preprocessing of the raw data into other forms of input for detector or classifier model [96].

4) SUPPORT VECTOR MACHINES (SVM)

Support vector machines (SVM) is one of the popularly known machine learning algorithms that is based on statistical learning concepts [117], [118]. It has enjoyed wide applications in diverse areas of study such as text categorization, state

estimation, face recognition, image recognition, modeling and control as well as cetacean signal classification among others [117]. SVM is defined in [118] as systems which utilize hypothesis space of a linear functions in a high dimentional feature space, trained with a learning algorithm from optimization theory that implements a learning preference derived from statistical learning theory. The SVM formulation uses the structural risk minimization (SRM) principle [119] which minimizes the upper bound on the expected risk. It has good classification ability through the use of the kernel function mapping technique [119]. The SVM classifier's performance is heavily subject to parameter selection and settings. Detail explanation on the working principles of the SVM can be found in [117]–[120].

The SVM technique has been used to solve classification problem in the analysis of signals of cetacean species. The SVM classifier was used in [121] for classification of North Atlantic Right whale up-calls. The authors proposed a novel algorithm by integrating two feature extraction techniques; DWT and MFCC. The MFCC was enhanced with the introduction of DWT which assist in separating ocean noise from the up-calls. A 92.27% detection rate was achieved with this algorithm. SVM produce good performance for datasets with overlapping characteristics. The performance output of a GMM and SVM classifiers were compared in [44] using detection error tradeoff (DET) curve. The equal error rates (EERs) from the plotted DET of three species (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%) indicated that GMM perform better than SVM. It was proposed that adding more species to the dataset may improve the performance of SVM.

A multi-class SVM classifier was developed in [122] to classify vocalizations 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 maximized decision function with respects to the reference class. The CS-SVM was structured to recognize 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 labeled and unlabeled 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 labeled feature vectors and a similar number of noise-only feature vectors. The training set were used in the optimization 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 unlabeled 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 unlabeled dataset was reported not to be as good as the result obtained from labeled dataset. This was attributed to the difference in the selected feature set. Therefore, the method can be further tested with real live data set that will indicate the location and method of recording of the data.

Humpback whale vocalizations were recently classified into song and non-song in [123] using three machine learning techniques: SVM, NNs, and Naive Bayes classifier. Humpback whale vocalizations can be grouped into either song or non-song. The song vocalizations are series of calls organized into hierarchical structure that can go on for several minutes while non-song vocalizations include feeding cries, bow-shaped and downsweep or meow moans. 70-second duration 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 file in the dataset which was recorded from the Gulf of Maine. The performance of each of the technique was evaluated using the accuracy, receiver operating characteristics (ROC) curve, and area under ROC curve (AUC). The performance of the three classifiers were compared, the MFCC-based SVM led with 94% accuracy while the MFCC-NN led with 94.27% AUC. The authors intend to ascertain the generalization of their approach to data from other regions of the world.

B. THRESHOLD-BASED TECHNIQUES

We categorized techniques that correlates their model output with an already known template of the signal in view to determine when detection or classification occur. For techniques under this categorization, the detection or classification patterns are set for the model through a define threshold. They require a *priori* hypothesis of the structural pattern of the signal. The model will then search for correlation in the structure of the dataset and the known template. Usually, this is done correlating the spectrogram of the signal recorded with an already human annotated template that has set the parameters for the signal of interest. The threshold as defined by a human expert from the known template of the signal decides the output of the model. Detection or classification will occur when an energy within a specified frequency band has exceeded the preset threshold. Techniques such as SPCC, DTW,and MF among others falls under this categorization.

1) SPECTROGRAM CROSS-CORRELATION (SPCC)

Spectrogram Cross-Correlation (SPCC) is a correlation technique that is popular among existing methods for bioacoustics detection and classification due to its implementation simplicity. It only require a single sound sample of call type to be detected. It can be effectively applied to either continuous detection or isolated vocalization task, where recordings have been presegmented into separate files. A sliding window is applied across a long recording, with correlation peaks signifying target detection [38]. A spectrogram conveniently represents signals by interpreting it in a waveform as a non-negative function of instantaneous frequency and time.

Generally, an input sound signal is converted into a spectrogram, a conditioning procedure; level equalization and normalization are then applied. A template vocalization is cross correlated directly with a spectrogram [124]:

$$S(t,f) = \left[\sum_{n=0}^{N-1} w(n)x(t+n)exp\left(\frac{-2\pi jnf}{N}\right)\right]^2, \quad (20)$$

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

$$d(t) = \sum_{t_1}\sum_{f} S(t+t_1,f)k(t_1,f), \quad (21)$$

A threshold is then applied and the times at which the detection function goes over the threshold are considered detection events (presence of sound of interest) [125]–[127].

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, [109], [126],
- When the desired output is to minimize the number of missed calls (false negatives) [126],

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 [38], [127]. It is also subject to ocean acoustic propagation effects, i.e., signal distortion as the sound propagates through the ocean medium.

The performance of SPCC technique developed in [109] was compared with that of the matched filter and HMM using a set of 114 songs sample of bowhead end notes. Each of the technique 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 a better performance than the matched filter. The SPCC was also compared with neural network technique but with a lager dataset, the neural networks however performed better than the SPCC. The better performance of neural networks is due to its ability to manage time variation in bowhead vocalization better than the SPCC perhaps as a result of its large training set. The SPCC works fairly well with small training set, therefore, this technique can fit in well in situation where small data recordings is available.

Mellinger in [126] used two different approach to develop SPCC method which was applied to detect right whale calls: the manual parameter choice and the automated optimization procedure. For the manual parameter choice procedure, parameters controlling the spectrogram correlation procedure were selected by hand, in a series of sequential steps, while in the automated optimization procedure, parameters for the spectrogram correlation were chosen by running an optimization process to obtain the set of parameters that will perform best. The performance of the automated optimization procedure was substantially better than the spectrogram correlation of the manually chosen parameter. This is because the automatic optimization procedure uses the whole datasets to choose its parameters while the manual procedure only use a small subset of the datasets to choose it parameters. The two SPCC approaches were compared to neural network method with the latter performing better than the former for calls with poor SNR.

SPCC performed best on short sounds recording when used for automatic detection of North Pacific right whale calls [36]. It detects 17 calls out of 18 calls samples. However, the proportion of false and missed detections increased as the recording duration increases because longer duration recordings contain longer period of noise relative to the number of right whale calls present. This stress the fact that SPCC works better when few datasets are available.

Mellinger and Clark [127] developed detector for automatic detection of bowhead vocalization using SPCC and matched filter techniques. A set of bowhead sound were analyzed with both techniques. SPCC performed relatively better than the matched filter in a substantially noisy recordings. The matched filter on the other hand performs better when the noise present in the recording is flat-spectrum. The same set of bowhead sounds were tested with a NN method. The NN performed better. However, SPCC and matched filter work better than neural network when a few data set are available.

2) MATCH FILTERING (MF)

The match filtering (MF) technique is similar to the spectrogram cross correlation; it is also a correlated-based technique. It is used in radar system and image processing analysis. In this technique, synthetic waveforms or synthetic spectrograms are used rather than sounds edited out of the original recordings [38], [109]. The matched filter detects signal of interest by maximizing the SNR of the input signal with respect to the noise present in recordings. This is followed by cross-correlating the sound signal with the known template signal. Generally, the objective of matched filtering technique is to detect presence of sound $s(t)$ in the received signals $x(t)$ which is contaminated with additive noise $n(t)$ as [128], [129]:

$$x(t) = s(t) + n(t). \quad (22)$$

With respect to cetacean detection; in Equation (22), $x(t)$ represents the dataset to be analyzed for the presence of species, $s(t)$ represents the known template signal associated with the species of interest, and $n(t)$ represents all other signals (such as shipping sounds, sounds from other species) present in the dataset. There is prior knowledge of the signal of species of interest (which serves as template), but the shape of such templates may vary within the species of interest. Due to this prior knowledge, the detection can be achieved based on a matched filter. A matched filter can be achieved using a finite impulse response (FIR) digital filter which has an impulsed response, $h(t) = s(t - t_0)$. That is, the time-reverse of the

template signal, such that the output SNR when $x(t)$ is applied to the input is maximized [130]. The resulting output of the filter $y(t)$ is expressed as:

$$y(t) = y_s(t) + y_n(t), \tag{23}$$

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, $y_n(t) = n(t) * h(t)$ is the noise output of $x(t)$ [128], [130]. The condition for optimal detection is that the output signal component $y_s(t)$ must be substantially greater than the output noise component $y_n(t)$. 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 compare to the average power of the output noise $y_n(t)$. This is equivalent to maximizing the peak pulse SNR as shown in Equation (24); [128]

$$r_0 = \frac{|y_s(t_0)|^2}{E[n^2(t)]}. \tag{24}$$

where E is the signal energy. The matched filter is an optimal detector if $n(t)$ is a white Gaussian noise random variable. In other words, it is optimum detector for detection of sound with white Gaussian background noise and a known signal [127], [130]. However, it shows a poor performance level than other techniques when there is variation in the recordings which can be as a result of harmonic interference such as ship noise [109]. Matched filtering also gives optimal performance for signals with known source. However, the exact source signal of marine mammals are not known which implies that matched filtering may not be an optimal detector for cetacean signal analysis. Despite this shortcoming with respect to it has been applied in the detection of some cetacean species where the matched filter performs cross-correlation between the input signal $x(t)$ and the 'targeted' sound of interest $s(t)$. The target sound of interest, selected by a human expert serves as template for the matched filter design. Any match with this template defines the signal of interest. Measured parameters such as mean frequency and bandwidth are used to synthesize a filter kernel representing the sound type [127]. It strengths and weaknesses are similar to that of SPCC [38]. Although, it requires more efforts to construct the pattern templates but it is easy to implement. Stochastic matched filter (SMF) [131], [132] method, which is a better version of the traditional MF has been proposed in analyzing PAM of cetacean signals. SMF is attractive due to it ability to both detect and classify. It was used in analyzing recorded data in noisy underwater environment containing Antarctic Blue whales [132]. It has been shown to give an improved performance when compared with MF.

Stafford, Fox and Clark in [133], developed a matched filter using recorded blue whale calls (for template design) from the U.S. Navy's Sound Surveillance and System (SOSUS) for detecting and localizing such calls in noisy environments. The developed matched filter was applied to real-time recordings from three different SOSUS arrays off the coast of the Pacific Northwest to detect and locate blue whale calls. A match was found between the calls recorded by patrol aircraft and the SOSUS which were confirmed to be the same. The matched filter developed has the advantage to detect the calls of interest from noisy data. This is made possible because the matched filter was designed from average values of numerous blue whale calls obtained from different SOSUS arrays. This makes it to be more robust for detection than a kernel developed from a single calls.

MF and HMM techniques were applied on a set of 189 recorded underwater acoustic signals and white Gaussian noise in [108] to detect the presence of bowhead whale notes. The performance of the two techniques were compared where the HMM selected 97% and the matched filter selected only 84%. Both methods give almost the same misclassification output (noise that were wrongly classified as bowhead notes); HMM method 51% and matched filter method 49%. These noises are in same frequency band as the bowhead notes and sometimes resemble a portion of a note.

3) DYNAMIC TIME WARPING (DTW)

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 [134]. DTW can be applied for the analysis of temporal sequences of video, audio or image data. It has been explored in many areas of applications like speech recognition, handwriting and online signature matching, sign language recognition, gestures recognition, data mining and time series clustering (time series databases search), computer vision and computer animation, surveillance, protein sequence alignment and chemical engineering, music and signal processing [135]. Suppose we have two time series, a sequence $\mathbf{R}$ of length n and a sequence $\mathbf{G}$ of length m, where

$$\mathbf{R} = (r_1, r_2, \ldots, r_i, \ldots, r_n) \text{ and } \mathbf{G} = (g_1, g_2, \ldots, g_j, \ldots, g_m). \tag{25}$$

The optimal match between these two sequences can be found using DTW, an $n - by - m$ distance matrix $\mathbf{Z}$ is constructed as shown in Equation (26), 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.

$$\mathbf{Z} = \begin{bmatrix} z(r_1, g_n) & \cdots & z(r_m, g_n) \\ z(r_1, g_{n-1}) & \cdots & z(r_m, g_{n-1}) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ z(r_1, g_1) & \cdots & z(r_m, g_1) \end{bmatrix}, \tag{26}$$

The path through the matrix that minimizes the total cumulative distance between them is the optimal path. The shorter the distance, the more similarity between points, and vice versa [136]. The distance matrix $\mathbf{Z}$ is solved to get a contiguous set of matrix $\mathbf{W}$ with elements $\mathbf{W} = (w_1, w_2, \ldots, w_k,)$ that represents a mapping between $\mathbf{R}$ and $\mathbf{G}$. The connection of

each element is called the warping path (warping distance of dynamic time) $\underset{\sim}{W}$. The k^{th} element of $\underset{\sim}{W}$, $w_k = (i.j)_k$ is the alignment of the i^{th} point of series $\underset{\sim}{R}$ and j^{th} point of series $\underset{\sim}{G}$. 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 the following conditions [135]–[137].

1) Boundedness condition: The length of warping path $\underset{\sim}{W}$ should be within this range $\max(m, n) \leqslant K \leqslant m + n - 1$;
2) Boundary conditions: The starting point of warping path $\underset{\sim}{W}$ is $\underset{\sim}{w}_1 = (1, 1)$ and the end point is $\underset{\sim}{w}_k = (m, n)$, that is, the alignment path starts at the bottom left and ends at the top right. This guarantees that the alignment does not consider partially one of the sequences.
3) Continuity condition: Suppose the previous point is $\underset{\sim}{w}_{k-1} = (i', j')$, next point $\underset{\sim}{w}_k = (i, j)$ in warping path, then there must be $(i - i') \leqslant 1$ and $(j - j') \leqslant 1$. This condition restricts the allowable steps in the warping path to adjacent cells. It ensures important features are not omitted by the alignment.
4) Monotonicity condition: Suppose the previous point is $\underset{\sim}{w}_{k-1} = (i', j')$, point $\underset{\sim}{w}_k = (i, j)$ in warping path, then there must be $0 \leqslant (i-i')$ and $0 \leqslant (j-j')$. This condition preserves the time-ordering of points. This guarantees that features are not repeated in the alignment.

The optimal path is the path that minimizes the warping cost

$$DTW(\underset{\sim}{R}, \underset{\sim}{G}) = \min \left\{ \sqrt{\sum_{k=1}^{k} \underset{\sim}{w}_k} \right. \quad . \tag{27}$$

Details on the working procedures of DTW are explained in [135], [137]. DTW's ability to distinguish the differences in signals of similar contours but different length makes it a good technique for classification of signals of different cetacean species. It was first applied for classification of 15 bottlenose dolphin whistles in [28] by comparing the frequency contours. The frequency contours of the signals were extracted using the STFT. The algorithm was able to correctly classify the 15 whistles into five classes.

DTW showed great performance in automatic classification of killer whale pulsed calls by presenting precise measurement of differences in the calls [138]. The pulsed calls are complex sound with many harmonics which make their classification more challenging than whistles. Five calls with high SNR from previously classified sounds of captive killer whale from Marineland of Antibes, France were used for the experiment in the work. DTW was used to relate the pulsed calls contours' fundamental frequencies of 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 [139] broaden what was done in [138] by investigating the effectiveness of DTW algorithms on more natural recordings that include diverse collection of species. The DTW was implemented using four different approaches; (1) Ellis method, (2) Sakoe-Chiba method, (3) Itakura method, and (4) Chai-Vercoe method on large dataset. The four algorithms give good classification between 70% to 90% despite the presence of biphonic calls and over 100 calls. The results show the versatility of DTW for 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 [52] developed a detector using DTW and LPC algorithms for continuous recording of Bryde's whale short pulsed calls (< 3:1s long). The data were weekly recorded for a period of five months on sighting of the Bryde's whale in a single site in the Gordon's bay harbour, False bay, South-West of South Africa. 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 sound) is used to test the detector performance. Each template has a k numbers of sample 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 when 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 performance. However, the DTW-based detector performed lower than the LPC-based detector.

C. SUMMARY OF THE TECHNIQUES

In conclusion to this section, it must be noted that the list of the survey in this work are some of the commonly used techniques for detection and classification of cetacean species, it is by no means the entire techniques. It is noted that while all the threshold-based techniques do not require any specific feature extraction method, some of the statistical-based techniques can as well carry out detection and classification on their own. Furthermore, it has been proven that the techniques for analysis of cetacean species signals can be very adaptive as recently shown in the work of [52] and [39] where LPC-based and EMD-based methods were deployed for detection and classification. These methods have been previously known to be deployed for extraction of features from signals.

Table 2 shows summary of the characteristics of the detection and classification techniques reviewed in this paper. The conclusions to the information provided in the table are arrived at from the survey of past work. 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 cetacean consists of two suborder in their taxonomy; *Odontocete* and *Mysticete*. The species applicable to each of the technique are grouped according to their taxonomy. In some of the techniques, feature extraction 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 remark on each of the technique are also included in the table.

V. OUTPUT PARAMETERS

The metrics used to determine the performance of each method develop and tested varies. The detection and classification of cetaceans sounds are important as mentioned earlier. Different methods have been developed as discussed in this review to carry out this important tasks. The detector or classifier output are characterized by certain parameters that determine their accuracy level. However, the performance measurements reporting style varies among authors. This can be attributed to the use of different techniques for analysis of these signals.

Like every scientific research, the outcomes of researches in this field are reported in terms of observable parameters. These parameters or metrics are to show the level of accuracy attained in the detector/classifier designed. Though, there are a number of reporting metrics such as true positive, true negative, false positive, false negative also known as *missed calls*, false positive rate (FPR), and true positive rate (TPR) also known as *Sensitivity* that are usually adopted by authors to report the output of their work. Authors do adopt their own style in reporting outputs. No method is 100% perfect, any method will produce false negative, false positive [48] depending on analyzed signals, feature extraction technique or types of detector and classifier.

The common metrics used in reporting output are as follows.

1) **False positives**: This is when a detector wrongly detects a signal as signal of interest; that is, the number of sounds wrongly detected as sound of interest.
2) **False negatives**: This occur when the sound of interest are missed. The false negatives are also known as the *missed calls*.
3) **True positives**: This is rightful detection of sound of interest by the detector; that is the number of sounds correctly detected.
4) **True negatives**: This is rightful prediction of the absence of sound of interest by the detector.
5) **False Positive rate (FPR)**: This is the proportion of sound of interest that is falsely detected by the the detector; it can be mathematically expressed as

$$FPR = \frac{false\ positives}{true\ positives + false\ positives}.$$

The best FPR is 0.0 while the worst is 1.0.

6) **True Positive rate (TPR)**: This is the proportion of sound of interest that is rightly detected by the the detector; it is also known as sensitivity. It can be mathematically expressed as

$$TPR = \frac{true\ positives}{true\ positives + false\ negatives}.$$

The best TPR is 1.0 while the worst is 0.0.

7) **Error rate (ERR)**: This is computed as the total number of all wrong predictions (false positives and false negatives) divided by the total number of the dataset; it can be mathematically expressed as

$$ERR = \frac{false\ positives + false\ negatives}{total\ number\ of\ the\ dataset}.$$

The best ERR is 0.0 while the worst is 1.0.

8) **Accuracy (ACC)**: this is computed as the total number of correct prediction (true positives and true negatives) divided by the total number of the dataset; it can be mathematically expressed as

$$ACC = \frac{true\ positives + true\ negatives}{total\ number\ of\ the\ dataset}.$$

The best ACC is 1.0 while the worst is 0.0.

9) **Precision (PREC)**: this is computed as the total number of rightful detection divided by the total number of detection (both true detection and false detection) in the dataset; it can be mathematically expressed as

$$PREC = \frac{true\ positives}{true\ positives + false\ positives}.$$

The best PREC is 1.0 while the worst is 0.0.

Each designed detector or classifier has a unique set of threshold or sensitivity subject to what it is intended to be achieved. For instance, in a survey of a relatively rare species such as *Right Whales*, the detector may be configured in such a way that there as low number of *missed calls* (false negatives) as possible due to availability of few datasets but such detector may have a large number of false positives. On the other hand, in a survey of a common species with abundant datasets available, where 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 false positive detection and achieve high TPR [24]. However, some authors such as in [39], have been observed to report the output of their research using some other metrics to compare performance of their detector or classifier. Though, this is does not really matter, in as much as there is a clear reporting of the metrics use in determining the performance level of their detector or classifier.

VI. FINDINGS, CHALLENGES AND FUTURE RESEARCH PROSPECTS

The threats faced by cetaceans are enormous due to increasing human anthropogenic activities. Ecosystem managers are concerned about mitigating these threats but have challenges of inadequate information about them. This is because cetaceans spent most of their time in water. However, they use vocalizations essentially for their daily activities such as communications, echolocation and social interactions. Researchers have thus explored the advantage of these vocalizations to study them from a distance. PAM is used for their monitoring and observations. However, PAM system usually resulted in having large datasets can be difficult to

TABLE 2. Summary of surveyed detection and classification techniques.

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 sound is considered as an entity with exclusive properties that characterizes 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 vocalizations.	•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. •Minimizing the amount of training can enhance the efficiency in terms of the classifier.
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 overfitting 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 overfitting if training goes on for too long.
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.
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 detection of call types when a relatively few instances of call types are known. •Performs well when the desire output is to minimized missed calls (false negative).
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 recognizing 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 recognized.

manually analyzed. Design of automatic techniques for the detection and classification of cetacean vocalizations has greatly assisted with the processing of large acoustics dataset that are acquired from long time recordings. It has also helped eliminates human bias and errors thus leading to fast computational time and relative consistencies. Therefore, it has assisted the ecosystem managers in having more information about their ecology. Thus, enhancing knowledge about their biological behaviour, measurement of range and seasonal presence, estimation of abundance of species within a specific area and so on. These techniques are based on signal processing, pattern recognition and recently machine learning algorithms. No detection or classification technique has given 100% output in terms of performance level. We observed that the performances of the techniques are influenced by various factors. These factors include feature extraction method deployed, species involved, the quality of recording, size of the recording, location of recording and computational

TABLE 3. Summary of past work surveyed on detection and classification techniques.

Reference	Task Performed	Recording method	Feature Extraction Technique	Detection Classification Technique	Species analyzed	Data Collection Location	Sound Type	Remark
[1]	•Detection	Fixed PAM	•MFCC	•HMM	•Bryde's whale.	•Hauraki Gulf, New Zealand.	•Pulsed calls.	Hidden Markov model Tookit (HTK) was applied for the detection of Bryde's whales vocalization. The method proved effective despite duration difference in Bryde's whale vocalization and overlapping passage vessel sounds. The work showed that HTK could further be explored for analysis of signals from other species of cetaceans. A further work is needed to ascertain the robustness of HTK to detection of other species of cetaceans that produce characteristics calls like the Bryde's whale.
[11]	•Detection •Classification	Mobile PAM	•STFT	•Threshold	•Short-beaked common dolphin •Melon-headed whale, •Bottlenose dolphin.	•North Atlantic regions.	•Whistles.	Noise cancellation technique was applied to spectrogram of data to search for connected regions which rise 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 whistles detection or partially detected whistles. However, classifier performance deteriorates with increase in the number of species mix in the dataset.
[12]	•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 sounds 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 narrow bandwidth in whistles, however the introduction of MUSIC algorithm and unpredictability measure improved the results to 18.1% CER and 90.3% TPR. The viability of MUSIC algorithm and unpredictability measure can be further explored in future research to analyze high frequency signals with narrow bandwidth.
[25]	•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 Network (CNN) was developed for the detection and classification of whistles of killer whales and long-finned pilot whales. No feature extraction technique used to extract features directly. Though, STFT was used to create whistle contours in annotated the spectrogram. The data inputs to the models were the time-frequency spectrograms that qualify the whole information of whistles. Therefore, the feature extraction structure and the computed featured were learned directly from the training data. A graphic user interface (GUI) was developed to aid visualization of the results of the detection and classification procedure. The accuracy of this method is however dependent on large training data size.
[141]	•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 GMM classifier for classification different odontocetes species using recordings of their echolocation clicks was proposed. The performance of the algorithm differs from species to species.
[93]	•Classification	Fixed PAM	•MFCC •LPC.	•HMM	•Humpback whale.	•Ste. Marie Island, Madagascar.	•Songs.	HMM was applied to analyze humpback whale songs. The training data size directly influence classification output. Large training set lead to high performance. However, there are other instances where the sound type also influence the classification output. The data used have 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.
[37]	•Classification	Not stated	•MFCC	•HMM •GMM	•Killer whale.	Not stated.	•Pulsed calls.	HMM and GMM were used for classification a set of 75 calls of Northern Resident Killer whales into call seven types. Their results establish that both GMMs and HMMs are greatly effective in the task of automatic classification of killer whale call styles, though HMM performs more 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 call types detected. GMM on the other hand, considers the whole sound as an entity with exclusive properties that characterizes each class. The details of the data used were not given, these could have helped in giving more insight into the work.
[126]	•Detection	Fixed PAM	•STFT	•SPCC •NN	•Right whales.	•Jacksonville Florida, USA, •Massachusetts, USA.	•Pulsed calls.	The performances of SPCC and NNs techniques were compared on Right whale up calls. Due to influence of SNR on detector performance, testing were done separately for calls of different SNR. Results show NN performing better than the SPCC with only 6% error rate compare to SPCC 26%. However, NNs require more dataset for training while SPCC require only a few dataset to give good performance. Good explanation of detection process using SPCC technique for cetacean sound analysis is given.
[109]	•Detection	Fixed PAM	•STFT •MFCC	•SPCC •HMM •NN •MF	•Bowhead whale.	•Alaska, USA.	•Songs.	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 dataset for training. The suitability of SPCC for extensive analysis of cetacean signals cannot be guaranteed due to it poor handling of time variation.
[114]	•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-dimentional feature map were used for classification of killer whale vocalizations. 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 network 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 relative distribution of the feature space and topological correlations among categories. The unsupervised do not require any FE technique because the NNs are able to detect patterns in their inputs.
[76]	•Detection •Classification	Fixed PAM	•HHT	•HHT	•Sperm whale.	Strait of Gibraltar, Spain.	•Clicks.	HHT ability to decompose non-stationary signal make it one of the alternatives of time-frequency methods. It is used to detect sperm whale clicks in continuous signal. HHT provides better time-frequency localization than STFT and WT which leads easy interpretation of result than STFT and WT. However, the method is faced with masking problem in the presence of high energy component in the signal.
[39]	•Detection •Classification	Not stated	•EMD	•EMD	•Beluga whale, •Sperm whale, •Humpback whale.	•Bering Sea, USA.	•Whistles, •Clicks, •Songs.	A novel method of for detection and classification using only EMD process. 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 Hilbert transform and manual labeling of pre-processed data by an expert. However, it performs poorly with the presence of extreme values in the signals. Therefore, this method can further be explored on analysis of other species of cetacean to ascertain it robustness, particularly as regard data size.

TABLE 3. ***(Continued.)*** **Summary of past work surveyed on detection and classification techniques.**

Reference	Task Performed	Recording method	Feature Extraction Technique	Detection Classification Technique	Species analyzed	Data Collection Location	Sound Type	Remark
[42]	•Detection	Mobile PAM	•MAP	•HMM	•Bryde's whale.	False bay, South Africa.	•Pulsed Calls.	A novel feature extraction method was adapted with HMM for detection of Bryde's whale pulsed calls. The new feature extraction method draw on the strength of three simple but robust parameters: the mean, relative amplitude and relative power (MAP) of the signal to be detected. The selection of the parameters were based on empirical observation of the calls to be detected. The result besides showing low computational complexity also gave a higher sensitivity when compared with existing LPC-HMM and MFCC-HMM detector.
[32]	•Detection •Classification	Not stated	•CWT	•BP-NN	•Sperm whale, •Long-finned pilot whale.	•Not stated.	•Clicks.	Back propagation (BP) NN is used for classification of clicks from complex overlapping sperm whales and long-finned pilot whales vocalizations found in the same dataset. CWT approach was used to decompose picked clicks of sperm whale and long-finned pilot whale and 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 data size greatly influence the performance.
[71]	•Detection	Not stated	•CWT •STFT	•STWE	•Sperm whale.	•Strait of Gibraltar, Spain.	•Clicks.	A novel detection algorithm Short-Time Windowed Energy (STWE) for characterization of particular shape present in time-frequency domain of sperm whale clicks. Further work is recommended on this technique to determine it viability for global application in cetacean signal analysis.
[44]	•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 cetaceans species' clicks are classified using GMM and SVM. TEO was to locate individual clicks while MFCC was used deployed in the construction of feature vectors for the classifier. The results from each classifier were compared using detection error tradeoff (DET) curve. DET curve is efficient plots for highlighting divergence between similar systems. However for SVM, some clicks might have being wrongly classified due to absence of distributional tactic.
[101]	•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 varies from 64 to 512. The classifier accuracy increases with increase in the number of mixtures per GMM. The overall and mean species accuracy were impacted with reduction in the amount of training data. The method developed here can further be tested on data from other location to know the impact of the relativity in the species signals.
[142]	•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 overcome problems of fixed-size input. The method can be tested with dataset that contain real-life noise to ascertain it effectiveness.
[33]	•Detection •Classification	Not stated	•No FE used	•NNs	•Sperm whale.	•Eastern Caribbean, Dominica. •Galapagos Island, Ecuador.	•Clicks.	A robust NNs technique using CNNs and RNNs approach were used to build 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 used. This method is promising for analysis of large sperm whale clicks from raw recordings.
[132]	•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, accurate estimation of the background noise is needed in order to maximize the SMF output.
[52]	•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 novel because LPC method has been hitherto used for feature extraction.
[110]	•Classification	Mobile PAM	•MFCC	•HMM	•Common dolphin.	•Not stated.	•Whistles.	The work demonstrate the use of HMM for classification of common dolphin signature whistles. There is need to further look into the suitability of this technique for signature whistle classification with large dataset.
[116]	•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 categorize bioacoustics recordings containing bottlenose dolphins whistles and killer whales pulsed calls into biologically significant categories. However, the method can only address peculiarities of acoustics perception rather than behaviour of the species. The method cannot classify individual signature whistles.
[122]	•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) is 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 live recording as the location and method of recording of the data used were not stated.
[143]	•Classification	Fixed PAM	•MFCC	•SVM	•Humpback whale.	•Chi-chi island, Japan.	•Songs.	SVM classifier centered on cepstral coefficients 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 [144] where cepstral features are used on GMM classifier.
[138]	•Classification	Fixed PAM	•STFT	•DTW	•Killer whale.	•Marineland of Antibes, France.	•Pulsed calls.	DTW showed great performance in automatic classification of killer whale pulsed calls by presenting precise measurement of differences in the calls. The dataset used here was from captive killer whales. A further work has been done using more natural recordings that include diverse collection of species in [139] to test the robustness of this method.
[139]	•Classification	Fixed PAM	•STFT	•DTW	•Killer whale.	•Northern Resident, Canada.	•Pulsed calls.	An extension of what was done in [138] to ascertain robustness of DTW calls classification. Four approaches were used to develop the DTW which were tested on a larger dataset recorded on open sea. The results achieved show the versatility of DTW for analysis of calls killer whales signals. Further experiments can be carried out for calls of other species of cetaceans.
[145]	•Detection	Fixed PAM	•MFCC	•GMM	•Blue whale.	•Guafo Island, Chile.	•Pulsed calls.	GMM used for detection of Blue whale calls through clustering of features into sets. The work demonstrated the strength of unsupervised GMM in detecting Blue whale calls from recording containing sounds from other sources by separating the calls in clusters. Other clustering approaches could be explore in later work to confirm the number of clusters that can be formed from a specie sound.
[121]	•Detection •Classification	Not stated	•MFCC •DWT	•SVM	•North Atlantic Right whale.	•Florida, USA.	•Pulsed calls.	An integrated algorithm which combines two feature extraction techniques as well as SVMs as classifier is proposed for detection and classification of North Atlantic Right whale calls. A more precise and faster computational time were achieved with this algorithm.
[146]	•Detection •Classification	Fixed PAM	•MFCC	•HMM	•Blue whale.	•Corcovado Gulf, Chile.	•Songs, •Pulsed calls.	A new approach of HMM for detection and classification of Blue whale is introduced using the Kaldi speech recognition toolkit. The kaldi speech recognition toolkit looks promising for analysis of other cetacean signals.
[45]	•Detection	Mobile PAM	•EMD	•HMM	•Bryde's whale.	•False Bay, South Africa.	•Pulsed calls.	EMD method was used to extract features from Bryde's whale pulsed calls. These features were fed into HMM to carry out detection. The EMD-based FE method enhance the performance efficiency of the HMM when compared with existing LPC-HMM and MFCC-HMM.

requirements. We also observe that the procedures adopted in implementation of the techniques by researchers varies as reported in some of the work. Some of the techniques may have more than one algorithms for their implementation. An example is the DTW technique used for the classification of killer whale vocalizations in [139], four algorithms are used to implement the DTW, with each giving different outputs. Furthermore, researchers often use more than one techniques to carry out detection and classification on a set of data and compare performance output of each of the technique deployed. An instance of such scenario is in [46] where five different classifiers: SVM, CNN, Long-short-term memory (LSTM) network, logistic regression, and decision tree are used to classify large volume of fin whale vocalizations. The comparison is carried out using multiple metrics including accuracy, precision, recall, and so on. Recently, it has been proven that the techniques for analysis of cetacean species signals can be very adaptive as shown in the work of [52] and [39] where LPC-based and EMD-based methods were deployed for detection and classification. These methods have been previously known for extraction of features from signals in analysis of cetacean vocalization. Another promising future approach to this area of research is the deployment of deep learning techniques as seen in recent works [25], [33], [41], [46], [96] on analysis of signals of various cetacean species. The deep learning avoids the preprocessing stage by extracting features directly from the raw data [96]. It has been reported to be efficient in carrying out detection of species signals in challenging datasets [41] due to it ability to manage complex, diverse, and unstructured data.

Despite the availability of different detection and classification techniques, there are still some challenges facing this research area. The comparison among different techniques for performance outputs can be difficult due to the feature extraction technique deploy. These feature extraction techniques often universal [26] in approach (e.g. MFCC, LPC, HHT) but some adjustments can be done to them to fit in properly to the characteristics of signals to be analyzed. Feature vectors play pivotal role to the output of any detector or classifier [38]. Thus the feature extraction techniques must be carefully selected. Background noise as a result of harsh recording surroundings, hydrophone type deploy and sound propagation unpredictability can be inimical to the detection and classification process. Some of the feature extraction techniques can be very susceptible to noise which will in turn reduce the precision of the system. Furthermore, there is need to make available more wide-ranging, expert-certified species sound catalogs such as the one available on *MobySound.org*. The collection of sound can be laborious and expensive for signal analysts to carry out. This can be addressed if experts (biologists) can have a collaborative efforts to develop harmonized catalogs for different species. These catalogs can lead to launching of centralized database (with information such as date/time of recordings, location, and so on) that can be made accessible to researchers working on development of automatic detectors and classifiers.

All the techniques have been used in a number of fields for different applications. The non-stationary characteristics of the cetacean signals often influence the use of some of these techniques in this subject area. There are a number of prospective areas for future research (deep learning techniques as stated earlier for example) which can lead to discovery of new techniques, thus increasing available options for cetacean signal analysis. We also observed that despite increase in efforts to develop robust techniques, there is need for collaborative efforts that will standardize outputs of research by different authors. The difference in the reporting of work done by researchers are sometimes confusing. The metrics use to report the performance level of detector or classifier by different authors can also be standardize for ease of comparison. The implementation procedures are not usually clearly stated. Most of the techniques are species-specific in application and in some instance, location-based. There is need to focus on developing robust techniques that can cover many species despite the variability in the nature of their sounds. More research efforts focusing on potentials of adopting techniques for analyzing non stationary signals in other field subject areas can give good lead to new discovery. Continuous improvement on feature extraction techniques will enhance accuracy of existing detection and classification techniques. There is need for researchers to deep more into exploring the opportunities of adapting feature extraction techniques used in other research areas into this field. Such prospective techniques such as dynamic mode decomposition (DMD) [140] must be those that have the characteristics of extracting feature vectors from non-stationary and non-linear signals. There is need for continuous research efforts that will lead to more improved techniques that can fit into a number different species with similar characteristics and features. The computational complexity of some the techniques need to be improved.

We surveyed many previous works on automatic techniques for the detection and classification of different cetacean species. Table 3 shows a summary of a few of the papers surveyed in this work. In these surveyed work, we noted the tasks implemented; detection or classification or both. The recording method used is also stated; either fixed PAM or mobile PAM. The feature extraction, detection, and classification techniques as well as the species involved are also stated. In some instances, more than one technique is used in a work. This is usually done to compare the performances between the techniques so as to take a more informed position on a particular technique. This is also applicable to the species used for analysis where more than a species is tested on same techniques in order to observe the difference in their performances. General remark on what was done in each of the survey is given. This include suggestions on areas of possible improvements in future research. The remark column is also to give readers a quick overview of each of the surveyed work. Though, there are instances where some of these information such as recording methods or location are not stated in the literature. The missing information are

mostly observed in situation where authors are provided with datasets from some institution data catalogs.

VII. CONCLUSION

The importance of cetaceans have led to increase interest of researchers in them over the years. Despite their importance, they are continually exposed to threat from human anthropogenic activities such as shipping, offshore exploration, geophysical seismic surveys and naval sonar operations. Ecosystem managers are concerned about mitigating these threats but have challenges of inadequate information about them. This is because cetaceans spent most of their time in water. These creatures use sounds for communication, echolocation and other social activities. PAM have provided a good alternative for their monitoring. However, PAM usually lead to large volumes of recording which can be difficult to manually analyze. The development of automatic techniques for the automatic detection and classification has led to faster and accurate analysis of cetacean signals. Human bias usually encountered in manual detection and classification process have also being eliminated. The success of any PAM is however dependent on the technique deployed to analyzed the recorded signals. Many automatic detection and classification techniques have been developed by different researchers with mixed outcomes. However, there is no work on a detail review of the existing techniques that can serve as guide to anyone new to this subject area. This review is intended to fill in this gap. We highlighted the procedural steps for carrying out feature extraction, detection and classification of cetacean vocalizations. A review of most of the existing techniques and methods for the design of feature extractor, detector and classifier of cetacean vocalization have been carried out. The differences in the outcomes of the surveyed techniques are influenced by a couple of factors which are discussed. Thus, there is no single technique that can detect and classify the vocalizations over 82 known species of cetacean.

In this review, we look into the different existing techniques so as to give an overview of each technique with respect to their strengths and weaknesses. The techniques are subjective in terms of performance which is determined by the type of signal to be analyzed, species to be detected or classify, feature extraction techniques deployed, location of recording of data, and the expected outputs from the results. We have seen from past efforts that some techniques are robust than another. In deciding on a technique to be used in carrying out an analysis, a thorough evaluation, appraisal, critical and far-reaching knowledge of the technique is needed. This review is intended to give a quick overview and serves as a guide about this subject area for future improvement of existing automated techniques for the detection and classification of cetacean vocalizations.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation (NRF) South Africa under Grant 116036. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

REFERENCES

[1] R. Putland, L. Ranjard, R. Constantine, and C. Radford, "A hidden Markov model approach to indicate Bryde's whale acoustics," *Ecol. Indicators*, vol. 84, pp. 479–487, Jan. 2018.
[2] 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, P. S. Hammond, N. D. Merchant, P. D. O'Hara, J. Purser, A. N. Radford, S. D. Simpson, L. Thomas, and M. A. Wale, "Impacts of anthropogenic noise on marine life: Publication patterns, new discoveries, and future directions in research and management," *Ocean Coastal Manage.*, vol. 115, pp. 17–24, Oct. 2015.
[3] N. D. Merchant, K. L. Brookes, R. C. Faulkner, A. W. J. Bicknell, B. J. Godley, and M. J. Witt, "Underwater noise levels in UK waters," *Sci. Rep.*, vol. 6, no. 1, p. 36942, Dec. 2016.
[4] H. B. Blair, N. D. Merchant, A. S. Friedlaender, D. N. Wiley, and S. E. Parks, "Evidence for ship noise impacts on humpback whale foraging behaviour," *Biol. Lett.*, vol. 12, no. 8, Aug. 2016, Art. no. 20160005.
[5] R. A. Dunlop, "The effect of vessel noise on humpback whale, megaptera novaeangliae, communication behaviour," *Animal Behav.*, vol. 111, pp. 13–21, Jan. 2016.
[6] 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," *Mar. Pollut. Bull.*, vol. 63, nos. 1–4, pp. 18–26, 2011.
[7] N. D. Merchant, E. Pirotta, T. R. Barton, and P. M. Thompson, "Monitoring ship noise to assess the impact of coastal developments on marine mammals," *Mar. Pollut. Bull.*, vol. 78, nos. 1–2, pp. 85–95, Jan. 2014.
[8] 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 Ecol. Evol.*, vol. 25, no. 7, pp. 419–427, Jul. 2010.
[9] W. J. Richardson, C. R. Greene Jr, C. I. Malme, and D. H. Thomson, *Marine Mammals and Noise*. New York, NY, USA: Academic, 2013.
[10] L. S. Weilgart, "A brief review of known effects of noise on marine mammals," *Int. J. Comparative Psychol.*, vol. 20, no. 2, pp. 159–168, 2007.
[11] D. Gillespie, M. Caillat, J. Gordon, and P. White, "Automatic detection and classification of odontocete whistles," *J. Acoust. Soc. Amer.*, vol. 134, no. 3, pp. 2427–2437, Sep. 2013.
[12] P. Peso Parada and A. Cardenal-López, "Using Gaussian mixture models to detect and classify dolphin whistles and pulses," *J. Acoust. Soc. Amer.*, vol. 135, no. 6, pp. 3371–3380, Jun. 2014.
[13] W. M. Zimmer, *Passive Acoustic Monitoring of Cetaceans*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[14] T. A. Jefferson, M. A. Webber, and R. L. Pitman, *Marine Mammals of the World: A Comprehensive Guide to Their Identification*. Amsterdam, The Netherlands: Elsevier, 2011.
[15] J. E. Reynolds, *Biology of Marine Mammals*. Washington, DC, USA: Smithsonian Institution Press, 2007.
[16] M. Hindell and R. Kirkwood, *Marine Mammals: Fisheries, Tourism and Management Issues*. Clayton, NSW, Australia: Csiro Publishing, 2003.
[17] J. S. Smith, L. K. Hill, and R. M. Gonzalez, "Whale watching and preservation of the environment in central Baja California, Mexico," *FOCUS Geography*, vol. 62, p. 1, Sep. 2019.
[18] E. Parsons, S. Baulch, T. Bechshoft, G. Bellazzi, Phil Bouchet, A. M. Cosentino, and C. A. J. Godard-Codding, "Key research questions of global importance for cetacean conservation," *Endangered Species Res.*, vol. 27, no. 2, pp. 113–118, Feb. 2015.
[19] Y. Xian, "Detection and classification of whale acoustic signals," Ph.D. dissertation, Dept. Elect. Comput. Eng., Duke Univ., Durham, NC, USA, 2016.
[20] A. S. Kavanagh, M. Nykänen, W. Hunt, N. Richardson, and M. J. Jessopp, "Seismic surveys reduce cetacean sightings across a large marine ecosystem," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.
[21] R. Filadelfo, J. Mintz, E. Michlovich, A. D'Amico, P. L. Tyack, and D. R. Ketten, "Correlating military sonar use with beaked whale mass strandings: What do the historical data show?" *Aquatic Mammals*, vol. 35, no. 4, pp. 435–444, Dec. 2009.
[22] M. Bittle and A. Duncan, "A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring," in *Proc. Acoust.*, 2013, pp. 1–8.
[23] 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," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 1982–1994, 2009.

[24] D. Mellinger, K. Stafford, S. Moore, R. Dziak, and H. Matsumoto, "An overview of fixed passive acoustic observation methods for cetaceans," *Oceanography*, vol. 20, no. 4, pp. 36–45, Dec. 2007.
[25] J.-J. Jiang, L.-R. Bu, F.-J. Duan, X.-Q. Wang, W. Liu, Z.-B. Sun, and C.-Y. Li, "Whistle detection and classification for whales based on convolutional neural networks," *Appl. Acoust.*, vol. 150, pp. 169–178, Jul. 2019.
[26] R. Gibb, E. Browning, P. Glover-Kapfer, and K. E. Jones, "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring," *Methods Ecol. Evol.*, vol. 10, no. 2, pp. 169–185, Feb. 2019.
[27] D. R. McGaughey, D. Marcotte, M. J. Korenberg, and J. A. Theriault, "Detection and classification of marine mammal clicks," in *Proc. OCEANS MTS/IEEE SEATTLE*, Sep. 2010, pp. 1–10.
[28] J. R. Buck and P. L. Tyack, "A quantitative measure of similarity for tursiops truncatus signature whistles," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 2497–2506, Nov. 1993.
[29] O. Adam, "Segmentation of killer whale vocalizations using the Hilbert-Huang transform," *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, Dec. 2008, Art. no. 245936.
[30] G. Qiao, M. Bilal, S. Liu, Z. Babar, and T. Ma, "Biologically inspired covert underwater acoustic communication—A review," *Phys. Commun.*, vol. 30, pp. 107–114, Oct. 2018.
[31] T. M. Yack, J. Barlow, S. Rankin, and D. Gillespie, "Testing and validation of automated whistle and click detectors using PAMGUARD 1.0," Nat. Ocean. Atmos. Admin., Silver Spring, MD, USA, Tech. Rep. NOAA-TM-NFS-SWFSC-443, 2009.
[32] J.-J. Jiang, L.-R. Bu, X.-Q. Wang, C.-Y. Li, Z.-B. Sun, H. Yan, B. Hua, F.-J. Duan, and J. Yang, "Clicks classification of sperm whale and long-finned pilot whale based on continuous wavelet transform and artificial neural network," *Appl. Acoust.*, vol. 141, pp. 26–34, Dec. 2018.
[33] 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," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.
[34] C. Erbe, "Underwater acoustics: Noise and the effects on marine mammals," in *A Pocket Handbook*, vol. 164. Victoria, BC, Canada: Jasco Applied Sciences, 2011. [Online]. Available: http://www.jasco.com
[35] K. S. Norris and W. E. Evans, "Directionality of echolocation clicks in the rough-tooth porpoise, Steno bredanensis (Lesson)," in *Marine Bio-Acoustics*. New York, NY, USA: Pergamon Press, 1967, pp. 305–316.
[36] 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," *Can. Acoust.*, vol. 33, no. 2, pp. 25–34, 2005.
[37] J. C. Brown and P. Smaragdis, "Hidden Markov and Gaussian mixture models for automatic call classification," *J. Acoust. Soc. Amer.*, vol. 125, no. 6, pp. EL221–EL224, Jun. 2009.
[38] 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.
[39] 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," *J. Acoust. Soc. Amer.*, vol. 144, no. 6, pp. 3181–3190, Dec. 2018.
[40] T. Guilment, F.-X. Socheleau, D. Pastor, and S. Vallez, "Sparse representation-based classification of mysticete calls," *J. Acoust. Soc. Amer.*, vol. 144, no. 3, pp. 1550–1563, Sep. 2018.
[41] Y. Shiu, K. J. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, and H. Klinck, "Deep neural networks for automated detection of marine mammal species," *Sci. Rep.*, vol. 10, no. 1, p. 607, Dec. 2020.
[42] 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," *Ecol. Informat.*, vol. 57, May 2020, Art. no. 101087.
[43] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, "Marine mammal species classification using convolutional neural networks and a novel acoustic representation," 2019, *arXiv:1907.13188*. [Online]. Available: http://arxiv.org/abs/1907.13188
[44] 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," *Can. Acoust.*, vol. 36, no. 1, pp. 41–47, 2008.
[45] 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," *J. Acoust. Soc. Amer.*, vol. 147, no. 2, pp. EL125–EL131, Feb. 2020.
[46] 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 Sens.*, vol. 12, no. 2, p. 326, Jan. 2020.
[47] P. Rickwood and A. Taylor, "Methods for automatically analyzing humpback song units," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1763–1772, Mar. 2008.
[48] 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," *Biol. Rev.*, vol. 88, no. 2, pp. 287–309, May 2013.
[49] W. Rayment, T. Webster, T. Brough, T. Jowett, and S. Dawson, "Seen or heard? A comparison of visual and acoustic autonomous monitoring methods for investigating temporal variation in occurrence of southern right whales," *Mar. Biol.*, vol. 165, no. 1, p. 12, Jan. 2018.
[50] R. Aubauer, M. O. Lammers, and W. W. L. Au, "One-hydrophone method of estimating distance and depth of phonating dolphins in shallow water," *J. Acoust. Soc. Amer.*, vol. 107, no. 5, pp. 2744–2749, May 2000.
[51] O. Adam, "Advantages of the Hilbert Huang transform for marine mammals signals analysis," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2965–2973, Nov. 2006.
[52] O. O. Ogundile and D. J. Versfeld, "Analysis of template-based detection algorithms for inshore Bryde's whale short pulse calls," *IEEE Access*, vol. 8, pp. 14377–14385, 2020.
[53] O. E. NOAA. (2001). *North Pacific Blue Whale Call*. Accessed: Apr. 22, 2020. [Online]. Available: https://cutt.ly/6yTaRud.
[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 Trans. Signal Process.*, vol. 46, no. 5, pp. 1202–1207, May 1998.
[55] M. N. Murty and V. S. Devi, *Pattern Recognition: An Algorithmic Approach*. London, U.K.: Springer, 2011, doi: 10.1007/978-0-85729-495-1.
[56] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Proc. Sci. Inf. Conf.*, Aug. 2014, pp. 372–378.
[57] A. Djebbari and F. Bereksi Reguig, "Short-time Fourier transform analysis of the phonocardiogram signal," in *Proc. 7th IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, vol. 2, Dec. 2000, pp. 844–847.
[58] J. Jun Lee, S. Min Lee, I. Young Kim, H. Ki Min, and S. H. Hong, "Comparison between short time Fourier and wavelet transform for feature extraction of heart sound," in *Proc. IEEE. Region 10 Conf. Multimedia Technol. Asia–Pacific Inf. Infrastruct. (TENCON)*, vol. 2, Sep. 1999, pp. 1547–1550.
[59] R. X. Gao and R. Yan, "Non-stationary signal processing for bearing health monitoring," *Int. J. Manuf. Res.*, vol. 1, no. 1, pp. 18–40, 2006.
[60] Z. Xiao, M. Zhang, L. Chen, and H. Jin, "Detection and segmentation of underwater CW-like signals in spectrum image under strong noise background," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 287–294, Apr. 2019.
[61] M. P. Fargues and R. Bennett, "Comparing wavelet transforms and AR modeling as feature extraction tools for underwater signal classification," in *Proc. Conf. Rec. 29th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Oct./Nov. 1995, pp. 915–919.
[62] 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 Trans. Image Process.*, vol. 25, no. 6, pp. 2739–2751, Jun. 2016.
[63] 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.
[64] 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 Univ.-Erbil Sci. J.*, vol. 3, no. 2, pp. 80–84, Aug. 2019.
[65] 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, 2019.
[66] L. Chun-Lin, "A tutorial of the wavelet transform," NTUEE, Taiwan, Taipei, Tech. Rep., 2010.

[67] G. G. Yen and K.-C. Lin, "Wavelet packet feature extraction for vibration monitoring," *IEEE Trans. Ind. Electron.*, vol. 47, no. 3, pp. 650–667, Jun. 2000.

[68] O. Duzenli, "Classification of underwater signals using wavelet-based decompositions," Naval Postgraduate School, Monterey, CA, USA, Tech. Rep. 19980722 033, 1998.

[69] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[70] A. Dogra, "Performance comparison of different wavelet families based on bone vessel fusion," *Asian J. Pharmaceutics (AJP): Free Full Text Articles Asian J Pharm*, vol. 10, no. 4, 2017.

[71] 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, Dec. 2005.

[72] M. van der Schaar, E. Delory, A. Català, and M. André, "Neural network-based sperm whale click classification," *J. Mar. Biol. Assoc. United Kingdom*, vol. 87, no. 1, pp. 35–38, Feb. 2007.

[73] 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," in *Proc. WEC*, vol. 2, 2005, pp. 188–195.

[74] A. Daviu and J. Alfonso, "Operation of the Hilbert-Huang transform: Basic overview with examples," Escuela Técnica Superior Ingenieros Ind., Dept. Ingeniería Eléctrica, Univ. Politècnica de València, Valencia, Spain, Tech. Rep., 2013.

[75] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[76] O. Adam, "The use of the Hilbert-Huang transform to analyze transient signals emitted by sperm whales," *Appl. Acoust.*, vol. 67, nos. 11–12, pp. 1134–1143, Nov. 2006.

[77] Z. Peng, W. T. Peter, and F. Chu, "An improved Hilbert–Huang transform and its application in vibration signal analysis," *J. Sound Vib.*, vol. 286, nos. 1–2, pp. 187–205, 2005.

[78] 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 *Proc. 2nd Int. Congr. Image Signal Process.*, Oct. 2009, pp. 1–5.

[79] L. Rabiner and R. Schafer, "Digital speech processing," *Froehlich/Kent Encyclopedia Telecommun.*, vol. 6, pp. 237–258, Dec. 2011.

[80] L. Wang, Z. Chen, and F. Yin, "A novel hierarchical decomposition vector quantization method for high-order LPC parameters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 212–221, Jan. 2015.

[81] N. Desai, K. Dhameliya, and V. Desai, "Feature extraction and classification techniques for speech recognition: A review," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 12, pp. 367–371, 2013.

[82] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2B, pp. 637–655, 1971.

[83] H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in speech recognition system," in *Proc. 6th Int. Conf.-Cloud Syst. Big Data Eng. (Confluence)*, Jan. 2016, pp. 498–502.

[84] S. Mazhar, T. Ura, and R. Bahl, "Effect of temporal evolution of songs on cepstrum-based voice signature in humpback whales," in *Proc. OCEANS-MTS/IEEE Kobe Techno-Ocean*, Apr. 2008, pp. 1–8.

[85] S. K. A and C. E, "Keyword spotting system for tamil isolated words using multidimensional MFCC and DTW algorithm," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2015, pp. 0550–0554.

[86] S. Gupta, J. Jaafar, W. W. Ahmad, and A. Bansal, "Feature extraction using MFCC," *Signal Image Process., Int. J.*, vol. 4, no. 4, pp. 101–108, 2013.

[87] L. Shi, I. Ahmad, Y. He, and K. Chang, "Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments," *J. Commun. Netw.*, vol. 20, no. 5, pp. 509–518, Oct. 2018.

[88] R. K. Km and S. Ganesan, "Comparison of multidimensional MFCC feature vectors for objective assessment of stuttered disfluencies," *Int. J. Adv. Netw. Appl.*, vol. 2, no. 05, pp. 854–860, 2011.

[89] M. R. Hasan, M. Jamil, and M. Rahman, "Speaker identification using MEL frequency cepstral coefficients," *Variations*, vol. 1, no. 4, pp. 565–568, 2004.

[90] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," 2010, *arXiv:1003.4083*. [Online]. Available: http://arxiv.org/abs/1003.4083

[91] 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 *Proc. IEEE Conf. Inf. Commun. Technol.*, Apr. 2013, pp. 1263–1268.

[92] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.

[93] F. Pace, P. White, and O. Adam, "Hidden Markov Modeling for humpback whale (Megaptera Novaeanglie) call classification," in *Proc. Meetings Acoust. (ECUA)*, 2012, vol. 17, no. 1, Art. no. 070046.

[94] V. Kandia and Y. Stylianou, "Detection of sperm whale clicks based on the Teager–Kaiser energy operator," *Appl. Acoust.*, vol. 67, nos. 11–12, pp. 1144–1163, 2006.

[95] Y. Xian, A. Thompson, Q. Qiu, L. Nolte, D. Nowacek, J. Lu, and R. Calderbank, "Classification of whale vocalizations using the weyl transform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 773–777.

[96] 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," *Appl. Sci.*, vol. 9, no. 5, p. 1020, Mar. 2019.

[97] R. Sridharan. (2014). *Gaussian Mixture Models and the EM Algorithm*. [Online]. Available: http://people.csail.mit.edu/rameshvs/content/gmm-em.pdf

[98] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *Int. Comput. Sci. Inst.*, vol. 4, no. 510, p. 126, 1998.

[99] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[100] X. C. Halkias, *Detection Tracking Dolphin Vocalizations*. New York, NY, USA: Columbia Univ., 2009.

[101] 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," *J. Acoust. Soc. Amer.*, vol. 121, no. 3, pp. 1737–1748, Mar. 2007.

[102] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[103] Z. Ghahramani, "An introduction to hidden Markov models and Bayesian networks," in *Hidden Markov Models: Applications in Computer Vision*. Singapore: World Scientific, 2001, pp. 9–41.

[104] D. Jurafsky, *Speech & Language Processing*. New Delhi, India: Pearson Education, 2000.

[105] G. Cybenko and V. Crespi, "Learning hidden Markov models using nonnegative matrix factorization," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3963–3970, Jun. 2011.

[106] 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," *J. Acoust. Soc. Amer.*, vol. 138, no. 6, pp. 3941–3950, Dec. 2015.

[107] 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," *J. Acoust. Soc. Amer.*, vol. 138, no. 1, pp. EL105–EL109, Jul. 2015.

[108] B. A. Weisburn, S. G. Mitchell, C. W. Clark, and T. W. Parks, "Isolating biological acoustic transient signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, Apr. 1993, pp. 269–272.

[109] D. K. Mellinger and C. W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *J. Acoust. Soc. Amer.*, vol. 107, no. 6, pp. 3518–3529, Jun. 2000.

[110] S. Datta and C. Sturtivant, "Dolphin whistle classification for determining group identities," *Signal Process.*, vol. 82, no. 2, pp. 251–258, Feb. 2002.

[111] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996.

[112] V. Cheung and K. Cannons, "An introduction to neural networks," Univ. Manitoba, Winnipeg, MB, Canada, Tech. Rep., 2002.

[113] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[114] S. O. Murray, E. Mercado, and H. L. Roitblat, "The neural network classification of false killer whale (Pseudorca crassidens) vocalizations," *J. Acoust. Soc. Amer.*, vol. 104, no. 6, pp. 3626–3633, Dec. 1998.
[115] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.
[116] V. B. Deecke and V. M. Janik, "Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls," *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 645–653, Jan. 2006.
[117] X. Yin, Y. Hou, J. Yin, and C. Li, "A novel SVM parameter tuning method based on advanced whale optimization algorithm," in *Proc. J. Phys., Conf.*, 2019, vol. 1237, no. 2, Art. no. 022140.
[118] V. Jakkula, "Tutorial on support vector machine (SVM)," School EECS, Washington State Univ., Seattle, WA, USA, Tech. Rep., 2006, vol. 37.
[119] C. Junli and J. Licheng, "Classification mechanism of support vector machines," in *Proc. 5th Int. Conf. Signal Process. 16th World Comput. Congr. (WCC-ICSP)*, vol. 3, Aug. 2000, pp. 1556–1559.
[120] V. Vapnik, "The nature of statistical learning theory springer New York Google scholar," Tech. Rep., 1995.
[121] A. K. Ibrahim, H. Zhuang, N. Erdol, and A. M. Ali, "A new approach for North Atlantic right whale upcall detection," in *Proc. Int. Symp. Comput., Consum. Control (ISC)*, Jul. 2016, pp. 260–263.
[122] 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," *Can. Acoust.*, vol. 36, no. 1, pp. 34–40, 2008.
[123] 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 *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 865–870.
[124] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, A. V. Oppenheim and R. W. Schafer, Eds. Upper Saddle River, NJ, USA: Prentice-Hall, 1975.
[125] R. A. Altes, "Detection, estimation, and classification with spectrograms," *J. Acoust. Soc. Amer.*, vol. 67, no. 4, pp. 1232–1246, Apr. 1980.
[126] D. K. Mellinger, "A comparison of methods for detecting right whale calls," *Can. Acoust.*, vol. 32, no. 2, pp. 55–65, 2004.
[127] D. K. Mellinger and C. W. Clark, "Methods for automatic detection of mysticete sounds," *Mar. Freshwater Behav. Physiol.*, vol. 29, nos. 1–4, pp. 163–181, Jan. 1997.
[128] Q. Xue, Y. H. Hu, and W. J. Tompkins, "Neural-network-based adaptive matched filtering for QRS detection," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 4, pp. 317–329, Apr. 1992.
[129] J. C. Bancroft, "Introduction to matched filters," CREWES Res., Dept. Geosci., Univ. Calgary, Calgary, AB, Canada, Tech. Rep., 2002, vol. 14.
[130] G. Turin, "An introduction to matched filters," *IRE Trans. Inf. Theory*, vol. 6, no. 3, pp. 311–329, Jun. 1960.
[131] P. Courmontagne, "The stochastic matched filter and its applications to detection and de-noising," in *Stochastic Control*. Rijeka, Croatia: IntechOpen, 2010.
[132] 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 *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 2319–2323.
[133] K. M. Stafford, C. G. Fox, and D. S. Clark, "Long-range acoustic detection and localization of blue whale calls in the northeast pacific ocean," *J. Acoust. Soc. Amer.*, vol. 104, no. 6, pp. 3616–3625, Dec. 1998.
[134] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.
[135] P. Senin, "Dynamic time warping algorithm review," Dept. Inf. Comput. Sci., Univ. Hawaii Manoa Honolulu, HI, USA, 2008, vol. 855, nos. 1–23, p. 40.
[136] G. Cleuziou and J. G. Moreno, "Kernel methods for point symmetry-based clustering," *Pattern Recognit.*, vol. 48, no. 9, pp. 2812–2830, Sep. 2015.
[137] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2001, pp. 1–11.
[138] J. C. Brown, A. Hodgins-Davis, and P. J. O. Miller, "Classification of vocalizations of killer whales using dynamic time warping," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. EL34–EL40, Mar. 2006.
[139] J. C. Brown and P. J. O. Miller, "Automatic classification of killer whale vocalizations using dynamic time warping," *J. Acoust. Soc. Amer.*, vol. 122, no. 2, pp. 1201–1207, Aug. 2007.
[140] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *J. Fluid Mech.*, vol. 656, pp. 5–28, Aug. 2010.
[141] 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," *J. Acoust. Soc. Amer.*, vol. 129, no. 1, pp. 467–475, Jan. 2011.
[142] W. Luo, W. Yang, and Y. Zhang, "Convolutional neural network for detecting odontocete echolocation clicks," *J. Acoust. Soc. Amer.*, vol. 145, no. 1, pp. EL7–EL12, Jan. 2019.
[143] S. Mazhar, T. Ura, and R. Bahl, "Vocalization based individual classification of humpback whales using support vector machine," in *Proc. OCEANS*, Sep. 2007, pp. 1–9.
[144] 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 *Proc. Ocean Eng. Symp.*, 2003.
[145] A. Cuevas, A. Veragua, S. Espanol-Jimenez, G. Chiang, and F. Tobar, "Unsupervised blue whale call detection using multiple time-frequency features," in *Proc. CHILEAN Conf. Electr., Electron. Eng., Inf. Commun. Technol. (CHILECON)*, Oct. 2017, pp. 1–6.
[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, 2019.

AYINDE M. USMAN (Member, IEEE) received the B.Eng. degree in electrical engineering and the M.Eng. degree in electrical and electronics engineering from the University of Ilorin, Ilorin, Nigeria, in 2007 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, Stellenbosch University, Stellenbosch, South Africa. He is also a Lecturer with the Department of Electrical and Electronics Engineering, University of Ilorin. His research interests include machine learning, digital communication, signal processing, and the Internet of Things.

OLAYINKA O. OGUNDILE (Member, IEEE) received the B.Eng. degree in electrical engineering from the University of Ilorin, Ilorin, Nigeria, in 2007, the M.Sc. degree in communication engineering from The University of Manchester, U.K., in 2010, and the Ph.D. degree from the University of the Witwatersrand, Johannesburg, South Africa, in 2016. He was a Postdoctoral Research Fellow with the School of Electrical and Information Engineering, University of the Witwatersrand, in 2016. From 2016 to 2017, he was with the Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa. He is currently a Lecturer with the Department of Computer Science, Tai Solarin University of Education, Ijebu Ode, Nigeria. His research interests include digital communication, digital transmission techniques, channel estimation, forward error correction, wireless sensor networks, visible light communication, and the Internet of Things.

DANIEL J. J. VERSFELD (Member, IEEE) received the B.Eng. and M.Eng. degrees in electronic engineering from North-West University, South Africa, in 1999 and 2001, respectively, and the Ph.D. degree from the University of Johannesburg, South Africa, in 2011. He is currently an Associate Professor with the Department of Electrical and Electronic Engineering, Stellenbosch University, Stellenbosch, South Africa. His research interests include algebraic coding for digital communications and signal processing applied to communications.