

A hidden Markov model with selective time domain feature extraction to detect inshore Bryde's whale short pulse calls

O.O. Ogundile^{a,b,*}, A.M. Usman^a, O.P. Babalola^a, D.J.J. Versfeld^a

^a Department of Electrical and Electronic Engineering, Stellenbosch University, Stellenbosch, South Africa

^b Department of Computer Science, Tai Solarin University of Education, Ogun State, Nigeria

ARTICLE INFO

Keywords:

Bryde's whale
HMM
LPC
MAP
MFCC
Sound detection

ABSTRACT

An Increase in the study of cetaceans' sounds has motivated the development of different automated sound detection and classification techniques. Passive acoustic monitoring (PAM) is extensively used to study these cetaceans' sounds over a period to understand their daily activities within their ecosystem. Using PAM, the gathered sound datasets are usually large and impractical to manually analyse and detect. Thus, hidden Markov models (HMM) is one of the popular tools used to automatically detect and classify these cetaceans' sounds. Nonetheless, HMM rely heavily on the employed feature extraction method such as Mel-scale frequency cepstral coefficients (MFCC) and linear predictive coding (LPC). In most cases, the more reliable the extracted feature vector from the known sound label, the higher the sensitivity of the HMM. Although these aforementioned feature extraction methods are widely used, their design is based on filters and requires windowing, fast Fourier transforms (FFT), and logarithm operations. Consequently, this increases the computational time complexity of the HMM. Here, we describe a selective time domain feature extraction method that can be easily adapted with the HMM. This proposed feature extraction method uses a combination of some simple but robust parameters such as the mean, relative amplitude and relative power/energy (MAP), which are selected based on empirical observations of the call to be detected. The performance of this proposed MAP-HMM was verified using the acoustic dataset of continuous recordings of an inshore Bryde's whale (*Balaenoptera*) short pulse calls collected in a single site in False Bay, South-West of South Africa. Aside from exhibiting a low computational complexity, the proposed MAP-HMM offers superior sensitivity and false discovery rate performances in comparison to the LPC-HMM and MFCC-HMM.

1. Introduction

Marine ecologists have shown over the years that cetaceans use sound for communications, positioning, hunting and avoiding predators within their ecosystems. These sounds can be in the form of whistles, clicks and pulse calls (Brown et al., 2007). Passive acoustic monitoring (PAM) offers good options to study cetaceans' vocal repertoire, movements and response to anthropogenic activities. In PAM, the sound datasets can be gathered over a long period, usually in months and years. These datasets can be collected in remote areas, even in adverse weather conditions (Putland et al., 2018; Constantine et al., 2018).

Usually, the datasets accumulated using PAM are large and can be difficult to analyse, detect and classify by human experts. As a result, there is an increase in the development of different automated sound detection and classification techniques for cetaceans. Ordinarily, these techniques vary depending on the cetaceans species, sound duration,

range of frequency and sound type. Machine learning (ML) tools such as the hidden Markov model (HMM) (Jurafsky and James, 2018) are popularly used for detection and classification of cetaceans' sounds including the obscure species. The HMM is a flexible technique that allows sound to be easily detected and classified from a series of specified observations (Yao et al., 2009; Putland et al., 2018). Yet, the detection and classification capabilities of the HMM depends on the extracted feature vector from a known sound label (the known sound label is often identified by a human expert). In most cases, the more reliable the extracted feature vector, the higher the detection accuracy of the HMM. The Mel-scale frequency cepstral coefficients (MFCC) is a feature extraction method widely used in signal processing and more often adapted with the HMM (Florian et al., 2010; Martin and Jeroen, 2003; Shi et al., 2018; Khalid et al., 2014). MFCC is widely used because it exhibit a low complexity in comparison to the other methods such as the linear predictive coding (LPC) (Makhoul, 1975; Rabiner and Shafer,

* Corresponding author at: Department of Electrical and Electronic Engineering, Stellenbosch University, Stellenbosch, South Africa
E-mail address: ogundileoo@gmail.com (O.O. Ogundile).

1978; Min and Tewfik, 2010), that can be combined with the HMM. Nevertheless, the design of these feature extraction methods are based on filters and require some other complex steps such as windowing, fast Fourier transform (FFT), discrete cosine transform (DCT) and logarithm operations. Consequently, it tends to increase the computational time complexity of the HMM.

This article introduces a selective time domain feature extraction method that can be easily adapted with the HMM. This proposed feature extraction method uses a combination of some simple but robust characteristics of the known sound signal; such as, the mean, relative amplitude and relative power/energy (MAP). The operation of the proposed MAP feature extraction method is not complex as it does not require any FFT, DCT and logarithm operations. In addition, the dimension of the extracted feature vector using the MAP method is small in comparison to the MFCC and LPC methods as shown in Section 4.3. Subsequently, it is less complex to adapt with the HMM because the larger the feature vector dimension, the more complex is the HMM processes. The performance of this proposed MAP-HMM was verified on five different 2 h recordings of an inshore Bryde's whale (*Balaenoptera*) short pulse calls collected in a single site in False bay, South-West of South Africa, from December 2018 to July 2019.

Bryde's whales discovered in 1913 off South Africa are members of the cetacean family (Olsen, 1913). Bryde's whales are presently classified as a single species referred to as *Balaenoptera edeni* (*B. edeni*). Consequently, two subspecies have been widely suggested: *Balaenoptera edeni edeni* (*B. e. edeni* - inshore form of *B. edeni*) and *Balaenoptera edeni brydei* (*B. e. brydei* - large offshore form of *B. edeni*) (Constantine et al., 2018; Committee on Taxonomy, 2017; Rosel and Wilcox, 2014; Smith et al., 2013). Many literature have described the population and geographical distribution (Constantine et al., 2018; Smith et al., 2013; Kerosky et al., 2012), nomenclature (Committee on Taxonomy, 2017), genetic and phylogenetic characteristics (Rosel and Wilcox, 2014), and vocal repertoire (Rice et al., 2014; Širović et al., 2014; Viloria-Gomora et al., 2015). Nonetheless, only a few have proposed an automated detection technique for the vocal repertoire of Bryde's whales. Bryde's whales cannot be easily distinguish from other species (such as the Sei whale) based on physical appearance (Viloria-Gomora et al., 2015; Putland et al., 2018). Therefore, identifying the Bryde's whales by sound is motivated in this study, since they produce characteristic calls. Hence, the development of automated detection sound models with high sensitivity and a low false discovery rate for Bryde's whales is paramount.

To the best of the authors knowledge, only Putland et al. (2018) developed an automated detection technique to indicate Bryde's whale acoustic. In Putland et al. (2018), a HMM approach is proposed to automatically detect Bryde's whales call using the MatlabHTK toolkit, where they used the default parametrization for the feature extraction technique (that is, the MFCC feature extraction technique is used). However, both the MFCC and LPC feature extraction techniques are commonly used with the MatlabHTK toolkit. Whereas, as mentioned, this article proposes a low complexity and performance efficient feature extraction technique (MAP) that can be easily combined with the HMM (MAP-HMM) to automatically detect Bryde's whales short pulse calls. The result of this proposed work is innovative and produces good sensitivity while maintaining a low false discovery rate in comparison with the conventional techniques such as the MFCC-MAP and LPC-HMM. Moreover, this proposed MAP-HMM technique can be extended to other cetaceans with closely defined characteristic call.

The remainder of this article is organised as follows. In Section 2, the recording location, recording set-up and data preprocessing are briefly described. Section 3 explains the HMM and its adaptation with the k-means clustering technique (MacQueen, 1967; Forgy, 1965) and the Gaussian mixture model (GMM) (Reynolds, 2009; Duda et al., 2001). The proposed feature extraction technique is explained in Section 4. Besides, the section briefly explains the MFCC and LPC extraction techniques used for comparison. Section 5 presents the simulation

test results and discussion. The article is concluded with discernible remarks in Section 6.

2. Data recording and processing

Bryde's whales has been previously observed off South Africa during marine mammal surveys (Penry et al., 2018, 2016; Olsen, 1913; Best, 1977). In fact, it was historically first sighted off South Africa (Olsen, 1913). PAM was carried out from December 2018 to August 2019 in an area of approximately 13 km², close to Gordon's bay harbour, False bay (34°08'47.3"S 18°48'10.4"E), South Africa. During this period, a recurring short pulse call of the inshore Bryde's whales were collected using dipping hydrophones (Aquarian Audio H2A-XLR Hydrophone, sensitivity: -180 dB re: 1V/μPa, frequency range: 10Hz-100kHz) that was connected to a Zoom H1N recorder, operating at 96kHz at 24 bit resolution. The deployment was to dip the hydrophone from an inboard engine sail boat which is 8 m long, assuming varying conditions. That is, sailing (2-4kts/h), dropping the sails (less than 1kt/h), heaving to (less than 2kts/h). The recording followed standard protocols as documented by the Department of Environmental Affairs, South Africa. The recordings were only carried out when the presence of the inshore Bryde's whale is visually confirmed with the physical features that corresponds to the descriptions in Penry et al. (2018). In contrast, to perform a comparative test, recordings were also conducted when the Bryde's whale was not present in the locality, and the short pulse call was not observed in the recordings. Furthermore, to preserve the relative amplitude and phase of these recordings, it was saved as raw samples (.wav format). The raw samples were filtered with a 3rd order Butterworth bandpass filter using MATLAB R2019a. This eliminates unwanted signal frequencies (background noise and DC component) below 10Hz and beyond 8000Hz. Fig. 1 depicts the waveform of the recorded inshore Bryde's whale short pulse call off South Africa as viewed from a Sonic Visualiser (Ogundile and Versfeld, 2020). For more details on the call characteristics of this Bryde's whale short pulse call



Fig. 1. Spectrogram and time-series representation of the Bryde's whale short pulse calls.

and the PAM recordings, readers are referred to [Ogundile and Versfeld \(2020\)](#).

3. Hidden Markov model

The HMM is a ML tool widely used for speech analysis and recognition ([Rabiner, 1989](#); [Rabiner and Juang, 1986](#)). It can be viewed as a stochastic state machine, where each change in the hidden states end with an emission of a symbol. The HMM tool is flexible, and can easily model and classify the spectral and temporal features of a set of sound signals. For a given set of feature vector extracted from the sound signal, the HMM calculates three main parameters (HMM parameters): the start probability π , the transition matrix ($\mathcal{T}$) and the parameters associated with the emission distribution (ϕ). HMM represents a continuous signal as a sequence of states across time as specified by the $\mathcal{T}$ ([Brown and Smaragdis, 2009](#)). This $\mathcal{T}$ represent the signal with probability values which optimally decides the switching between the states. The $\mathcal{T}$ is developed by calculating the maximum-likelihood (M-L) estimate of the parameters of a HMM from the set of feature vector ([Bilmes, 1998](#)). This process can be performed using the Baum-Welch (BM) algorithm ([Baum et al., 1970](#); [Churbanov and Winters-Hilt, 2008](#)). This M-L estimation is modelled with some Gaussian emission distribution parameters (ϕ) such as the mean, covariance and mixture weight which assume random or flat values at the start of the M-L estimation process. However, HMMs are sensitive to random values of these Gaussian parameters, as its hamper the performance of the model. Therefore, k-means clustering ([MacQueen, 1967](#); [Forgy, 1965](#)) with GMM ([Reynolds, 2009](#); [Duda et al., 2001](#); [Bilmes, 1998](#)) are used sequentially to initialise the Gaussian parameters. The HMM detection system can be viewed as shown in [Fig. 2](#). At convergence, the trained HMM parameters, π , $\mathcal{T}$, and ϕ can be matched with the sound to be recognised using the Viterbi algorithm (V-alg) ([Viterbi, 1967](#); [Churbanov and Winters-Hilt, 2008](#)). As shown in [Fig. 2](#), the V-alg takes the trained HMMs parameters (π and $\mathcal{T}$) and the observable feature vector (the feature vector must have been refined using the HMMs emission distribution ϕ parameter) of the sound to be recognised and it returns all the most likely matching hidden states of the models. Implementing HMM with BM and V-alg have been used in several applications. Nonetheless, the extracted feature vector plays a key role in the performance of any HMM. A reliable feature vector showcase the detection ability of the HMM. The more reliable the extracted feature vector, the better the performance of the HMM. As such, this article focuses on the feature vector extraction blocks as we have highlighted in [Fig. 2](#).

4. Feature extraction techniques

This section briefly explains the LPC and MFCC feature extraction technique used in this article for comparison. Of interest, the proposed MAP feature extraction techniques is explained in details in this section.

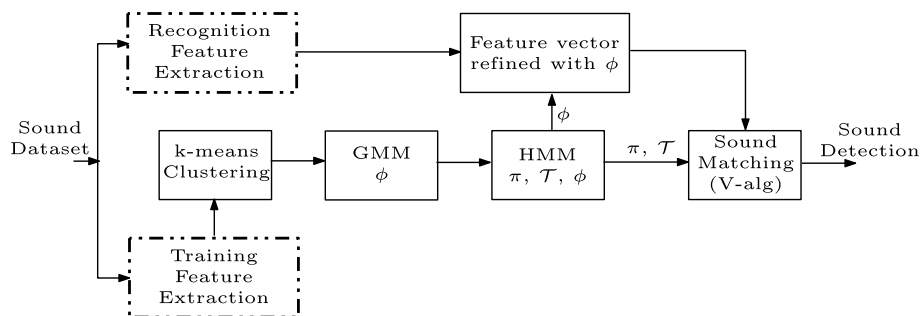


Fig. 2. HMM based detection system.

4.1. LPC

The LPC, also called inverse filtering is a signal processing technique used widely for linear prediction. It has been used for speech coding, synthesis and recognition, and in analysing short length of cetacean vocalisations ([Pace, 2008](#); [Min and Tewfik, 2010](#)). The concept of LPC is to use a linear combination of the previous regenerated sound signal (η^{th}) to predict a value for the current sound signal ($h(\eta)$) as ([Min and Tewfik, 2010](#); [Makhoul, 1975](#); [Rabiner and Shafer, 1978](#)):

$$\hat{h}(\eta) = \sum_{v, \eta=1}^{\mu} \lambda_v h(\eta - \gamma), \quad (1)$$

where $\hat{h}(\eta)$ is the approximated value of $h(\eta)$, $h(\eta - \gamma)$ is the previous η^{th} signal, λ_v is the LPC filter coefficients, and $\mu = v^{th} = \eta^{th}$ is the number of filter coefficients. These coefficients λ_v can be computed by reducing the sum of the squared differences between the linearly estimated signals and the original signals as defined as ([Min and Tewfik, 2010](#); [Makhoul, 1975](#); [Rabiner and Shafer, 1978](#)):

$$\varepsilon(\eta) = h(\eta) - \sum_{v, \eta=1}^{\mu} \lambda_v h(\eta - \gamma), \quad (2)$$

where $\varepsilon(\eta)$ is the error between $h(\eta)$ and $\hat{h}(\eta)$. The coefficients λ_v can be estimated from (2) employing a number of methods such as the auto-correlation method. Ordinarily, the value of $\mu = v^{th} = \eta^{th}$ ranges between 10 and 14. Accordingly, in this article, a value of $\mu = v^{th} = \eta^{th} = 12$ is assumed. This implies that the LPC-HMM used for comparison is a 12-dimensional model. Refer to [Makhoul \(1975\)](#) and [Rabiner and Shafer \(1978\)](#) for detail discussion on the LPC technique. Therefore, given a sample signal $\mathcal{F}$, with τ number of windows, where each window has approximately t number of sampling points, the LPC feature vector can be computed as:

$$\sigma_{LPC} = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} & \lambda_{1,3} & \dots & \lambda_{1,\mu} \\ \lambda_{2,1} & \lambda_{2,2} & \lambda_{2,3} & \dots & \lambda_{2,\mu} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{\tau,1} & \lambda_{\tau,2} & \lambda_{\tau,3} & \dots & \lambda_{\tau,\mu} \end{bmatrix}. \quad (3)$$

4.2. MFCC

The MFCC is a signal processing technique widely used for feature extraction in speech recognition, gesture recognition, speaker identification, cetacean vocalisation detection and analysis ([Gupta et al., 2013](#); [Majeed et al., 2015](#)). The MFCC technique is used to extract distinctive features from a speech signal. It makes use of two kinds of filters in the process, namely, the linear spaced filters and logarithm spaced filters ([Majeed et al., 2015](#)). The observable features are extracted by converting the signal from the time domain into the Mel frequency scale (frequency domain). The MFCC feature extraction process can be summarised into seven sequential steps; (1) Pre-emphasis (2) Framing (3) Windowing (4) FFT (5) Mel-scale filter bank (6) Logarithm

operation, and (7) DCT (Gupta et al., 2013; Majeed et al., 2015). Summarily, given a sound sample, the MFCC calculates the cepstral coefficient as defined as (Zheng et al., 2001):

$$\alpha_k = \sum_{i=1}^{\beta} \mathcal{X}_i \cos \left[k(i - 0.5) \frac{\pi}{\beta} \right], \quad k = 1, 2, \dots, \beta, \quad (4)$$

where $\mathcal{X}_i$ is the logarithmic energy of the i^{th} Mel-spectrum band and β is the total number of cepstral coefficients. The value of β usually ranges between 10 and 14, as in the case of the LPC technique. Therefore, for a fair comparison, $\beta=12$ is used in this article, which implies that the MFCC-HMM is a 12-dimensional model. Also, given a sample signal $\mathcal{F}$, with τ number of windows, where each window has approximately t number of sampling points, the MFCC feature vector can be computed as:

$$\sigma_{MFCC} = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \dots & \alpha_{1,\beta} \\ \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \dots & \alpha_{2,\beta} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{\tau,1} & \alpha_{\tau,2} & \alpha_{\tau,3} & \dots & \alpha_{\tau,\beta} \end{bmatrix}. \quad (5)$$

Note that despite the number of steps used to compute σ_{MFCC} , the MFCC technique exhibits a low computational time complexity in comparison to most of the feature extraction techniques used with HMM. Hence, it is widely used for real time speech analysis. See Zheng et al. (2001) and Majeed et al. (2015) for detail discussion on the MFCC technique.

4.3. Proposed MAP

The proposed feature extraction method is based on an empirical observation of the characteristic waveforms of the pulse call to be detected. In this case, the sound to be detected is a short pulse call of an inshore Bryde's whale that has distinctive features. For example, the relative amplitude of the short pulse calls have a theoretically observable lower and upper limits. Likewise, other characteristics of the pulse calls such as the call duration, mean, relative power, frequency range, among others, are distinctive and have empirical limits. Thus, the proposed MAP feature extraction method can be easily formed by carefully selecting some of these important characteristics of the pulse call in the time domain. Given a sample signal $\mathcal{F}$, the MAP feature vector can be formed by combining the following characteristic parameters in no specific order.

4.3.1. Mean, $\mathcal{M}$

The $\mathcal{M}$ is an important parameter that forms part of the MAP feature vector. The mean is typically used to average out or minimise the background noise or DC component not filtered (during filtering), that might potentially change the shape of the waveform. It is selected as a component of the MAP feature vector because for a given training sample, the mean of the inshore Bryde's whale pulse call do not exceed an empirically determined value. Accordingly, consider a sample signal $\mathcal{F}$, with τ number of windows, where each window has approximately t number of sampling points, the $\mathcal{M}$ is simply computed in the time domain for each window as:

$$\mathcal{M}_i = \frac{1}{t} \sum_{u=1}^t (\mathcal{F}_i)_u, \quad i = 1, 2, \dots, \tau. \quad (6)$$

where $\mathcal{F}_i$ is a window of the sample signal $\mathcal{F}$. That is, for the entire sample signal $\mathcal{F}$, the $\mathcal{M}$ can be represented as:

$$\mathcal{M} = \begin{bmatrix} \mathcal{M}_1 \\ \mathcal{M}_2 \\ \vdots \\ \mathcal{M}_\tau \end{bmatrix} = \begin{bmatrix} \frac{1}{t} \sum_{u=1}^t (\mathcal{F}_1)_u \\ \frac{1}{t} \sum_{u=1}^t (\mathcal{F}_2)_u \\ \vdots \\ \frac{1}{t} \sum_{u=1}^t (\mathcal{F}_\tau)_u \end{bmatrix} \quad (7)$$

4.3.2. Relative amplitude vector, $\mathcal{A}$

The relative amplitude vector is also an important parameter used in forming the MAP feature vector. The $\mathcal{A}$ is selected as a feature in the MAP feature vector because it determines the relative amplitude band of the short pulse call. Using similar notations as in Section 4.3.1, the $\mathcal{A}$ is expressed as:

$$\mathcal{A}_i = [\mathcal{A}_i^{\text{dif}}, \mathcal{A}_i^{\text{max}}, \mathcal{A}_i^{\text{min}}], \quad (8)$$

where $\mathcal{A}_i^{\text{max}}$, $\mathcal{A}_i^{\text{min}}$ and $\mathcal{A}_i^{\text{dif}}$ are defined as:

$$\mathcal{A}_i^{\text{max}} = \max(\mathcal{F}_i), \quad \mathcal{A}_i^{\text{min}} = \min(\mathcal{F}_i) \quad \text{and} \quad \mathcal{A}_i^{\text{dif}} = \mathcal{A}_i^{\text{max}} - \mathcal{A}_i^{\text{min}}, \quad (9)$$

are the maximum relative amplitude, minimum relative amplitude and relative amplitude difference respectively for each window of the sample signal $\mathcal{F}$. Note that $\mathcal{A}_i^{\text{dif}}$, $\mathcal{A}_i^{\text{max}}$, and $\mathcal{A}_i^{\text{min}}$ in (8) can be arranged in no specific order. For a τ number of windows of $\mathcal{F}$, the $\mathcal{A}$ can be represented as:

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_1^{\text{dif}} & \mathcal{A}_1^{\text{max}} & \mathcal{A}_1^{\text{min}} \\ \mathcal{A}_2^{\text{dif}} & \mathcal{A}_2^{\text{max}} & \mathcal{A}_2^{\text{min}} \\ \vdots & \vdots & \vdots \\ \mathcal{A}_\tau^{\text{dif}} & \mathcal{A}_\tau^{\text{max}} & \mathcal{A}_\tau^{\text{min}} \end{bmatrix}. \quad (10)$$

The $\mathcal{A}_i^{\text{min}}$ and $\mathcal{A}_i^{\text{max}}$ are selected in $\mathcal{A}$ because the Bryde's whale pulse call falls between an empirically defined relative amplitude band. Intuitively, the $\mathcal{A}_i^{\text{dif}}$ will assume a defined band as computed in (9).

4.3.3. Relative power, $\mathcal{P}$

It was observed that the inshore Bryde's whale short pulse calls have relative powers $\mathcal{P}$ that fall within an empirically determined range. This means that most training sample with a $\mathcal{P}$ below this range could either be the whale pulse call or noise with equal probability, depending on the other characteristics of the waveform. Hence, the $\mathcal{P}$ gives an insight to the type of call when combined with other parameters in the MAP feature vector. The $\mathcal{P}$ can be calculated as:

$$\mathcal{P}_i = \sum_{u=1}^t (\mathcal{F}_i)_u^2, \quad (11)$$

where all notations are defined as above. Similarly, for τ number of windows of $\mathcal{F}$, the $\mathcal{P}$ can be represented as:

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_1 \\ \mathcal{P}_2 \\ \vdots \\ \mathcal{P}_\tau \end{bmatrix} = \begin{bmatrix} \sum_{u=1}^t (\mathcal{F}_1)_u^2 \\ \sum_{u=1}^t (\mathcal{F}_2)_u^2 \\ \vdots \\ \sum_{u=1}^t (\mathcal{F}_\tau)_u^2 \end{bmatrix}. \quad (12)$$

Therefore, the proposed MAP technique extracts the observable features from the sound signal by combining these aforementioned parameters as:

$$\sigma_{MAP} = [\mathcal{M}, \mathcal{A}, \mathcal{P}]. \quad (13)$$

Note, in (13), σ_{MAP} is a sequence of feature vectors, which are 5-dimensional. This means that the MAP-HMM is a 5-dimensional model, as compared to the 12-dimensional LPC-HMM and MFCC-HMM. With respect to the HMM computations, the larger the dimension of the feature vector, the higher the computational time complexity of the

model. Whereas, after a particular number of coefficients, further increase in the dimension of the feature vector, do not necessarily improve the detection ability of the HMM. Thus, aside offering less computational time complexity, the MAP-HMM simplifies the operation of the HMM because it has a smaller feature vector dimension in comparison to the LPC-HMM and MFCC-HMM.

5. Test results and discussion

5.1. Sample training and testing

The recordings accumulated during the specified PAM period were analysed aurally and visually using the spectrogram in Sonic Visualiser.¹ The recordings are 2 h long, where each recording is divided into two portions: one portion (usually small) is used for training the model and the other portion is used for testing. Annotations were carried out on the training sample (the training samples were extracted from the small portion) and categorised into two classes; (1) Inshore Bryde's whale pulse call, and (2) Noise (note that any recording that cannot be aurally and visually recognised as the whale pulse call is regarded as noise). Likewise, the large portions of the five different 2 h recordings were all annotated for evaluation purposes. Therefore, a total of two HMMs were formed, one for the pulse call ($\pi_1, \mathcal{T}_1, \phi_1$) and the other for noise ($\pi_2, \mathcal{T}_2, \phi_2$). We varied the number of training sample k ($k = 18, 12$ and 6) used to train the HMMs in each case, where the training samples have varying time lengths depending on the call duration. Moreover, each training sample is divided into t number of sampling points, where performance analysis is carried out for different values of t . The HMMs were set to four states ergodic type with two mixtures each. The dimension of the models is determined by the extracted feature vector. In this case, the LPC-HMM and MFCC-HMM assumes 12-dimension, while the proposed MAP-HMM is a 5-dimensional model.

During testing, the trained pulse call HMM and noise HMM are combined before fed into the V-alg. The combined HMM ($\pi = \pi_1, \pi_2, \mathcal{T} = \mathcal{T}_1, \mathcal{T}_2, \phi = \phi_1, \phi_2$) contains eight states and four mixture models. The first 1 - 4 states represent the unwanted signal (noise), while states 5 - 8 represents the Bryde's whale call. The $\phi = \phi_1, \phi_2$ parameter is used to refine the extracted feature vector of the unknown signal to be detected. The $\mathcal{T} = \mathcal{T}_1, \mathcal{T}_2$ and $\phi = \phi_1, \phi_2$ parameters are used in the V-alg to specify the sequence of states, which ultimately classify the unknown signal as either the pulse call or noise. Note that an equal probability is specified in the transmission matrix $\mathcal{T}$ for the V-alg to switch from states 1 - 4 (noise) to states 5 - 8 (Bryde's whale call), and vice versa.

5.2. Comparative parameters

The performance of the proposed feature extraction technique is compared with the existing techniques based on these two standard means of measurement in sound processing Yack et al. (2009).

1. Sensitivity, $\mathcal{S}$: The $\mathcal{S}$ is used to measure the ability of automated detector to correctly detect the sound. It is represented as:

$$\mathcal{S} = \frac{\delta}{\delta + \omega}, \quad (14)$$

where δ is called the true positives (TP), which is the number of calls that were correctly detected. The notation ω is called the false negatives (FN), which is the number of times the automated sound detector misses the manually identified sound. A high value of $\mathcal{S}$ is desirable in all cases, which in turn gives a small value of the miss rate $\mathcal{M}$ defined as:

$$\mathcal{M} = 1 - \mathcal{S}. \quad (15)$$

2. False Discovery Rate, $\mathcal{F}_\mathcal{D}$: The $\mathcal{F}_\mathcal{D}$ measures the reliability of the automated detector. The $\mathcal{F}_\mathcal{D}$ is represented as:

$$\mathcal{F}_\mathcal{D} = \frac{\varphi}{\delta + \varphi}, \quad (16)$$

where φ is called the false positives (FP), which is the number of incorrect Bryde's whale call detections. A small value of the $\mathcal{F}_\mathcal{D}$ indicates that the detector is dependable.

5.3. Discussion

Tables 1–3, shows the performance of the LPC-HMM, MFCC-HMM and proposed MAP-HMM detectors as a function of k , and for different value of t . In Table 1, Table 2 and Table 3, the number of sampling points is varied to $t = 500, t = 300$ and $t = 100$ respectively. From Table 1 ($t = 500$), the MAP-HMM offers a $\mathcal{S}$ performance gain of 2.62% and 8% in comparison to the MFCC-HMM and LPC-HMM respectively at $k = 18$. Likewise, at $k = 12$ and $= 6$, the MAP-HMM obtained a superior performance gain of 2.98% and 11.11%, and 2.31% and 14.38% over the MFCC-HMM and LPC-HMM respectively. More so, the MAP-HMM is more reliable in comparison to the MFCC-HMM and LPC-HMM. Because at $= 18, = 12$ and $= 6$, the MAP-HMM obtained a $\mathcal{F}_\mathcal{D}$ of 0.11%, 0.12%, and 0.15% respectively, which are lower in comparison to 0.15%, 0.16%, and 0.25% for the MFCC-HMM and 1.31%, 1.47%, and 1.52% for the LPC-HMM.

Similarly, in Table 2 ($t = 300$) and Table 3 ($t = 100$), the MAP-HMM exhibit superior performance in comparison to the MFCC-HMM and LPC-HMM. However, the performance difference is not as wide as compared to the result in Table 1. For example, at $= 18$ in Table 2 ($t = 300$) and Table 3 ($t = 100$), the performance gain of the MAP-HMM over the MFCC-HMM and LPC-HMM is 1.77% and 6.96% and 1.09% and 1.75% respectively, which are low in comparison to the 2.98% and 11.11% in Table 1 ($t = 500$). As shown, as the number of sampling points reduces, the performance of the MFCC and LPC feature extraction methods slightly improves. Nonetheless, the smaller the value of, the

Table 1
Feature extraction performance combined with the HMM: $t = 500$.

1–103 k	$\mathcal{S}$ (%)			$\mathcal{M}$ (%)			$\mathcal{F}_\mathcal{D}$ (%)		
	MAP	MFCC	LPC	MAP	MFCC	LPC	MAP	MFCC	LPC
18	99.35	96.73	91.35	0.65	3.27	8.65	0.11	0.15	1.31
12	98.69	95.71	87.58	1.31	4.29	12.42	0.12	0.16	1.47
6	97.06	94.75	82.68	2.94	5.25	17.32	0.15	0.25	1.52

Table 2
Feature extraction performance combined with the HMM: $t = 300$.

k	$\mathcal{S}$ (%)			$\mathcal{M}$ (%)			$\mathcal{F}_\mathcal{D}$ (%)		
	MAP	MFCC	LPC	MAP	MFCC	LPC	MAP	MFCC	LPC
18	99.44	97.67	92.48	0.56	2.33	7.52	0.08	0.12	1.22
12	98.75	96.73	87.91	1.25	3.27	12.09	0.08	0.16	1.41
6	97.12	95.70	83.46	2.88	4.30	16.54	0.11	0.22	1.44

Table 3
Feature extraction performance combined with the HMM: $t = 100$.

k	$\mathcal{S}$ (%)			$\mathcal{M}$ (%)			$\mathcal{F}_\mathcal{D}$ (%)		
	MAP	MFCC	LPC	MAP	MFCC	LPC	MAP	MFCC	LPC
18	99.44	98.35	97.69	0.56	1.65	2.31	0.08	0.10	1.09
12	98.66	97.39	92.81	1.34	2.61	7.19	0.08	0.13	1.22
6	97.12	96.35	91.50	2.88	3.65	8.50	0.11	0.18	1.41

¹ <https://www.sonicvisualiser.com>

higher the number of windows (k), which in turns increase the number of times the MAP, MFCC, and LPC feature extraction operations are performed. That is, the value of k determines the number of times the whole feature extraction operation is performed. So, an increase in or decrease in, increases the computational time complexity of the whole detection process, which is not desirable for real-time detection or applications. Of interest, the MAP-HMM offered approximately similar performance in Table 2 ($k=300$) and Table 3 ($k=100$), which is better than the performance of the MFCC-HMM and LPC-HMM in Table 3 ($k=100$). Therefore, at $k=300$, the MAP-HMM obtained a performance that is better than the MFCC-HMM and LPC-HMM at $k=100$, with less computational time and delay complexity.

To buttress these results, we present the sensitivity and false discovery rate performance of the detectors with respect to k as depicted in Fig. 3a and Fig. 3b respectively. As previously explained, the LPC method offered the worst performance in both figures, while the MAP offered the best performance, even as k reduces. Note that the performance of the detectors improve, as the value of k increases in the figures and tables. This means that the performance of the detectors can further improve if we increase the value of k . Yet, we emphasise that at a

particular value of k , the performance of the HMM will saturate. As such, the detection accuracy of the model will not improve further, even if k keeps increasing. In some cases, the performance of the model can drop with further increment in the value of k . Summarily, the proposed MAP technique showcases two major advantages over the MFCC and LPC that will benefit the beaked whale research community. Firstly, it gives better detection accuracy at a reduced false discovery rate. Secondly, it reduces the computational complexity of the whole HMM detection process shown in Fig. 2; thus, it can be used in real-time sound detection. Finally, the proposed MAP is flexible and can be easily modified to analyse different cetacean vocalisation, where the parameters in the feature vector can empirically selected. Note, we employed mobile PAM method in recording the Bryde's whale short pulse calls. However, in practice, irrespective of the recording technique used (that is, fixed PAM or mobile PAM methods), the proposed MAP can be used as presented in Section 4.3 or slightly modified depending on the observable characteristics of the call, so long as the relative amplitude and phase of the signal is preserved as much as possible.

6. Conclusion

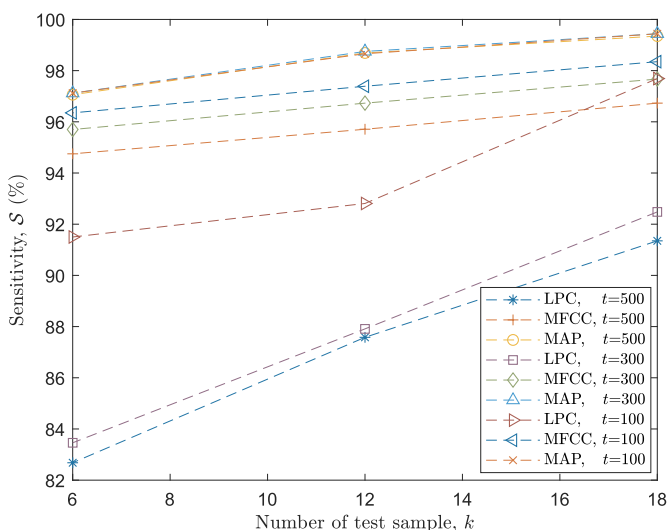
This article introduced a simple but robust feature extraction technique that can be easily adapted with the HMM. This proposed feature extraction method uses a combination of some simple but robust characteristics of the sound to be detected; such as, the mean, relative amplitude and relative power. The detection accuracy and false discovery rate of the proposed MAP-HMM was verified using the acoustic dataset of continuous recordings of an inshore Bryde's whale short pulse call collected in False bay, South Africa. The proposed MAP-HMM method proved to be very effective in comparison to the existing MFCC-HMM and LPC-HMM methods. It offered good sensitivity and false discovery rate performance, at a reduced computational time and delay complexity as compared to the studied methods in this article. Thus, we suggest that the MAP-HMM would be a useful real-time approach to investigating the movement of different obscure but vocal cetaceans, like the inshore Bryde's whales that produce characteristics call.

Acknowledgements

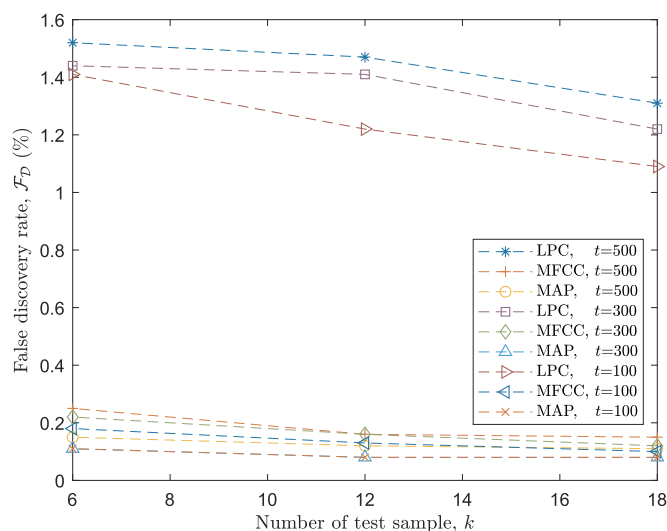
The financial assistance of the National Research Foundation (NRF) of South Africa towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF (Grant number: 116036).

References

- Committee on Taxonomy, 2017. List of marine mammal species and subspecies. Soc. Mar. Mammal. www.marinemammalscience.org (download: April, 2019).
- Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41, 164–171.
- Best, P., 1977. Two allopatric forms of Bryde's whale off South Africa. *Rep. Int. Whal. Comm.* 1, 10–38.
- Bilmes, J.A., 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. In: *Technical Report ICSI-TR-97-021*. 4 University of Berkeley.
- Brown, J.C., Smaragdis, P., 2009. Hidden Markov and Gaussian mixture models for automatic call classification. *J. Acoust. Soc. Am.* 125, EL221–EL224.
- Brown, J., Miller, J.O., P., 2007. Automatic classification of killer whale vocalizations using dynamic time warping. *J. Acoust. Soc. Am.* 122, 1201–1207.
- Churbanov, A., Winters-Hilt, S., 2008. Implementing EM and Viterbi algorithms for hidden Markov model in linear memory. *BMC Bioinformatics* 9.
- Constantine, R., Iwata, T., Nieukirk, S.L., Penry, G.S., 2018. Future directions in research on bryde's whales. *Front. Mar. Sci.* 5, 333.
- Duda, R., Hart, P., Stork, G.D., 2001. *Pattern Classification*, 2nd ed. .
- Florian, E., Martin, W.O.L., Schuller, B.O., 2010. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462.
- Forgy, E.W., 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classification. *Biometrics* 21, 768–769.
- Gupta, S., Jaafar, J., Wan Ahmad, W.F., Bansal, A., 2013. Feature extraction using MFCC.



(a) Sensitivity



(b) False discovery rate

Fig. 3. Performance comparison for the different feature extraction methods.

- Sign. Image Process. 4, 101–108.
- Jurafsky, D., James, H.M., 2018. Speech and language processing. In: Downloaded: 15th April, 2019, pp. 1–17.
- Kerosky, S.M., Širović, A., Roche, L.K., Baumann-Pickering, S., Wiggins, S.M., Hildebrand, J.A., 2012. Bryde's whale seasonal range expansion and increasing presence in the Southern California bight from 2000 to 2010. *Deep-Sea Res. I Oceanogr. Res. Pap.* 65, 125–132.
- Khalid, S., Khalil, T., Nasreen, S., 2014. A survey of feature selection and feature extraction techniques in machine learning. In: 2014 Science and Information Conference, pp. 372–378.
- MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, pp. 281–297.
- Majeed, S., Husain, H., Abdul Samad, S., Idbeaa, T., 2015. Mel frequency cepstral coefficients (MFCC) feature extraction enhancement in the application of speech recognition: a comparison study. *J. Theor. Appl. Inf. Technol.* 79, 38–56.
- Makhoul, J., 1975. Linear prediction: a tutorial review. In: *Proceedings of the IEEE*. 63, pp. 561–580.
- Martin, M., Jeroen, B., 2003. Features for audio and music classification. In: *Proceedings of the International Symposium on Music Information Retrieval*, pp. 151–158.
- Min, C.-H., Tewfik, A., 2010. Automatic characterization and detection of Behavioral patterns using linear predictive coding of accelerometer sensor data. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 220–223.
- Ogundile, O.O., Versfeld, D.J.J., 2020. Analysis of template-based detection algorithms for inshore Bryde's whale short pulse calls. *IEEE Access* 8, 14377–14385.
- Olsen, Ø., 1913. On the external characters and biology of Bryde's whale (*Balaenoptera brydei*), a new Rorqual from the coast of South Africa. In: *Proceedings of the Zoological Society of London*, pp. 1073–1089.
- Pace, F., 2008. Comparison of Feature Sets for Humpback Whale Song Classification. Master dissertation. University of Southampton.
- Penry, G. S., Findlay, K., & Best, P. (2016). A conservation assessment of *Balaenoptera edeni*. In Child MF, Roxburgh L, Do Linh San E, Raimondo D, Davies-Mostert HT, Editors. *The Red List of Mammals of South Africa, Swaziland and Lesotho*. South African National Biodiversity Institute and Endangered Wildlife Trust, South Africa.
- Penry, G.S., Hammond, P.S., Cockcroft, V.G., Best, P.B., Thornton, M., Graves, J.A., 2018. Phylogenetic relationships in southern African Bryde's whales inferred from mitochondrial DNA: further support for subspecies delineation between the two allopatric populations. *Conserv. Genet.* 19, 1349–1365.
- Putland, R., Ranjard, L., Constantine, R., Radford, C., 2018. A hidden Markov model approach to indicate Bryde's whale acoustics. *Ecol. Indic.* 84, 479–487.
- Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proceedings of the IEEE*. 77, pp. 257–286.
- Rabiner, L.R., Juang, B.H., 1986. An introduction to hidden Markov models. *IEEE ASSP Mag.* 3, 4–16.
- Rabiner, L., Shafer, R., 1978. *Digital Processing of Speech Signals*. Prentice-Hall International Inc., London.
- Reynolds, D., 2009. Gaussian mixture models. In: Li, S.Z., Jain, A. (Eds.), *Encyclopedia of Biometrics*. Springer US, Boston, MA, pp. 659–663.
- Rice, A.N., Palmer, K.J., Tielens, J.T., Muirhead, C.A., 2014. Potential Bryde's whale (*Balaenoptera edeni*) calls recorded in the northern Gulf of Mexico. *J. Acoust. Soc. Am.* 135, 3066–3076.
- Rosel, P.E., Wilcox, L.A., 2014. Genetic evidence reveals a unique lineage of Bryde's whales in the northern Gulf of Mexico. *Endanger. Species Res.* 25, 19–34.
- Shi, L., Ahmad, I., He, Y., Chang, K., 2018. Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments. *J. Commun. Netw.* 20, 509–518.
- Širović, A., Bassett, H.R., Johnson, S.C., Wiggins, S.M., Hildebrand, J.A., 2014. Bryde's whale calls recorded in the Gulf of Mexico. *Mar. Mammal Sci.* 30, 399–409.
- Smith, B.D., Kershaw, F., Minton, G., Rosenbaum, H.C., Leslie, M.S., Anderson, R.C., LeDuc, R.G., Baldwin, R., Brownell, J., Robert, L., Mansur, R.M., Collins, T., 2013. Population differentiation of 2 forms of Bryde's whales in the Indian and Pacific oceans. *J. Hered.* 104, 755–764.
- Vilorio-Gomora, L., Romero-Vivas, E., R., J. U., 2015. Calls of Bryde's whale (*Balaenoptera edeni*) recorded in the Gulf of California (L). *J. Acoust. Soc. Am.* 138, 2722–2725.
- Viterbi, A., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* 13, 260–269.
- Yack, T., Barlow, J., Rankin, S., Gillespie, D., 2009. Testing and Validation of Automated Whistle and Click Detectors Using PamGuard 1.0 NOAA Technical Memorandum NMFS. U. S. Department of Commerce.
- Yao, R., Johnson, M., Clemens, P., Michael, D., Glaeser, S., Osiejuk, T., Ebenezer, O.-N., 2009. A framework for bioacoustic vocalization analysis using hidden Markov models. *Algorithms* 2, 1410–1428.
- Zheng, F., Zhang, G., Song, Z., 2001. Comparison of different implementations of MFCC. *J. Comput. Sci. Technol.* 16, 582–589.