Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Wavelet Scattering Tranforms Applied to Whale Vocalisations

by

Marco Wiehann Rademan

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

Supervisor: D.J.J. Versfeld
Co-supervisor: J.A. Du Preez

2024

Declaration

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

Date: 14th May 2024

Abstract

Wavelet Scattering Tranforms Applied to Whale Vocalisations

M.W. Rademan

Department of Civil Engineering,
Stellenbosch University,
Private Bag X1, 7602 Matieland, South Africa.

Dissertation: PhD

2024

Uittreksel

Wavelet Scattering Tranforms Applied to Whale Vocalisations

M.W. Rademan

Departement Siviele Ingenieurswese,
Stellenbosch Universiteit,
Privaatsak X1, 7602 Matieland, Suid Afrika.

Proefskrif: PhD

2024

Contents

List of Figures

List of Tables

Nomenclature

Notation

Some notations are used interchangably, which is obvious from the context. Due to the nature of the publication format, notation is explained within the article it occurs.

$f(x)$	a continuous function of x
$f(x, y)$	a continuous multi-variable function of x and y
$f[n]$	a discrete-time (sampled) function indexed by n
$f[n, m]$	a multi-variable discrete-time function indexed by n and m
$\mathcal{F}\{f\}$	the Fourier transform operator acting on f
$\hat{f}$	the Fourier transform of f or an estimate of f, obvious from context
x^*	the complex conjugate of x
$p(x)$	a probability density
$p(x\|y)$	a conditional probability density
$\{a, b, ...\}$	a set of numbers
$A \setminus B$	set minus: the set A excluding the elements of B
$A \times B$	Cartesian product of sets
$\|x\|$	magnitude of a number (modulus), or size of a set
$\mathbb{R}$	the set of real numbers
$\mathbb{C}$	the set of complex numbers
$\mathbb{N}^+$	the set of natural numbers (excluding 0)
$\mathbb{Z}$	the set of integers
$x \sim p$	x is drawn from the probability density p
$\frac{d}{dx} f$	derivative with respect to x of a function f
$\mathcal{U}_j x$	scalogram of the j'th level of x
$\mathcal{S}_j x$	scattering coefficients of the j'th level of x
$x * y$	convolution of two functions
$x \otimes y$	circular convolution of two discrete functions
$(x)_{\downarrow r}$	a discrete function subsampled by a factor r
$\mathbf{x}$	a vector

$X[k,m]$	a time-frequency decomposition of x, with k the frequency index and m the time index
$S[k,m]$	power spectrum calculated from $X[k,m]$
$H[m]$	spectral entropy calculated from $S[k,m]$
$X(s,b)$	continuous wavelet transform of x, with s as scale and b as translation
ψ_λ	sometimes used as short-hand for $\psi(\lambda t)$
$\mathrm{RM}_m^F(x)$	rolling median filter operator with a window of size F operating on index variable m on a discrete function x
$\mathrm{RA}_m^F(x)$	rolling average filter operator with a window of size F operating on index variable m on a discrete function x
$x \triangleq y$	x is defined as y

Common Variables

Some variables are used interchangably, which is obvious from the context. Due to the nature of the publication format, variables are explained within the article it occurs.

a	Hello
a	Hello
a	Hello
a	Hello

Acronyms

GMM	Gaussian mixture model
DT	discrete-time
POI	point(s) of interest
EM	expectation maximization
SE	spectral entropy
PAM	passive acoustic monitoring
PSD	power spectral density
STFT	short-time Fourier transform
TPR	true positive rate
FPR	false positive rate
PDF	probability density function
SNR	signal-to-noise ratio
AMFM	amplitude and frequency modulated
BLED	band-limited energy detection
TF	time-frequency
CWT	continuous wavelet transform
DWT	discrete wavelet transform
SWT	stationary wavelet transform
WPD	wavelet packet decomposition
MODWT	maximal overlap discrete wavelet transform
MFCC	Mel-frequency cepstral coefficient
DNN	deep neural network
CNN	convolutional neural network
HMM	hidden Markov model
SOTA	state-of-the-art
LSTM	long-short-term memory
EWT	Endangered Wildlife Trust
WS	wavelet scattering
WST	wavelet scattering transform
FFT	fast Fourier transform
GPU	graphics processing unit
ROC	receiver operating characteristic
CLS	cosine-log scattering
LDA	linear discriminant analysis
ML	machine learning
NN	neural network
SVM	support vector machine

1D	1-dimensional
2D	2-dimensional
SIFT	scale-invariant feature transform
LPF	low-pass filter
IR	impulse response
FIR	finite impulse response
AUC	area under curve
SD	standard deviation
STPR	soft true positive ratio
STNR	soft true negative ratio
BPF	band-pass filter

1 Introduction

2 Time-Frequency Analyses

A time-frequency (TF) analysis is one of the most important tools in audio classification, data preparation and data manipulation. The main idea behind TF decompositions is to split a signal into frequency bands, which can viewed as time-varying functions.

Formally, given a 1-dimensional (1D) discrete-time (DT) signal $x[n]$, with an index variable n, a TF analysis decomposes x as

$$X[k,n] = \sum_{i} x[n]\upsilon_k[i-n] = x * \upsilon_k, \tag{2.1}$$

where $\upsilon_k \in \Upsilon$ is the k'th band-pass filter (BPF) in the decomposition filter dictionary Υ. Note that a low-pass filter (LPF) can be considered to be a BPF with a centre-frequency of 0.

Generally, the entire convolution $X[n,k]$ is not computed. Instead, only select samples are computed by speficying a hop-distance or downsample factor d. We indicate this by tranforming the index variable $m = dn$ as

$$X[k,m] \triangleq X[k,dn] = (x * \upsilon_k)_{\downarrow d}\,. \tag{2.2}$$

The is equivalent to applying a downsample operator $(\cdot)_{\downarrow d}$ to the convolution.

2.1 Fourier Transforms

A continuous-time (CT) Fourier transform of $x(t)$ is given as

$$\mathcal{F}\{x\} \triangleq \hat{x}(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt = \left(x * e^{j\omega t}\right)\Big|_{t=0}\,. \tag{2.3}$$

Derived from the conventional Fourier transform, a DT signal $x[n], n \in \{0, ..., N-1\}$

(length N) has a discrete-time Fourier transform (DTFT)

$$\text{DTFT}\{x\} \triangleq \hat{x}[k] = \sum_{n=0}^{N} x(t)e^{-j\frac{2\pi kn}{N}} = (x \otimes \upsilon_k)\Big|_{n=0}, \tag{2.4}$$

for which $\upsilon_k = e^{j\frac{2\pi kn}{N}}$, $k \in \{0, ..., N-1\}$ forms an orthogonal basis of x. The circular convolution operator $\otimes$ implies the infinite periodic extension of x beyond its boundaries of $n \in \{0, ..., N-1\}$.

Equation (2.4) describes the DTFT as convolution evaluated at the index 0 as a useful link to the generalised notion of TF analyses: Fourier transforms can be considered as complex sinusoidal filters being applied to the signal of interest, with periodic boundary extension.

The DTFT is commonly used to quickly compute convolutions for which the fast Fourier transform (FFT) is a fast alorithm with $N \log N$ computational complexity. In particular, to circularly convolve signals x and y, a fast computation is implemented as

$$x \otimes y = \text{FFT}^{-1}\{\hat{x} \cdot \hat{y}\}. \tag{2.5}$$

With the correct padding procedures, one can compute $x * y$ using (2.5).

Fourier transforms as decribed above are invertible operators, since they are orthogonal decompositions. However, for TF analyses, invertibility is not necessarily a requirement of the application. As such, this disseration does not consider invertibility for the remainder of the content.

2.2 Constructing Filters

The most popular TF analysis is the short-time Fourier transform (STFT).

In terms of equation (2.2), a STFT is computed with

$$\upsilon_k[n] = w[n]e^{j\frac{2\pi kn}{N}}, \tag{2.6}$$

where w is a window function defined by the user. Various FFTs are use to compute this transform, so the properties of w is typically specified by the FFT length, and the window type (Hann, Hamming, Blackman, etc.). The hop-length (or downsample factor) d is an additional parameter specified by the user.

The frequency domain content $\hat{w}$ entirely specifies the filter behaviour of υ_k, since the factor $e^{j\frac{2\pi kn}{N}}$ only serves to specify the centre frequency of υ_k. The resulting filter dictionary Υ is therefore comprised of constant-bandwidth analytic (complex) filters, linearly spaced in the frequency domain. As such, depending on the definition of filter bandwidth, there exists an optimal d which captures all of the necessary information of $|X[k,m]|$ such that no aliasing occurs across the index dimension m. This type of optimality is discussed, at length, for wavelet filterbanks in chapter 7.

The linearly-spaced filters and constant-bandwidth does not necessarily reflect and/or capture the natural phenomena typically present in audio signals. In particular, an audio signal can be expressed as a series of harmonics above its fundamental. This motivates the use of exponentially placed filters.

Formally, given a signal with a fundamental frequency most accurately captured in the frequency bin k_0, any change in frequency by a factor a will move the fundamental frequency content to the bin ak_0. This has dire consequences for lower frequency signals - the frequency resolution of lower frequency signals is significantly poorer compared to the frequency resolution of its own harmonics. As such, a small frequency shift of a lower frequency signal with ample harmonic content which cause large index shifts for its harmonics, i.e., the b'th harmonic will be shifted to index abk_0.

Suppose that we design Υ to contain Q filters per octave. A frequency change by a factor a shifts the frequency index of the fundamental to $(Q\log_2(a)+1)k_0$. Similarly, the b'th harmonic's frequency index is shifted to $(Q\log_2(a)+1)(Q\log_2(b)+1)k_0$. For example, if $a=2$, the fundamental's frequency index is shifted to k_0+Q (as opposed to $2k_0$ for the STFT). This property is referred to a frequency-shift invariance, since small frequency deformations of the order Q do not significantly change the frequency index structure of the TF decomposition. The STFT is much more sensitive to such deformations across its entire structure for harmonically rich signals.

An exponential filterbank design has increased stability to small frequency fluctuations, which can make it more useful for feature-extraction methods in a machine learning (ML) pipeline, since feature vectors are more stable against frequency deformations. The exponential filterbank is, in fact, the way wavelet filters are typically constructed. Additionally, the Mel-frequency cepstral coefficients (MFCC) filterbank attributes much of its sucess (compared to raw STFT magnitude coefficients) due to its exponentially placed filters.

2.3 Wavelet Filters

To construct a wavelet filterbank with $\upsilon_k \in \Upsilon$ which has the exponentially-spaced frequency properties discussed in section 2.2, we use a "template" wavelet $\psi(t)$, referred to as the mother wavelet. Conventionally, wavelet filters are defined as CT filters and then discretised in software implementations.

The mother wavelet is a BPF with the following properties:

1. Centre frequency of 1 rad/s.
2. Zero mean: $\hat{\psi}(0) = 0$.
3. Bandwidth defined by the user.

uppsi can the be scaled by a factor λ such that

$$\upsilon_k(t) = \lambda_k \psi(\lambda_k t). \tag{2.7}$$

Notational shorthand is often used:

$$\psi_{\lambda_k}(t) \triangleq \lambda_k \psi(\lambda_k t). \tag{2.8}$$

We construct a set $\lambda_k \in \Lambda$ such that $\lambda_{k+1} = 2^{\frac{1}{Q}} \lambda_k$, which ensures Q wavelets per octave. Since ψ has a frequency of 1 rad/s, λ_k becomes the centre frequency of ψ_{λ_k}, while also scaling its bandwidth by a factor of λ_k.

An additional advantage resulting from bandwidth scaling is that higher frequencies become more localised in time, allowing for better localisation precision as frequency increases. This results from the fact that ψ_{λ_k} maintains the same number of oscillations over scale, which is not true for the STFT.

Chapter 3 discusses the applications of such a filterbank in terms of the continuous wavelet transform (CWT), obtaining a scalogram. Note that traditional CWT notation typically uses the scale variable $s = \frac{1}{\lambda}$, where s is made continuous. Discretisations of the CWT approximate the transform by limiting s to the set $s_k \in \frac{1}{\lambda_k}$, $\lambda_k \in \Lambda$. Chapter 3 specifically considers the application of the CWT for signal detectors, while also using other techniques to improve traditional methods.

3 Wavelet Transforms as a Fourier Alternative

M.W. Rademan, D.J.J Verself, J.A. Du Preez

Accepted in Ecological Informatics on 15 May 2024

doi: doi

Abstract

Underwater passive acoustic monitoring systems record many hours of audio data for marine research, making fast and reliable non-causal signal detection paramount. Such detectors assist in reducing the amount of labor required for signal annotations, which often contain large portions devoid of signals.

Cetacean vocalization detection based on spectral entropy is investigated as a means of vocalization discovery. Previous techniques using spectral entropy mostly consider time-frequency enhancement of the entropy measure, and utilize the short time Fourier transform (STFT) as its time-frequency (TF) decomposition. Spectral entropy methods also requires the user to set a detection threshold manually, which call for knowledge of the produced entropy measures.

This paper considers median filtering as a simple, effective way to provide temporal stabilization to the entropy measure, and considers the continuous wavelet transform (CWT) as an alternative TF decomposition. K-means clustering is used to determine the threshold required to accurately separate the signal/no-signal entropy measures, resulting in a one-dimensional, two-class classification problem. The class means are used to perform pseudo-probabilistic soft class assignment, which is a useful metric in algorithmic development. The effect of median filtering, signal-to-noise ratio and the chosen TF decomposition are investigated.

The accuracy and specificity measures of the proposed detection technique are simulated using a pulsed frequency modulated sweep, corrupted by a sample of ocean noise. The results show that median filtering is particularly effective for low signal-to-noise ratios. Both the STFT and CWT prove to be effective TF analyses for signal detection purposes, each presenting with different advantages and drawbacks. The simulated results provide insight into configuring the proposed detector, which is compared to a conventional STFT-based spectral entropy detector using manually annotated humpback whale (*Megaptera novaeangliae*) songs recorded in False Bay, South Africa, July 2021.

The proposed method shows a significant improvement in detection accuracy and specificity, while also providing a more interpretable detection threshold setting via soft class assignment, providing a detector for use in development of adaptive algorithms.

3.1 Introduction

Underwater passive acoustic monitoring (PAM) systems consisting of hydrophones are often used to record marine activity [9, 79, 67]. PAM is a very effective tool which may aid in ecological monitoring, preservation and research. However, PAM systems present their own of set challenges relating to the field of digital signal processing. The recorded vocalizations are often from distant sources, which may result in poor signal-to-noise ratios. PAM systems record hundreds of hours of audio from multiple hydrophones, making it time-consuming to find and label vocalizations manually. Thus, reliable and computationally efficient signal detection is of great importance for data gathering purposes.

Many cetaceans, such as the humpback whale (*Megaptera novaeangliae*) and southern right whale (*Eubalaena australis*), produce vocalizations that are periodic in nature. These signals may be modelled as amplitude-frequency modulated (AMFM) sinusoids [21]. Recordings of these vocalizations are useful for conservation studies, like determining migration patterns or population estimates. It has become critical to monitor the effect of industry on marine mammals, and to take measures for ecological preservation. Numerous detection algorithms and monitoring systems have been developed to aid in this venture [78, 47].

Recent techniques which detect and/or classify cetacean calls often make use of machine learning (ML) models, using deep learning [25, 36] or hidden Markov models (HMM) [75, 51]. Due to the difficulty of obtaining large amounts of annotated data, neural network ML models often need to use transfer learning as in [36]. ML techniques often require additional feature extraction methods to reduce the dimensionality of the audio data, thus adding to computational complexity [52]. ML models cannot necessarily operate well outside the conditions of their training data. Training ML models may also be a cumbersome task [33], making simpler, unsupervised techniques which are more computationally efficient the preferred method of sifting through large amounts of data to identify points of interest. These points of interest allow biologists to ignore extensive portions of noise in the audio, thus greatly reducing manual work, before classifying the audio.

An example of such an efficient detector which is prevalent in PAM audio analysis software packages (PAMGuard [55] and Raven Pro [16]), is the band-limited energy detector (BLED) [79]. This detector is simple to compute, making it a popular choice for finding points of interest in long audio recordings. BLED has also seen use as part of multi-stage classifier systems [33].

Spectral entropy (SE) is another popular and simple method derived from Shannon's information entropy [68] to perform narrow-band signal detection, and has demonstrated success in both speech and cetacean vocalization detection, consistently outperforming BLED [19, 59, 27]. At its core, spectral entropy utilizes a time-frequency (TF) decomposition and treats each time-slice as a pseudo-probability density. SE will decrease in the presence of narrow-band frequency content. This method of narrow-band signal detection was first proposed by Powell and Percival in [58].

In real data, the SE measure may fluctuate due to changing noise conditions, resulting in less accurate detection. Various methods for speech detection aim to stabilize/enhance the SE as in [59, 27], which utilize weighting factors in the SE calculation as well as temporal averaging. For signal detection, the researcher must determine the appropriate SE threshold manually, which again increases labor, and requires some understanding of how to set the detection threshold properly. Furthermore, [59, 27] only consider the short-time Fourier transform (STFT) as a TF decomposition.

The continuous wavelet transform (CWT) has shown success in many applications, including feature extraction in ML models for electrocardiogram classification [70], analysis of periodic patterns in lake water levels [50] and detecting pachyderm movements using seismic sensors [56]. The CWT has previously been used in structural engineering applications as a TF decomposition for SE [13, 14], and may be considered as an alternative to the STFT for SE calculation.

Erbe and King [19] used SE to detect cetacean calls in the Canadian Arctic. They demonstrate its effectiveness compared to BLED, which SE outperforms on all accounts. Erbe and King used the simplest form of SE with the STFT, without any enhancement of the entropy measure, which will serve as a baseline comparison for this paper.

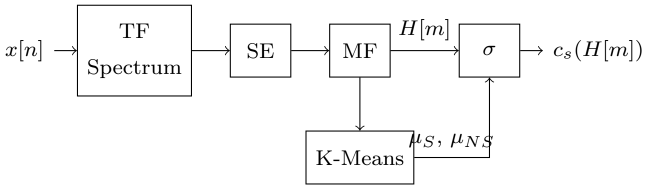

Figure 3.1: Block diagram of the proposed soft-output detector.

This paper compares the STFT and CWT as TF decompositions for the SE calculation, and proposes median filtering (MF) as a simple, yet effective, method of providing temporal stabilization to the SE measure. Additionally, a means of converting the SE to a pseudo-probability is proposed, using the k-means algorithm as a computationally simple way to distinguish between signal/noise SE segments. The pseudo-probability measure is more interpretable than SE, and is automatically scaled, making it robust against changes in ambient ocean conditions across multiple recorded files, which will typically require manually resetting the SE threshold. Refer to figure 3.1 for a description of the proposed solution.

To test the accuracy and specificity of our proposed solution compared to the baseline SE detector, this paper measures detector performance on a per-sample basis on the audio. Simulation is used to verify and design the proposed technique, and gain some insight on how SNR and hyper-parameters influence the detector in a controlled test environment. The proposed detector is also verified with manually annotated data, to demonstrate that it performs well in a practical setting. Per-sample accuracy is a necessary measure to test detectors, since software packages (PAMGuard; Raven Pro) allow the user to discard detections of short length. Thus, detecting only part of the signal is not suitable for all applications.

We demonstrate empirically that the L1-normalized CWT is a superior TF decomposition for similar practical applications for signal detection using SE. Median filtering is shown to be a simple and viable form of SE stabilization to improve signal detection accuracy and specificity. Additionally, we propose the K-means algorithm to provide appropriate scaling for the SE to be converted into a pseudo-probabilistic measure of signal presence, which may be used to develop adaptive algorithms and provide more intuitive threshold settings for non-experts. To the authors' knowledge, this paper is the first to introduce a soft-classification procedure for SE detectors through the k-means algorithm. Additionally, to the authors' knowledge, it is the first time the CWT is used in a SE detector for cetacean detection applications.

Improving SE-based detectors are of great importance for performing narrow-band signal detection. SE detectors offer good accuracy in identifying periodic signals, while requiring no training, and offering some robustness against fluctuation noise profiles. Depending on the TF decomposition parameters used, SE allows for accurate time-localized detections, which yields more detailed analysis on vocalization statistics, such as call length, and distance between calls, and may assist in creating accurate datasets via data augmentation for supervised ML models.

3.2 False Bay Data

The data used in this paper was recorded via a hydrophone at False Bay, on 1 July 2021 at midnight. Two humpback whale songs were extracted in the time frame between 00:00 and 00:30 hours. The hydrophones were anchored 6 kilometers off the coast. The hydrophone capsule recorded at 15 metres below sea level. The audio is sampled 32 kHz, which is resampled to 2 kHz. This is adequate for preserving the observed Humpback whale vocalizations, which consist of songs in the 150-1000 Hz band. The audio is normalized to have an average power of 1.

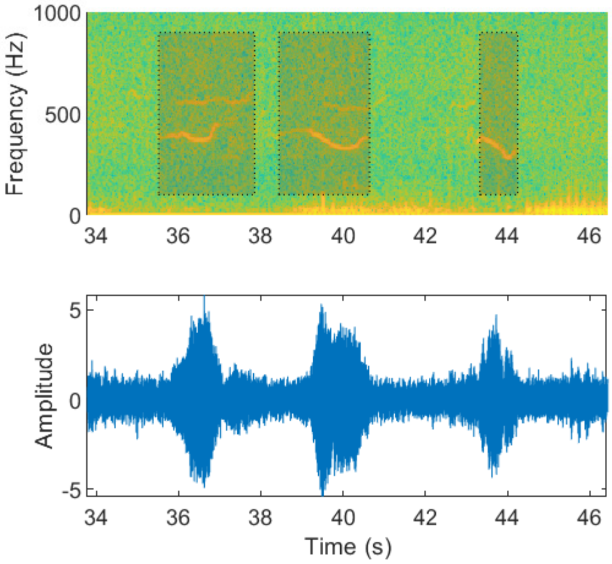

Figure 3.2: Audio snippet and spectrogram of a humpback song recorded in False Bay. The waveform shown is high-passed at 100 Hz, resampled and normalized. The signal annotations are shown on the spectrogram. The spectrogram is calculated before high-passing the signal.

Figure 3.2 shows a snippet of a recorded whale song. Many noise types can be observed. Low frequency anchor noise, caused by the swaying of the hydrophone in the water due to the anchoring system, contains the majority of signal power. This significantly corrupts all frequency content below 100 Hz. Thus, this noise is filtered out, since it does not influence the signals of interest, and is highly dependent on the anchoring system and the tides.

The remaining noise types are background noise, and impulsive noise, caused by snapping shrimp and fish. All other noise sources, such as waves, fish and boats are classified as background noise.

Due to the dominating anchor noise, the signal-to-noise ratio (SNR) is only calculated for frequencies above 150 Hz to the Nyquist limit (1 kHz). The recorded whale calls are estimated to have an average SNR of approximately 0 dB, which is calculated in the time-frequency regions of signal presence, utilizing a power spectrogram. The instantaneous SNR is estimated to range from -20 to 30 dB.

The songs were manually annotated. The starting and ending time of a vocalization is recorded, which therefore only indicates signal presence. Overlapping calls are annotated as one vocalization. The two songs have a combined length of 369 seconds, containing 120 annotated calls. The signals compose 32% of the recorded samples within the song segments. The remaining 68% of samples are considered as noise.

Figure 3.3 show the combined statistics of the humpback songs. Call length and the distance between calls are calculated from the manual annotations. The distance is measured between the endpoint of a vocalization and the start of the next vocalization. These statistics are useful in determining median filter window lengths limits.

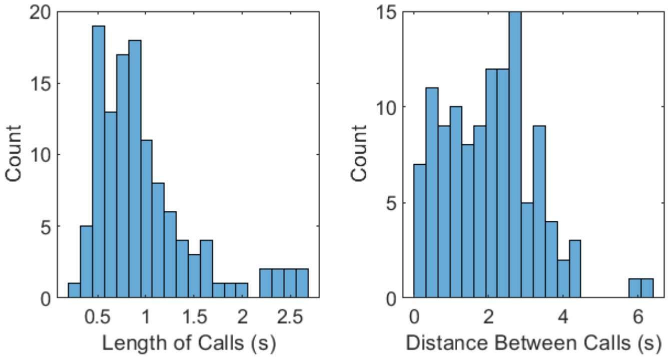

Figure 3.3: Histograms indicating the combined call length and distance between calls of the two humpback songs, for use in determining median filter window length limits.

3.3 Spectral Entropy

3.3.1 STFT

Given a discrete-time signal $x[n]$, the windowed STFT may be expressed as

$$X[k,m] = \sum_{n=m}^{m+N-1} w[m-n]x[n]e^{-j2\pi kn/N},$$

where w is the windowing function, N is the window size, $k \in \{0, 1, ..., N-1\}$ is the frequency index and m is the time index. The total number of frequency bins may be artificially increased by padding each window with additional zeros.

The power spectrum $S[k,m]$ may be derived from $X[k,m]$ as

$$S[k,m] = \frac{1}{N} X[k,m]X^*[k,m],$$

which may then be converted to a pseudo-probability density $P[k,m]$, referred to as the spectral distribution:

$$P[k,m] = \frac{S[k,m]}{\sum_{j=0}^{N-1} S[j,m]}. \tag{3.1}$$

The spectral distribution is used to construct a time-varying entropy measure $H[m]$, based on Shannon's entropy:

$$H[m] = -\sum_{k=0}^{N/2} P[k,m] \log(P[k,m]). \tag{3.2}$$

It is well known that the uniform distribution is the maximum entropy distribution on a finite interval [15]. For spectral entropy, this corresponds to the spectral distribution of white noise. In a sense, it is a measure of certainty for discrete probability distributions: a flat spectral distribution will have high entropy, whereas the spectral impulse from a sinusoid will have low entropy. Note that the summation only occurs to $N/2$ (even N assumed), due to the symmetry of the Fourier transform.

If a sinusoidal signal is known to be within a certain frequency range $f_0 \in (f_1, f_2)$, it may be beneficial to restrict the entropy measure between frequency indices:

$$H[m] = -\sum_{k=k_1}^{k_2} P[k,m] \ln(P[k,m]),$$

where $k_1 = \lfloor Nf_1/f_s \rfloor$ and $k_2 = \lfloor Nf_2/f_s \rfloor$, with f_s representing the sample frequency. This

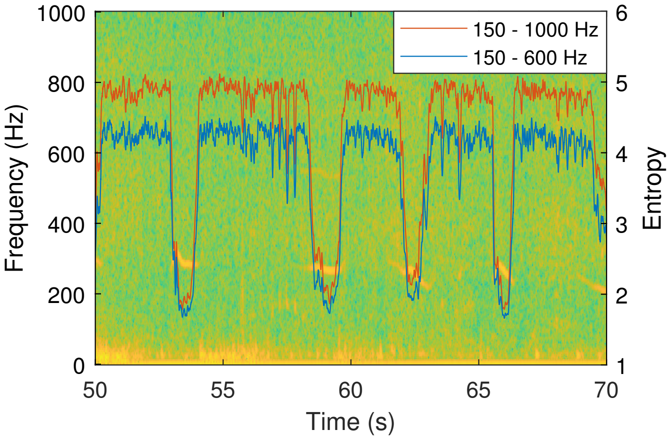

Figure 3.4: Spectral entropy (STFT) calculated for different frequency bin ranges.

will reduce the risk of signals within other bands contaminating the entropy measure. However, small frequency ranges are not advised, since a smaller frequency range will decrease the apparent “flatness" of the measurement, hence lowering the entropy difference between signal/no-signal segments.

3.3.2 CWT

The CWT is a form of TF decomposition which falls under the theory of multi-resolution analysis (MRA). Wavelet decomposition is a topic of many facets and contain a great deal of mathematical depth. Mallat [39] describes wavelet theory as an amalgamation of ideas from a variety of scientific fields, which were developed simultaneously.

From a signal processing point of view, the numerically calculated CWT may be interpreted as a set of finite impulse response (FIR) band-pass filters constructed to have a certain bandwidth and time support. This allows for more flexibility in the time-frequency resolution, depending on the chosen filter-bank parameters. The mother wavelet ψ may be considered as the main impulse response (IR) of the filters, which is scaled in time such that its centre frequency coincides with a frequency of interest.

The analysing wavelet used in this paper is the generalized Morse wavelet, described by Olhede and Walden in [53, 34], which is defined by its Fourier transform:

$$\mathcal{F}\{\psi_{\beta,\gamma}(t)\} = \Psi_{\beta,\gamma}(\omega) = U(\omega)\alpha_{\beta,\gamma}\omega^{\beta}e^{-\omega^{\gamma}},$$

with $\alpha_{\beta,\gamma} = (\frac{2e\gamma}{\beta})^{\beta/\gamma}$ as a normalizing constant, such that

$$\Psi(\omega_{\beta,\gamma}) = 2, \tag{3.3}$$

where $\omega_{\beta,\gamma} = (\frac{\beta}{\gamma})^{1/\gamma}$ is the peak frequency. The time-bandwidth product of the wavelet $P^2 = \beta\gamma$ determines the wavelet time support, which is proportional to $P = \sqrt{\beta\gamma}$. $U(\omega)$ represents the Heaviside step function:

$$U(\omega) = \begin{cases} 1, & \omega > 0 \\ 0, & \omega \leq 0 \end{cases},$$

which removes all negative frequency components, making the Morse wavelet analytic (contains one-sided frequency information).

The frequency symmetry of the wavelet is controlled by γ. A symmetric Morse wavelet has $\gamma = 3$. β further controls the time support of the wavelet, or equivalently, the filter bandwidth and IR length. Any analytic wavelet with a tunable bandwidth may also be used to yield similar results. The bandwidth must be set narrow enough so that the filters do not significantly overlap given a number of filters per octave. A filterbank that adequately covers the entire frequency range of interest without significant overlap between filters provides optimal results since a sharper frequency resolution creates a lower entropy.

The CWT of a signal $x(t)$ is defined as:

$$X(s,b) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t)\psi^* \left(\frac{t-b}{s}\right) dt, \tag{3.4}$$

where s is the scale parameter which expands/contracts the analyzing wavelet and b is the time-shift parameter. ψ^* denotes the complex conjugate of ψ. Equation (3.4) corresponds to band-pass-filtering $x(t)$ with the impulse response described by the scaled wavelet. The factor $|s|^{-1/2}$ provides L2 normalization, ensuring that energy is preserved with scaling.

In this paper, the MATLAB wavelet toolbox is used to perform CWT and frequency analysis. More information on how this toolbox performs CWT calculation may be found at [43]. The number of band-pass filters is specified by the number of voices per octave, which result in a discrete set of scale parameters $\{s_1, s_2, ..., s_N\}$ which are related to one

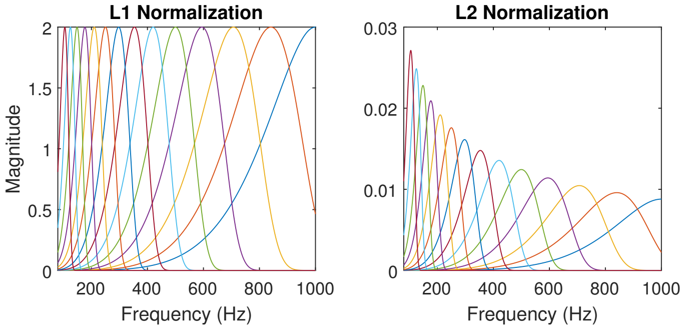

Figure 3.5: An example of a CWT filterbank, shown as the L1 and L2 normalized versions , with $\gamma = 10$, $\beta = 50$ and 4 voices per octave.

another by a factor $s_{n+1}/s_n = \sqrt[p]{2}$. This toolbox limits the ratio of the time/bandwidth parameters to $\beta_{max} = 40\gamma$.

Given a starting frequency f_1, then

$$s_1 = \frac{\omega_{\beta,\gamma}}{2\pi f_1}.$$

When calculating the CWT numerically, the MATLAB wavelet toolbox instead uses L1 normalization of the wavelets, thus preserving amplitude and not energy. The scaling factor correction $|s|^{-\frac{1}{2}}$ in (3.4) is therefore replaced by a factor of $|s|^{-1}$ in the MATLAB implementation. As such, all band-pass filters have a maximum gain of 2 which compensates for the analytic nature of the Morse wavelet, since the combined positive and negative frequency components of a real sinusoid captures the amplitude.

The L1 normalized CWT is defined as

$$X(s,b) = \frac{1}{|s|}\int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-b}{s}\right) dt. \tag{3.5}$$

More information on the fast numerical calculation of the CWT may be found in [8]. Figure 3.5 shows an example of the resulting MATLAB Morse wavelet filterbank (L1-normalized), and its L2-normalized version.

Given then the discretized/numerical calculation of the CWT with L1 normalization, $X[k,m]$, where m denotes the discrete-time index of the transform, and k denotes the index corresponding to s which refers the centre frequencies (f_c) of the band-pass filters, a new pseudo-power spectrum may be utilized for entropy calculation. For a sinusoid with amplitude A and a frequency corresponding to any k_0 of the band-pass filters' centre frequencies, the resulting amplitude of $|X[k_0, m]|$ will be A from (3.3). Thus, the power of a sinusoid observed within a band is

$$S[k, m] = \frac{1}{2} X[k, m] X^*[k, m] \tag{3.6}$$

since the average power of a sinusoid with amplitude A is $A^2/2$. For the CWT, (3.6) refers to the equivalent average power of a sinusoid at time index m and frequency band k.

The bandwidth of the filter influences the observed power, especially in the presence of noise. However, the equivalent sinusoidal power may be more useful in most applications. When a sinusoid is present, the equivalent sinusoidal power does not reduce as frequency increases (as with L2 normalization). However, higher frequency wavelets will be more sensitive to noise, due to their increased bandwidth.

SE is calculated by substituting the CWT equivalent sinusoidal power from equation (3.6) into equations (5.14) and (5.15).

3.4 Signal detection

3.4.1 K-means

Many variants of the k-means algorithm were independently developed by researchers in the 1950's [54]. The name was first coined by MacQueen in [37]. It has a rich history as a method of unsupervised classification, used in a wide set of applications. Its simplicity and interpretability make it a popular choice for many classification problems. In HMM classifiers, the k-means clustering algorithm is often used to initialize Gaussian mixture models (GMM) prior to data fitting [75, 51, 52]. In this paper, k-means is used to determine the class means of the entropy measure given signal/no-signal conditions. This eliminates the need for a user-set threshold on the entropy and provides class means which may be used to construct a pseudo-probabilistic measure of signal presence. Refer to appendix 3.A for details on the implementation for this application.

3.4.2 Soft Output

Given the class means μ_S (signal), μ_{NS} (no signal) determined by k-means clustering, the k-means decision boundary β is the midpoint:

$$\beta = \mu_S + \frac{\mu_{NS} - \mu_S}{2}$$

In classical K-means, the assigned class $c_h(H[m])$ results in a hard output:

$$c_h(H[m]) = \begin{cases} 1, & H[m] \leq \beta \\ 0, & H[m] > \beta \end{cases}$$

where a value of 1 corresponds to signal presence and a value of 0 representing no signal.

To provide a soft output, a function which smoothly transitions from one class to another in the interval $H[m] \in [\mu_S, \mu_{NS}]$ is required. Since an entropy measurement which is lower than μ_S is more probable to contain signal and, conversely, an entropy measurement which is greater than μ_{NS} is more probable to be noise, the transition function $\alpha(t)$ should satisfy the following criteria:

1. $\alpha(t)$ is monotonically decreasing.
2. Approaches hard classification in the limit: $\lim_{t\to\infty} \alpha(t) = 0$ and $\lim_{t\to-\infty} \alpha(t) = 1$.
3. Equal class probability at the boundary: $\alpha(t-\beta)|_{t=\beta} = \alpha(0) = \frac{1}{2}$.
4. Class assignment symmetry: $\alpha(t) = 1 - \alpha(-t)$.

The sigmoid function $\sigma(x)$ evaluated at $x = -t$ satisfies all the above criteria, defined as

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

$\sigma(x)$ is proposed in the style of logistic regression [17]. A piecewise-linear function may also satisfy the above criteria.

The shape of the sigmoid function may be set by specifying the gain $g > 0$ as a hyper parameter. A higher gain will result in a steeper shape, thus attributing more certainty to entropy measurements closer to the class means, and vice versa for lower gains. It may also be beneficial to normalize the entropy measure such that $H[m] = \mu_S$ and

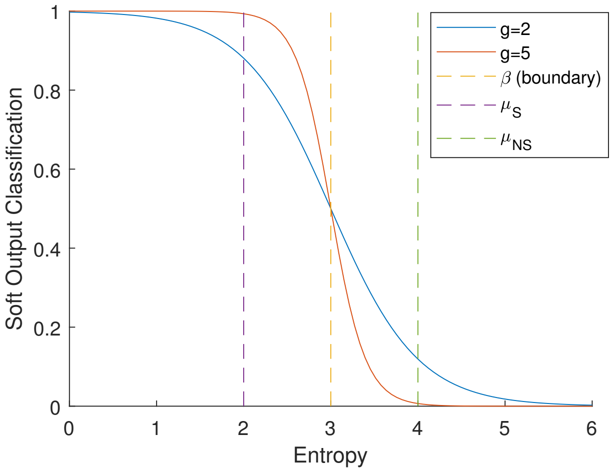

Figure 3.6: Example of the sigmoid function for soft output classification for various gains ($\mu_S = 2$, $\mu_{NS} = 4$).

$H[m] = \mu_{NS}$ map to ±1. The normalization will ensure consistency in the behaviour of the hyper parameter g over a range of class means. Figure 3.6 shows the effect of g on the classification output.

The soft-output classification function $c_s(H[m])$ is therefore

$$c_s(H[m]) = \sigma\left(-2g\frac{H[m]-\beta}{\mu_{NS}-\mu_S}\right).$$

To assign a specific pseudo-probability $c_s(H[m]) = p$ that corresponds to signal presence when $H[m] = \mu_S$, then the gain hyper parameter g is calculated as

$$g = \frac{1}{2}\frac{\mu_{NS}-\mu_S}{\mu_S-\beta}\ln\left(\frac{1}{p}-1\right) \tag{3.7}$$

For $g > 0$, equation (3.7) requires $p > \frac{1}{2}$.

3.4.3 Median Filtering

Median filtering (MF) is often used to suppress “salt-and-pepper" noise in image processing, and shown past success as a simplistic method of denoising images and signals [28, 23].

Median filtering may assist in suppressing the fluctuations of instantaneous spectral entropy measures prior to K-means, but preserve the “step responses" present in the entropy measure when transitioning from signal to no-signal and vice versa. The entropy of background noise in practical data may not be as consistent as the entropy of GWN. It will therefore help suppress any outliers within the filtering window.

If the window is chosen large enough, it will suppress the entropy fluctuations of signals that are shorter than the half the window length. This may be useful if a restriction on signal length is imposed.

Given a discrete-time signal $x[n]$, and an window of odd length M, the median-filtered signal $y[n]$ is

$$y[n] = \mathrm{med}(x[n-R], ..., x[0], ..., x[n+R]),$$

where $R = \frac{M-1}{2}$ is the half-window length, and $\mathrm{med}(x_1, ..., x_M)$ refers to the median of a sequence of M numbers x_1 to x_M. Note that if $M = 1$, then no median filtering is applied.

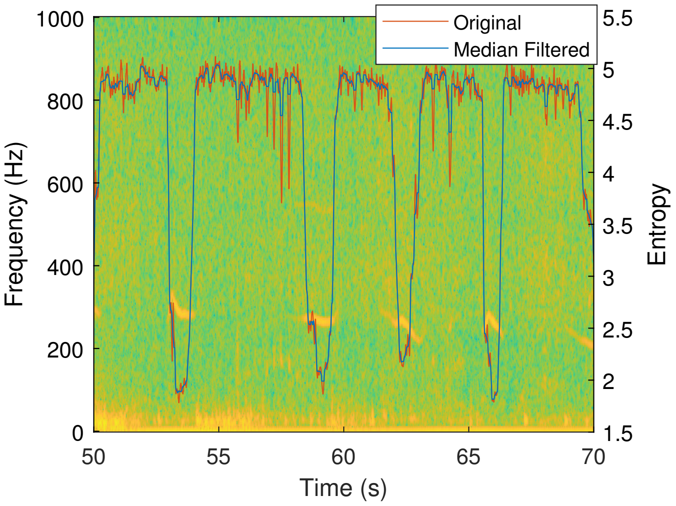

Figure 3.7: Median filtered entropy (STFT) with $M = 355$.

Figure 3.7 shows an example of applying median filtering ($M = 355$) on the STFT spectral entropy.

An example of the resulting class means applied to the entropy measure is shown for the STFT in figure 3.8.

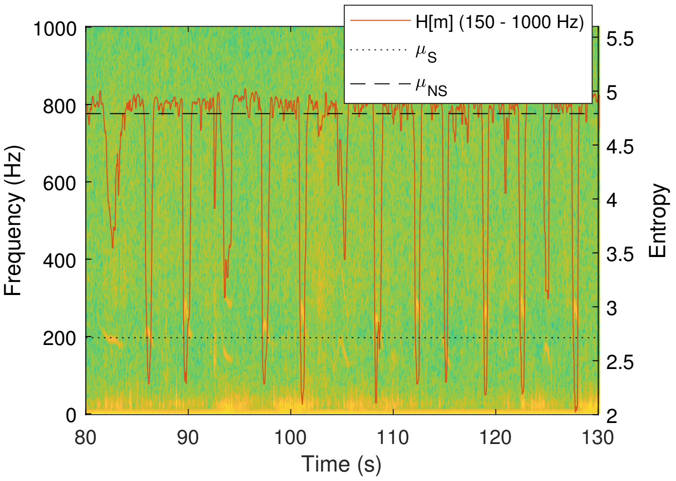

Figure 3.8: The resulting class means from the K-means algorithm on median filtered entropy with $M = 355$.

3.5 Simulation

3.5.1 Purpose

Measuring classification accuracy on a per-sample basis is difficult to do with practical data, since the exact start and endpoints of a signal cannot be known exactly. Furthermore, the instantaneous SNR of a whale vocalization must be estimated, and does not remain constant.

As such, it is preferable to corrupt known test signals with noise for sample-accurate measurements. An exact SNR also allows for detailed investigation on the effect of SNR on the algorithm, which grants insight into designing a detector for practical data with estimated parameters.

Simulation is used to determine the best TF-decomposition for the practical data with estimated instaneous SNRs and provide insight on how to choose the median filter window length.

3.5.2 Setup

A pulsed FM sweep ($x[n]$) from 150 - 800 Hz is mixed with a sample of noise ($w[n]$) collected in the same time frame as the humpback songs from False Bay. The SNR is calculated only with respect to relevant indices used for entropy calculation, which ranges from 150 - 1000 Hz. As such, the SNR described in this section is higher than what would

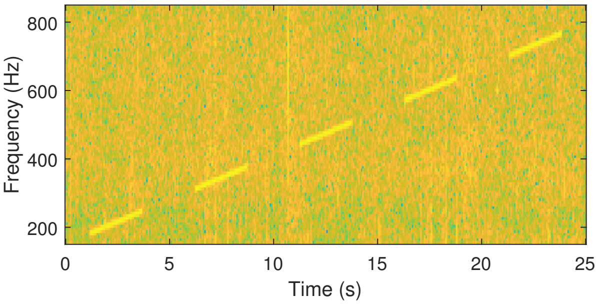

Figure 3.9: Spectrogram of the test signal (0 dB SNR).

be observed practically. This eliminates anchor noise, which contains most of the signal power, from the SNR calculation. The SNR is calculated using the segments where $x[n]$ is present.

For this simulation, $x[n]$ is non-zero in 10% of noise samples, pulsed evenly over a 25 second duration, sampled at 2 kHz. The STFT of the test signal is shown in figure 3.9.

All STFT's are performed with a Hamming window of length 256 with an overlap of 255. No zero-padding is applied for the FFTs. Figure 3.9 shows the STFT of the test signal mixed with the ocean noise with a SNR of 0 dB.

All CWTs are L1-normalized and are performed with a Morse wavelet with parameters $\gamma = 50$, $\beta = 2000$, over the same frequency range as the STFT (150 - 1000 Hz). 40 filters per octave are used, corresponding to a total of 110 band-pass filters, which is a comparable number of frequency bins used by the STFT entropy calculation (109 bins). The high time-bandwidth product ensures that the filters are narrow enough to reduce contamination between frequency bins. The hyper parameter g is set such that $p = 0.99$.

3.5.3 Measuring Soft-Output Accuracy

Given an observed signal $y[n] = w[n] + x[n]$ of length N, where $w[n]$ is noise and $x[n]$ is a sporadically appearing signal. The presence of $x[n]$ is encoded by the set $S_x = \{n_1, n_2, ..., n_L\}$, where $n_i \in \{0, 1, ..., N-1\}$ is the indices at which the signal is present. The cardinality (size) of the set is denoted as $|S_x|$.

The soft-output classification function $c_s(H[m])$ is used to determine the soft true positive

(STPR) and true negative (STNR) discovery rates:

$$\text{STPR} = \frac{1}{|S_x|} \sum_{n \in S_x} c_s(H[m])$$

$$\text{STNR} = \frac{1}{N - |S_x|} \sum_{n \notin S_x} 1 - c_s(H[m])$$

This is equivalent to calculating the conventional TPR and TNR discovery rates if $c_h(H[m])$ is used instead of $c_s(H[m])$.

3.5.4 Results

Figure 3.10 shows the STPR and STNR measures for various SNRs and median filter window lengths.

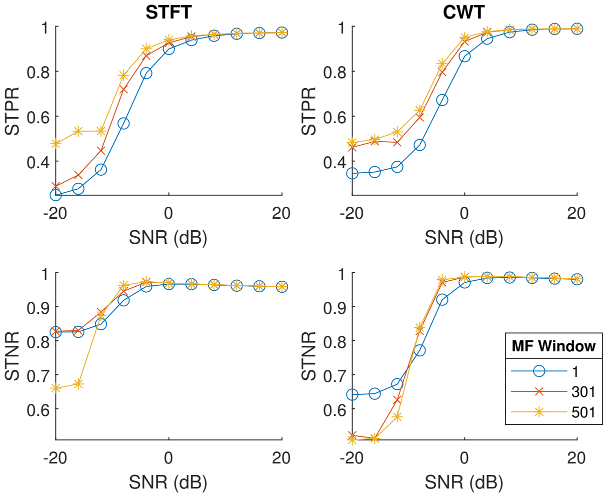

Figure 3.10: STFT and CWT soft discovery rates over a range of SNR's and median filter window lengths. The left-hand side of the figure shows the STPR and STNR of the STFT respectively. The CWT soft discovery rates are on the right-hand side.

Figure 3.11 shows the STPR and STNR measures for median filter window lengths for SNRs of -10, -5 and 0 dB. As shown in tables 3.1 and 3.2, the discovery rates for SNRs above 0 dB are greater than 90%, with median filtering only showing minor improvements of the measures. Median filtering is therefore critical for lower power signals, which may include the fringes of a whale call waveform, due to ramping in volume.

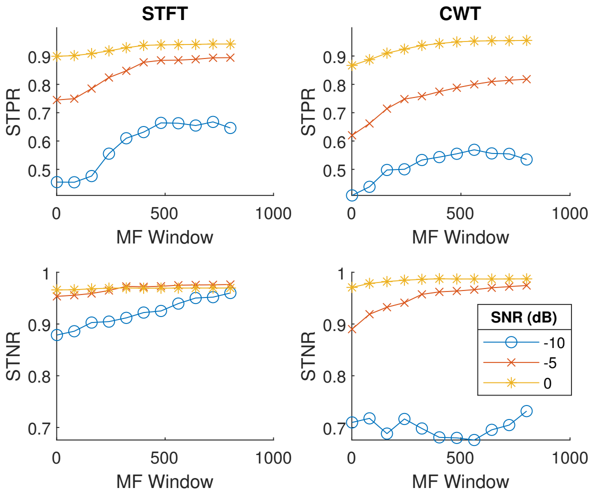

Figure 3.11: STFT and CWT soft discovery rates over a range median filter window lengths for low SNRs. The left and right sides of the figure show the results for the STFT and CWT respectively.

Median filtering increases the STPR for low SNRs. Below -10 dB, it has a negative effect on the STNR. This is acceptable, since most practical signals are expected to operate above -10 dB SNR.

Figure 3.12 shows how the class mean separation is affected by SNR and median filter window length for the STFT and L1-normalized CWT. Median filtering has no significant influence on class separation above 0 dB SNR, but negatively impacts the separation below 0 dB SNR. The STFT exhibits better class mean separation compared to the CWT. Despite this, tables 3.1 and 3.2 show superior CWT performance, due to the better time localization properties of the CWT.

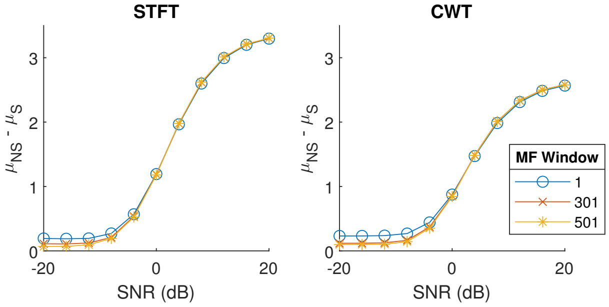

Figure 3.12: STFT and CWT difference in class means ($\mu_{NS} - \mu_S$) over a range of SNRs and median filter window lengths.

Table 3.1: Comparing the STPR of the STFT and CWT at varying SNR's and MF Window Lengths

		Soft True Positive Rate (%)				
	M	SNR				
		-10	-5	0	5	10
	1	45.5	74.5	89.9	94.5	96.3
STFT	301	59.3	84.3	92.7	95.7	96.6
	501	**66.4**	**88.4**	94.0	96.0	96.7
	1	40.8	62.0	86.7	95.5	98.1
CWT	301	52.4	75.0	93.2	97.8	98.7
	501	56.9	79.1	**95.1**	**98.0**	**98.7**

Table 3.2: Comparing the STNR of the STFT and CWT at varying SNR's and MF Window Lengths

		Soft True Negative Rate (%)				
	M	SNR				
		-10	-5	0	5	10
	1	87.9	95.3	96.6	96.5	96.2
STFT	301	91.2	97.2	96.9	96.6	96.3
	501	**92.9**	**97.4**	96.9	96.6	96.3
	1	71.0	89.0	97.1	98.5	98.5
CWT	301	70.8	95.6	98.7	98.8	98.7
	501	66.8	96.4	**98.7**	**98.8**	**98.7**

The CWT exhibits poorer STNR and STPR performance for SNRs below 0 dB. Utilizing the equivalent sinusoidal power from equation (3.6) unevenly weights some of the frequency bins according to filter bandwidth, which lowers the CWT's performance for low SNRs, since the noise power becomes significant compared to the signal power. This makes the L1-normalized CWT ineffective for low SNRs. The reduction in class mean separation compared to the STFT in figure 3.12 may also be explained by this phenomenon.

For SNRs below 0 dB, it is recommended to use the STFT with a large MF window ($M \in [500, 1000]$). Note that this will reduce the capability of detecting shorter signals. For SNRs above 0 dB, the CWT performs best, with a recommended window size greater than 300.

Median filtering has little to no significant effect above 10 dB SNR. If no median filtering is to be used ($M = 1$), the STFT is recommended for SNRs below 5 dB and the CWT for SNRs above 5 dB. For a change in sampling frequency, the MF window lengths may be adjusted accordingly.

3.5.5 A Note on L2 Normalization

Figure 3.13 compares the performance of the L1 and L2-normalized CWTs. The L2-CWT shows better STNR performance than the L1-CWT, since the power scaling effect is not present. However, the L2-CWT does not perform as well on STPR, due to the suppression of sinusoidal power at higher frequencies.

The the STFT outperforms the L2-CWT on all accounts, making the L1-CWT the superior option, since it performs better than the STFT for higher SNRs.

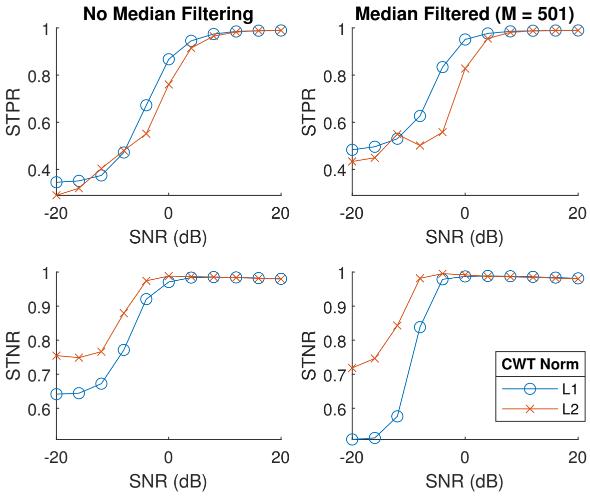

Figure 3.13: Comparing L1 and L2 CWT normalization for spectral entropy calculation. The left side of the figure shows the results without MF ($M = 1$). The right side of the figure shows with MF ($M = 501$). Note the large differences between L1 and L2 normalization for STNR and STPR.

3.6 Application to the False Bay Data

From simulation, it is determined that a larger median filter length and the L1-normalized CWT performs best for SNRs above 0 dB. As such, a median filter window lengths of 201 and 501 are tested for the proposed soft-classification technique, and compared to the baseline STFT-based SE (Hamming window of 256 samples; 255 sample overlap) used in [19]. No median filter is also tested on the CWT so that the STFT and CWT may be directly compared without enhancements. The detectors operate in the 130-1000 Hz frequency region, with the CWT having 40 filters per octave, $\gamma = 50$ and $\beta = 2000$.

The manual annotations are considered as the ground truth. However, it should be noted that these annotations may suffer from slight inaccuracies on a per-sample basis, which would reflect as a constant offset in the results, thus not contaminating this experiment significantly.

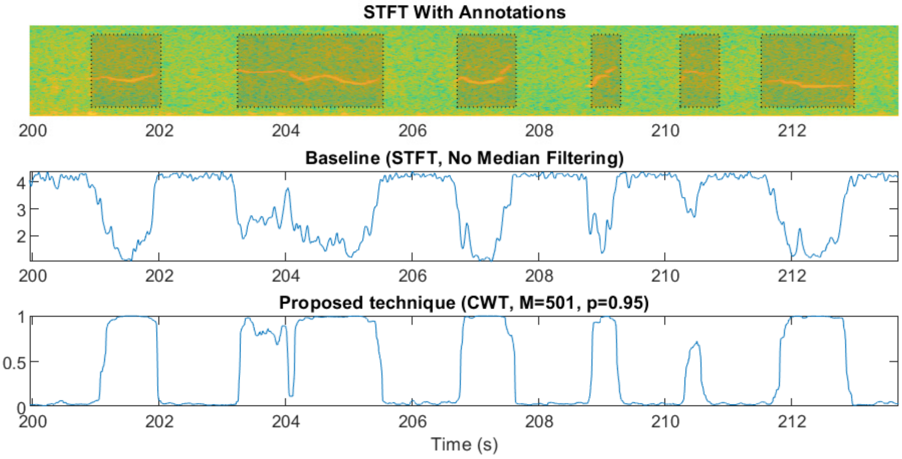

Figure 3.14: A snippet showing the measures produced by the entropy and the proposed soft-output signal detector.

A receiver operating characteristic (ROC) curve is used to show the performance of the proposed technique, as is done in [19]. This curve plots the true positive rate (TPR) and the false positive rate (FPR) on the same axes, determined by sweeping over a detection threshold parameter. For the baseline SE, this threshold is swept linearly between the maximum and minimum entropy values. The proposed technique sweeps the pseudo-probability threshold from 0 to 1. The sigmoid gain is set such that $p = 0.95$. Figure 3.14 shows the output the baseline SE and the proposed detector on a snippet of the recorded humpback song.

The ROC curve illustrates the trade-off between probability of detection (TPR) and probability of false alarms (FPR) for a varying detection threshold. Curves which are closer to the upper left corner of the ROC plot indicate greater detection capability, since it reduces the risk of a false detection while remaining accurate (true detections). Accuracy is measured on a per-sample basis, since accurate start and endpoints are important for many applications.

The TPR and FPR measures are calculated using:

$$\text{TPR} = \frac{\#\ \text{True Positives}}{\#\ \text{Signal Samples}};$$

$$\text{FPR} = \frac{\#\ \text{False Positives}}{\#\ \text{No-Signal Samples}}.$$

The ROC curve shown in figure 3.15 is constructed from the two recording segments containing the humpback songs. The thresholds which produce the ROC curve is shown in figure 3.16. For reference, the ROC curve of BLED is also shown, since it is widely used as a comparison in relevant literature [79, 19]. BLED is configured to calculate the energy in the 130-1000 Hz region via summation of the STFT produced for the baseline SE detector.

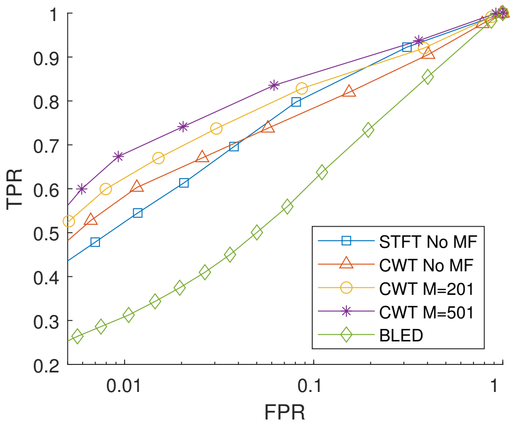

Figure 3.15: The ROC curves comparing the proposed technique and the baseline STFT SE. Median filter window lengths of 1, 201 and 501 are shown. The sigmoid gain is set such that $p = 0.95$. FPR is plotted on a log scale, which better illustrates the TPR differences for small FPR values.

From figure 3.15, it is clear that the proposed technique which uses a median filtering window of 501 samples and the CWT as a TF decomposition outperforms the conventional STFT-based SE. The proposed method shows an increase of around 10% in sensitivity (TPR) for FPRs below 3%, and remains a better detector overall. As shown previously [19], SE greatly outperforms a standard BLED detector.

In practice, the detector should be tuned such that the FPR is below 5%. From figure 3.16, this corresponds to a threshold of $c_s(H[m]) > 0.1$ for $p = 0.95$. This threshold will

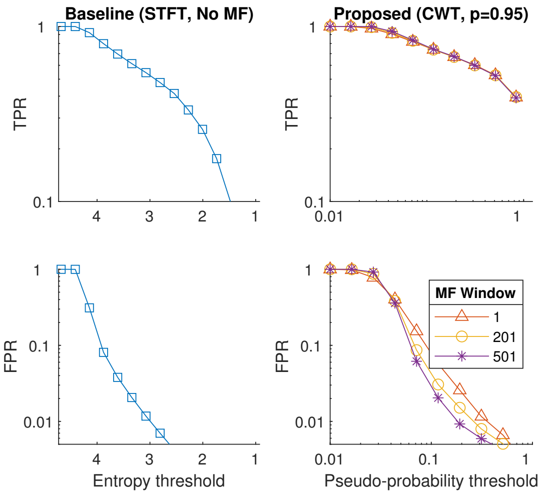

Figure 3.16: The thresholds used to produce the ROC curve and the TPR/FPR of the baseline SE and the proposed technique. Median filter window length of 1, 201 and 501 are shown on the right. The baseline entropy is shown on the left.

vary depending on the choice of p. Also note that the CWT SE without median filtering ($M = 1$) outperforms the STFT in the FPR $< 5\%$ region, indicating that it is the best TF-decomposition for similar practical applications.

3.7 Discussion

The proposed detector based on the CWT with median filtering outperforms the traditional SE detector as described in [19] for practical applications, verified by real and simulated experiments. The comparison method is done on a per-sample basis, which also indicates increased time localization capabilities of the proposed detector. This is to be expected, since the CWT improves time localization for higher frequencies as a trade-off for filter bandwidth, as opposed to the STFT which has a fixed time-localization as a function of window length [39]. Median filtering significantly decreases the FPR by suppressing fluctuations in the SE measure. Overall, the proposed detector indicates an approximate 10% increase of TPR for FPRs below 3%. As shown by [19], we have again demonstrated that SE detectors outperform BLED on all accounts.

It should be noted that the results shown by the ROC curve from figure 3.15 are lower compared to Erbe and King's results [19]. Our paper's method of measuring performance differs from much of classification literature, since it constrains the window overlap to produce a TF decomposition for every sample, whereas classification accuracy is usually measured on a per-window basis [75, 51, 52, 25, 74] with less overlap. For simple classification problems, accurate audio segmentation and time localization is not necessarily a priority. In [19], the window overlap is not maximal, thus detections are limited to the time localization accuracy determined by the overlap length. The obtained results are more similar to [27], which also measure accuracy on a per-sample basis.

This paper introduces the k-mean clustering algorithm as a method to provide a pseudo-probabilistic measure of signal presence. This idea may be adapted to any binary signal detector, such as BLED [79], to assist in providing a more intepretable measure of signal presence which does not depend on the magnitude of test statistic produced by the detector. The pseudo-probabilistic measure yields a threshold setting which may be more intuitive to non-experts, while also providing a value to be used to create adaptive algorithms which deal with sporadically appearing signals. Possible fields of application include denoising [26, 22], instantaneous frequency estimation [82] and classification (for example, as an additional input to a deep neural network) [25]. Additionally, the CWT may offer better computational efficiency compared to the STFT if per-sample signal segmentation accuracy is required [8], which could be relevant in preparing datasets for supervised learning.

The addition of the k-means algorithm presents another advantage: utilizing the class-mean separation ($\mu_{NS} - \mu_S$) as metric to quickly determine whether there are points of interest within an audio segment. Furthermore, the k-means algorithm may be restricted to converge within a set number of iterations. Whether the algorithm converged indicates if there are points of interest within an audio segment. By segmenting the long recording and testing for appropriate class mean separation and k-means algorithm convergence, the proposed technique provides a fast way of discarding large portions of noise in long audio recordings. This method may be robust to changes in noise conditions, as long as the changes in noise are not experienced in a given segment of a long audio recording. The use of such a system remains to be explored.

Automatic threshold setting is a possible extension of this study. This may be achieved by fitting a 2-component GMM to the SE distribution, with k-means clustering acting as the seed values of the fitting procedure, similar to HMM literature [75, 51, 52]. If a good approximation of the SE distribution is achieved, an appropriate threshold can be automatically determined for a fixed FPR specified by the user [77].

A SE-based detector may not be appropriate for detection in all whale species, since it is restricted to detecting vocalizations with periodic oscillations. This implies that some broadband impulsive noises, such as very short clicks, may not be detected. However, the detector will be insensitive to impulse noise disturbances, such as snapping shrimp [22, 79].

3.8 Conclusion

The performance of whale vocalization detection in the presence of noise based on spectral entropy has been evaluated in this paper. Through simulation and comparison with manual annotations, the effectiveness of the CWT as a TF-decomposition for SE calculation is demonstrated. The use of median filtering also indicates significant improvement of the conventional STFT-based SE detector. Through simulation, this paper shows that the L1-normalized CWT outperforms the STFT and L2-normalized CWT for practical SNRs.

The performance of the proposed detector shows an increase of approximately 5-10% in accuracy (TPR), for false alarm rates (FPR) below 5%, which is ideally what the detector should be tuned to. Furthermore, we propose a method of converting the SE entropy measure to a pseudo-probability using k-means clustering and automatic scaling, which is both more interpretable to non-experts while providing a foundation for developing adaptive algorithms for sporadically appearing cetacean vocalizations, which must take signal presence into account.

3.9 Acknowledgements

This research is funded by the National Research Foundation (NRF) of South Africa (grant numbers: MND210609609887; 129224). The findings of this study and the opinions reflected therein are those of the authors and do not necessarily represent the NRF.

Appendices

3.A K-means Implementation

For a known signal and noise profile, such as white noise, it is possible to estimate the expected values of the spectral entropy measure $H[m]$ given signal/no-signal conditions. However, in a real-world setting, the noise profile cannot be assumed to be a stationary process such as Gaussian white noise (GWN) or colored white noise (CWN). The variability of AMFM signals also further complicates this estimation. Setting appropriate thresholds on the entropy measure have to be done either manually, or via an automatic process.

K-means is used to estimate these expected values in real-world conditions. A spectral entropy measurement $H[m]$ may belong to 2 classes: no signal (C_{NS}); signal (C_S). Note that it is required to have signal segments from both classes present in $H[m]$, otherwise this process will fail. It is assumed that the data satisfies the condition $\mathbb{E}\{H[m]|C_S\} < \mathbb{E}\{H[m]|C_{NS}\}$, where $\mathbb{E}$ denotes the expected value, implying that signal presence will always lower the entropy measure.

Class-membership is assigned using a hard class-assignment function:

$$w[m] = \begin{cases} 1, & H[m] \in C_S \\ 0, & H[m] \in C_{NS} \end{cases}.$$

Given the class means μ_S, μ_{NS} and the entropy measure $H[m]$ of length N, the assignment function assigns the class with the closest mean:

$$w[m] = \begin{cases} 1, & (H[m] - \mu_S)^2 \leq (H[m] - \mu_{NS})^2 \\ 0, & \text{otherwise} \end{cases}. \tag{3.8}$$

The K-means algorithm for this application is shown below:

1. Assign initial means:

$$\mu_S = \min_m (H[m]);$$

$$\mu_{NS} = \max_m (H[m]).$$

2. Perform class assignment using equation (3.8).

3. Update the means according to:

$$\mu_S = \frac{\sum_{m=0}^{N-1} w[m]H[m]}{\sum_{m=0}^{N-1} w[m]};$$

$$\mu_{NS} = \frac{\sum_{m=0}^{N-1}(1 - w[m])H[m]}{\sum_{m=0}^{N-1} 1 - w[m]}.$$

4. Repeat steps 2 and 3 until the maximum change in means converges to a specified value ϵ. Terminate if it exceeds a predefined number of iterations.

Step 1 guarantees the means converge to their correct classes, and do not swap around in the process. It assumes that the maximum and minimum are not too extreme in that only a single point is associated with the initialization, and that signal/no-signal segments exist. The difference in class means may be used to determine whether a signal is present in the audio clip at all, since the class means will converge to similar values if no signal segments exist.

4 Feature Extraction with Wavelet Transforms

TF decompositions are often used to extract features for signals. In audio, MFCCs are features calculated from the STFT of the signal. A “TF” decomposition, as referred to in this dissertation, does not necesarrily require a time variable: spatial variables are often used. In this sense, a convolutional neural network (CNN) is can also be considered as a “TF” decomposition, where multi-dimensional index variables (or index vectors) are used for “time” and frequency:

$$X[\mathbf{k},\mathbf{m}] = (x * \upsilon_{\mathbf{k}})_{\downarrow\mathbf{d}}\,,\ \upsilon_{\mathbf{k}} \in \Upsilon. \tag{4.1}$$

In this case, we abuse notation to indicate that the operator $(\cdot)_{\downarrow\mathbf{d}}$ downsamples by different amounts for each dimension of $\mathbf{m}$.

The downsampling factor $\mathbf{d}$ limits the bandwidth of X, which, as discussed in chapter 2, is selected so as to retain all information carried by the filters in Υ. Although not notated, different filters (different $\mathbf{k}$) may have different values of $\mathbf{d}$. Notation to support such a case is neglected for clarity, but is later expanded upon in chapter 7.

4.1 The Importance of Invariance

The decomposition X is often further processed by additional filters $\Phi_i[\mathbf{k},\mathbf{m}]$, where i is the index of the post-processing filter:

$$\tilde{X}_i[\mathbf{k},\mathbf{m}] = X * \Phi_i. \tag{4.2}$$

The filters Φ_i are not necesarrily multi-dimensional, and may only operate along a single dimension of X. If Φ_i is an averaging (low-pass) filter, some additional stability can be observed in $\tilde{X}_i$ compared with X. This allows for the introduction of invariances to various deformations applied to X, which can be greatly beneficial for some applications. In a sense, cascased structures of X and $\tilde{X}_i$ are defined in a CNN for which the network learns the required invariances.

It is widely known that enforcing invariance in feature extraction and ML pipelines increase performance. Enforcing invariance in NNs has been shown to improve the network, while also improving the interpretability of the network. For example, when classifying images, invariance to scale, shear, rotation and translation deformations, depending on image properties, may be useful when extracting features. Specifically, deformations of a specified order should approximately map to the same point in feature space:

$$\Gamma\left(\delta\left(x\left[\mathbf{n}\right]\right)\right) \approx \Gamma\left(x\left[\mathbf{n}\right]\right), \tag{4.3}$$

where Γ is the feature extraction operator, and δ is the deformation operator applied to x.

Invariance filters (Φ_i) can also have additional properties that makes it favourable for audio applications. For example, many applications “blur” (low-pass) the STFT magnitude or power spectrum, which can reduce the impact of noise fluctuations, and can also serve to suppress small frequency fluctuations across adjascent frequency bins (if the blurring kernel Φ_i filters across frequency as well). The blurring operation can improve the performance of signal detectors and classifiers. However, as a cost, time localisation and/or frequency localisation is reduced.

A new problem is introduced when enforcing invariances via the Φ_i filters - information is lost via averaging. Various techniques can be used to further retain information, while also retaining invariance. Wavelet scattering is such a method. In a sense, multilayer CNNs topologies with skip connections can be though of as a method to recover lost information in the deeper layers.

4.2 MFCCs Reframed as Shift-Invariant Features

Since MFCCs are the most widely used feature extraction method for audio signals, save for CNN frontends which also act as a feature extractor when operating on the STFT. For this reframing, we ignore the discrete cosine transform (DCT) applied to the filtered values, since this linear transformation only serves to decorrelate the filter coefficients.

The magnitude spectrum used in MFCC calculation is expressed as a TF decomposition as defined by equations (2.2) and (2.6):

$$X[l,m] = |\upsilon_l * x|_{\downarrow d}, \tag{4.4}$$

where υ_l is the l'th STFT filter.

The Mel spectrogram TF decomposition X_{Mel} is then calculated as

$$X_{\mathrm{Mel}}[k,m] = \log\left(X[l,m] * \Phi_k[l]\right)\Big|_{l=0}, \tag{4.5}$$

where Φ_k is the k'th MFFC triangular filter operating along the STFT index l. The evaluation at $l = 0$ indicates that this operation is a multiplication and summation only.

We can view this process as a TF decompistion with STFT filters $\upsilon_l \in \Upsilon$, which is modified with the Mel-scale triangular filters Φ_k. υ_l provides time-shift invariance due to the STFT window, whereas Φ_k selectively averages frequency content of the STFT, thereby providing some frequency-shift invariance.

4.3 Wavelets as a MFCC Generalisation

We can construct a wavelet filterbank which, similar to MFCCs, introduces time and frequency-shift invariance in a similar manner to the Mel spectrogram. This process is constructed in the opposite order (frequency-shift, then time-shift invariance) to the Mel spectrogram. However, its properties remain similar [7].

We construct wavelet band-pass filters $\upsilon_k \in \Upsilon$ which is spaced in frequency such that the desired invariance properties is obtained (as discussed in section 2.3). We then utilise a single low-pass filter $\Phi[\mathbf{m}]$ which then provides invariance in the time and/or spatial dimensions:

$$X[\mathbf{k},\mathbf{m}] = \left|x * \upsilon_{\mathbf{k}}\right|_{\downarrow\mathbf{d}},\ \upsilon_{\mathbf{k}} \in \Upsilon; \tag{4.6}$$

$$\tilde{X}[\mathbf{k},\mathbf{m}] = X[\mathbf{k},\mathbf{m}] * \Phi[\mathbf{m}] \tag{4.7}$$

The notation in equations (4.6) and (4.7) conforms the general TF description of chapter 2, although the specific wavelet scattering operators in chapters 5 to 7 follow a different standard. These operators are known as wavelet scattering, which formalise the notation of "averaging a TF-decomposition" in terms of digital signal processing. Chapters 5 to 7 discuss the specifics, flavours and implementation of scattering operators at length.

5 Introducing Wavelet Scattering

M.W. Rademan, D.J.J Verseld, J.A. Du Preez

Accepted in the Journal of the Acoustical Society of America (JASA) on 15 May 2024

doi: doi

Abstract

Current literature trends for whale classification in passive acoustic monitoring systems are mainly concerned with high accuracies obtained via deep learning approaches. This requires preparation of large and accurate annotated datasets, which are often not available when surveying new regions, nor are they wholly accurate when they are. In contrast to classification, many efforts are still being made in improving generalised detectors for the purpose of disregarding noise. As such, there is a disparity between detectors and classification models.

This paper aims to bridge the gap between detection and classification, while exploring the application of wavelet scattering features not yet popularised in the field of passive acoustic monitoring. This paper improves on the classical spectral entropy detector regarding noise adaptivity and uses wavelet scattering for entropy calculation. The same scattering decomposition is further fed to a classification stage: a linear classifier using scattering features provided by windows identified by the entropy detector.

The proposed method works well for small datasets and aims to form the foundation of wavelet scattering classification/detection using non-neural network approaches. The purpose of this work is to provide a practical classification and detection system for new regions in which very little annotated data is available.

5.1 Introduction

Underwater passive acoustic monitoring (PAM) systems are an effective non-invasive tool to monitor marine activity, with many research endeavours to automate cetacean monitoring and detection over the last decades. ML is a particularly prevalent approach to detecting and classifying cetacean calls from PAM audio [64], although some recent publications remain focused on traditional approaches that seem to oppose ML trends [79, 80]. The disparity between the approaches are often due to a lack of data or a high uncertainty of the data quality [45], which are detrimental to NN and deep learning models.

Although the dataset and the species of interest for classification is not the focus of this paper, we apply the proposed method to a dataset containing annotated Antarctic blue whales (*Balaenoptera musculus intermedia*) and fin whale calls (*B. physalus*). Blue and fin whales are considered globally endangered species of whale following whaling activities in the 1900s [11, 10]. This launched an effort of gathering many hours of PAM data in the southern Antarctic ocean [76] for the purpose of evaluating detectors and classifiers, which has been made publicly available [46].

Large annotated datasets may not necessarily be accessible to automate detection and classification in conservation studies, which motivates the further improvement of “white-box" strategies as opposed to using NNs. For point(s) of interest (POI) detection purposes, BLED, the SE detector and the spectrogram correlator [46] are mostly used in practise, often with software such as PAMGuard [55]. These detectors require the user to choose many different hyperparameters for the purpose of adaptivity to noise conditions and signal level. They do not necessarily classify signals, and instead pinpoint POIs, which suffer from many false positive detections. On the other hand, very accurate NN approaches exist for a variety of species [36, 48, 81, 63], but require large datasets and careful data preparation.

Non-NN approaches have also obtained very accurate results, most of which utilise MFCCs as features [73, 51]. A hidden Markov model (HMM) is deployed in the aforementioned studies, thereby allowing the modelling of time-dependent sound characteristics without requiring a NN.

At the center of the recent work revolving around PAM classification and detection lies a form of TF-decomposition, whether it be for feature extraction or improved detection. Both MFCC features and sound detectors (BLED, SE) typically use the STFT as a base for TF-decomposition. However, CWTs may be be used a drop-in replacement for the

STFT, as is shown by the recent work of Rademan *et al.* [60]. Wavelet scattering (WS) is an extended form of the CWT and may be used to extract features that provide more information than MFCCs, while also sharing many similarities with the structure of CNNs [7]. In fact, WS may be considered as a manual definition of a CNN front-end, while the time-decimated CWT amplitude coefficients (equivalent to first level scattering) mimic MFCC features.

WS has not yet been well established in the field of PAM, but has seen use in many medical [35, 49], fault diagnosis [29] and other audio [62, 69, 6] applications. A recent study by Michau et. al. has shown exceptionally accurate results in birdsong classification via WS features with a support vector machine (SVM) classifier [44].

Automated classification is often the end goal of many PAM systems, but identifying POIs without classification is a necessary step to assist in the annotation and sound exploration process, especially when surveying new and unknown oceanic regions. As such, generalised detection is often performed with BLEDs and SE detectors [59]. SE detectors have consistently shown better performance than energy detectors [19]. The previous approach by Rademan *et al.* [60] modifies SE detectors to stabilize the entropy measure and provide pseudo-probabilistic output, thereby improving endpoint detection accuracy and the interpretability of the detector's output. SE detectors naturally motivate extension by including feature extraction and classification using the same TF-decomposition used for SE calculation, ultimately reducing computational resources.

A closely related study performed indicates state-of-the-art (SOTA) results in blue and fin whale call detection within given audio segments using a long-short-term memory (LSTM) NN with scattering features [65]. Our work differs vastly from that of Sattar as it aims to incorporate detectors as well as a non-NN model for use in small datasets. Our work classifies on a per-call basis, which is more comparable to other ML techniques. We evaluate different blue whale calls and fin whale calls separately and our method distinguishes itself by focusing on the applicability to very small datasets. We evaluate the proposed method in the context of a fully annotated dataset and show it is effective for a variety of call types. Our evaluation method does not include dataset cleaning, thereby allowing for results that reflect on the full spectrum of signal-to-noise ratio (SNR), annotation errors, very large noise variations (> 60 dB) and various vocalisation types. Although we do not compensate for analyst errors as done by Miller *et al.* [45], the accuracy of the "ground truth" labels is discussed at length in this study.

In this paper, we introduce a method which combines an improved traditional SE detector with WS to train a linear classifier. This paper aims to introduce WS to the world of PAM classification and detection, as it is a feature extraction method which has not permeated into this field as of yet. Since WS is a very flexible TF decomposition, it may serve as input to a variety of traditional detectors, while also providing advantages in the visualisation of waveforms compared to the STFT. We show that the proposed method is effective for very small datasets (10s of samples) and apply simple chi-square feature selection [72] to further improve results for smaller datasets. Refer to Figure 5.1 for a visualisation of the processes employed in this study.

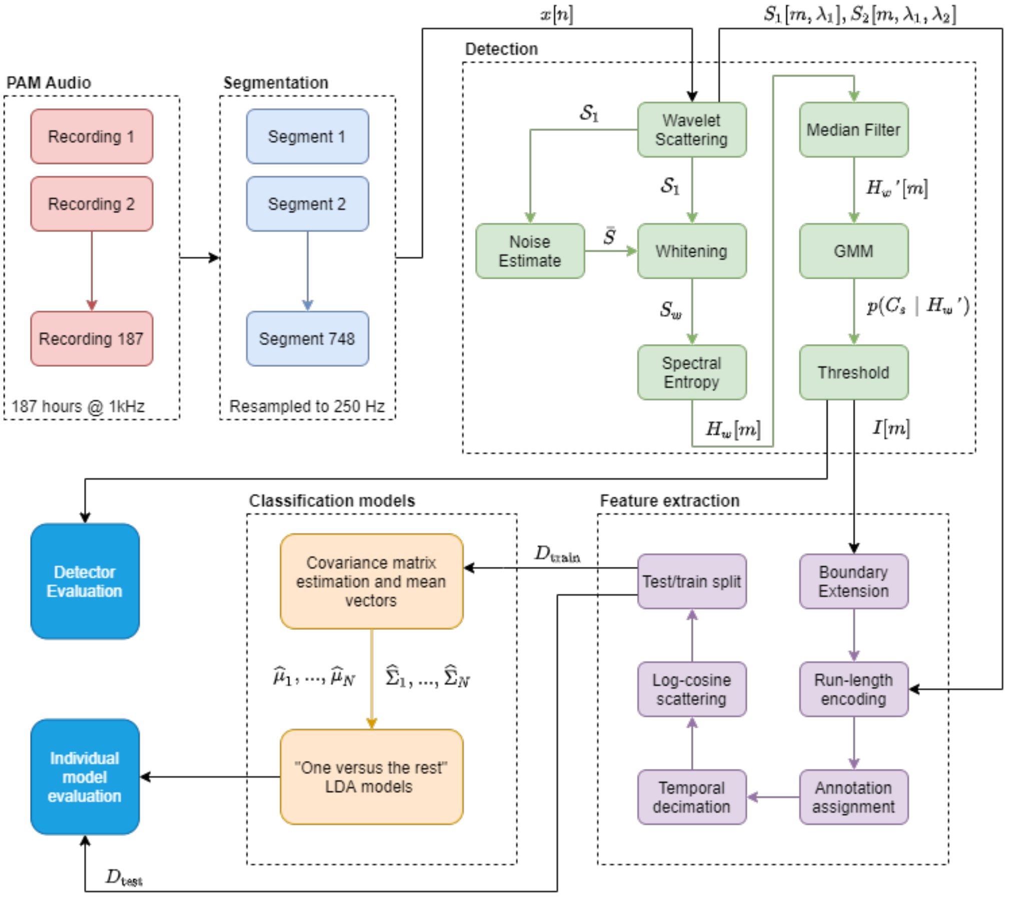

Figure 5.1: The detection and classifying system. The audio database constists of 1-hour long recordings, which are segmented for computational efficiency. The detection system provides input to the feature extraction and classification steps. The classifier and detector are evaluated separately in this study.

5.2 Data

In this paper, we use an open source dataset recorded over the period of one year in the Casey Islands of Antarctica in 2017, as detailed in [46]. This dataset contains 187 audio files of approximately 1 hour duration each. Recordings took place over the entire year. The hour-long audio recordings are spaced at intervals of approximately 42 hours apart, sampled at 1 kHz. This dataset was chosen instead of the other datasets presented in [46] due to its reliability, abundance of blue whale annotations and annotation end-point accuracy. Other datasets contained many annotations which were not significantly above the noise level, have inaccurate time localisation regarding the starting and endpoints of the annotation, and/or did not contain enough annotations.

We do not perform any annotation cleaning, as this will unfairly bias our results towards high SNR detection and classification. However, inaccurate analyst annotations are a significant concern [45], where some studies discard annotations based on estimated SNR to improve the “ground truth" used for accuracy measurements [66]. Analyst annotation inaccuracies are discussed in greater detail in the later sections of this paper.

For this study, we split each hour-long recording into segments of 2^{18} samples after re-sampling to 250 Hz, with the end segment often having less samples. Each segment is approximately 17 minutes in length. Segmentation is performed for memory conservation during computation.

Annotations are provided that give the time-frequency bounding boxes of blue and fin whale vocalisations, among others. In this paper, we mainly focus on the blue whale vocalisations as this is the class with the greatest abundance in this dataset. We also include fin whale vocalisations during detector evaluation. However, due to low SNR and annotation overlap, very few fin whale samples reach the classification stage. The Casey Islands dataset contains a total of 2988 blue whale annotations, separated as *Bm-A*, *Bm-B*, *Bm-Z* and *Bm-D* types. Figure 5.2 shows a 1-st level WS decomposition ($\mathcal{S}_1$) of high SNR examples of these vocalisation types. Table 5.1 summarises the number of annotations within this dataset.

We estimate the number of false positives by investigating a subset of annotations (at least 100 when possible) and determining which are more probable to be noise than a vocalisation according to their $\mathcal{S}_1$ coefficients as portrayed by Figures 5.2 and 5.5. This task is difficult to perform with absolute precision, so we provide a lower and upper bound for these estimates such that they are a fair reflection of the uncertainty. The number

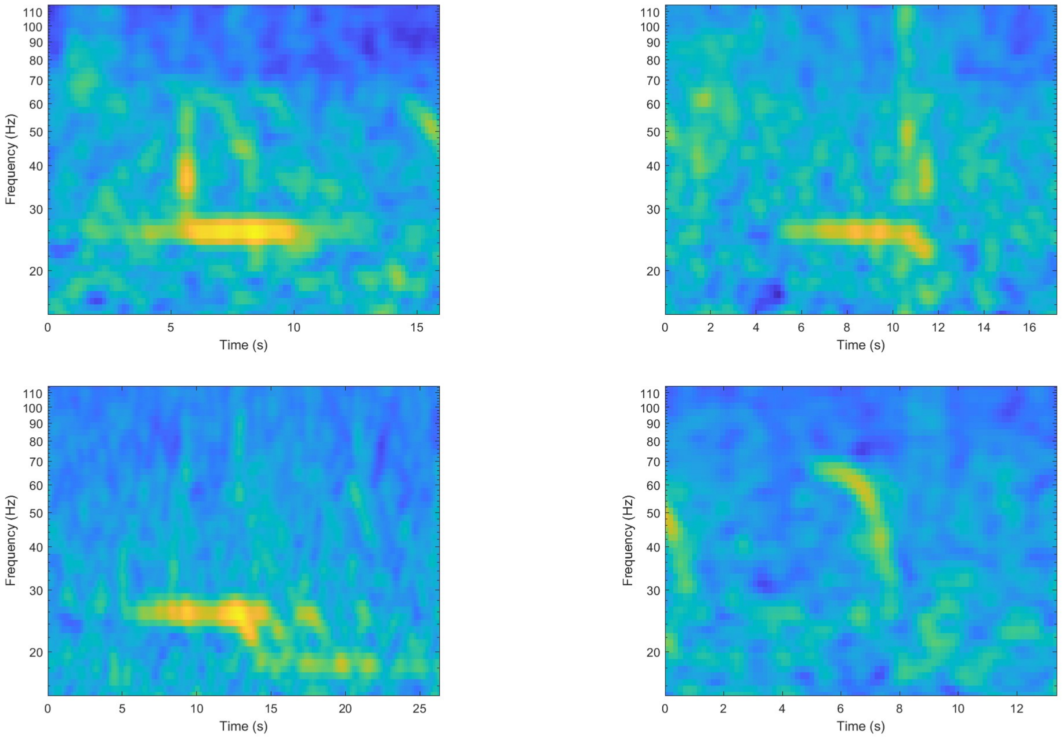

Figure 5.2: High SNR examples of the first level scattering transform to illustrate the characteristics of blue whale calls. Note the logarithmic frequency scale. *Bm_ Ant_ A* (a) is a constant frequency pulse which occurs around 20 Hz. *Bm_ Ant_ B* (b) and *Bm_ Ant_ Z* (c) are similar to *Bm_ Ant_ A*, but with the addition of frequency modulated tails. *Bm_ D* (d) is a frequency modulated downsweep which typically starts around the 80 Hz and sweeps to 30 Hz.

Table 5.1: Total number of annotations in the Casey 2017 dataset.

Vocalisation	**Number of annotations**	**Estimated False positives (%)**
Bm_Ant_A	1741	5 - 10
Bm_Ant_B	558	5 - 10
Bm_Ant_Z	119	1 - 2
Bm_D	553	8 - 15
Bp_20Hz	78	1 - 3
Bp_20Plus	214	1 - 3
Total	3263	5 - 10

of missed detections require more investigation and expertise and is out of scope for this study. The false positive proportions are indicated in Table 5.1, with the total proportion of false positives calculated by considering the proportion of annotations per class.

Table 5.2: Estimated erroneous (false positives) proportion of annotations per class that could be considered noise. This table does not reflect on whether annotations are of the correct class (i.e., misclassifications). The total proportion of false positives are calculated according to the number of annotations per class.

Vocalisation	**False positives (%)**
Bm_Ant_A	5 - 10
Bm_Ant_B	5 - 10
Bm_Ant_Z	1 - 2
Bm_D	8 - 15
Bp_20Hz	1 - 3
Bp_20Plus	1 - 3
Total	5 - 10

5.3 Wavelet Scattering

5.3.1 Filter Bank Construction

This paper utilises a modified version of the wavelet scattering transform, as detailed by Andén and Mallat [7]. Given a mother wavelet $\psi(t)$ with energy concentrated at 1 rad/s, the wavelet dilated by a factor λ is notated by

$$\psi_\lambda(t) = \lambda\psi(\lambda t). \tag{5.1}$$

The multiplicative scaling of λ ensures that dilation results in constant peak Fourier amplitude in the frequency domain, therefore resulting in L1-normalisation. When ψ is centered at 1 rad/s, then λ is also the center frequency of the filter (in rad/s) specified by ψ.

A typical choice of ψ, also the choice used in this paper, is the Morlet wavelet, given by

$$\psi(t) = (e^{jt} - \beta)\theta_\sigma(t), \tag{5.2}$$

with σ indicating the time support of ψ and j is the imaginary unit. The shifting factor β ensures that $\psi(t)$ satisfies the admissibility criterion (zero-mean). A Gaussian with zero mean and standard deviation σ is denoted by $\theta_\sigma(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}$, thus resulting in $\beta = \frac{\theta_{1/\sigma}(-1)}{\theta_{1/\sigma}(0)}$.

A filter bank is constructed with Q wavelets per octave. To ensure adequate overlap with adjacent filters, we set $\sigma = Q$. The set of dilation factors Λ in the filter bank is constructed with $\lambda_{n+1} = 2^{\frac{1}{Q}}\lambda_n$, where λ_0 sets the first center frequency in the filter bank, and $\lambda_k \leq \omega_{\max}$ is the last filter specified by the maximum frequency $\omega_{\max}$.

Discretised filter impulse responses are constructed using

$$\psi_\lambda[n] = \begin{cases} \psi_\lambda\left(\frac{n}{f_s}\right), \ n \in \{-N, ..., N\} \\ 0, \ \text{otherwise} \end{cases}, \ \lambda \in \Lambda. \tag{5.3}$$

$N = \left\lceil \frac{5Q}{\lambda_0} \right\rceil$ provides an adequate coverage of the lowest-frequency wavelet – 5 standard deviations of $\psi_{\lambda_0}[n]$.

The operator $\mathcal{U}$ is used to define the scalogram of a discrete signal $x[n]$ sampled at a frequency of f_s:

$$\mathcal{U}x[n,\lambda] = |x[n] * \psi_\lambda[n]|\,, \;\; \lambda \in \Lambda. \tag{5.4}$$

The scalogram extracts the band-limited amplitudes of each frequency band in Λ. Since the modulus operator $|\cdot|$ is effectively amplitude demodulating a band-limited analytic signal, the maximum bandwidth resulting from $\mathcal{U}$ is the bandwidth of $\psi_{\lambda_{\max}}$. We define this bandwidth as 2 standard deviations of $\Psi_{\lambda_{\max}}(\omega)$, corresponding to $2\frac{\lambda_{\max}}{Q}$ rad/s. This allows the critical down-sampling of $\mathcal{U}x$ by a factor of $d = \left\lfloor \frac{\pi f_s Q}{2\lambda_{\max}} \right\rfloor$. We denote the down-sampled scalogram as $\mathcal{U}_{\downarrow d}x$.

Additional to the scalogram, a further filtering and decimation step is performed to obtain the scattering coefficients. This is accomplished by utilising a Gaussian low-pass filter $\varphi(t)$ with a user-specified standard deviation of $\frac{T}{2\pi}$ seconds (cutoff frequency of $\frac{1}{T}$ Hz). T is referred to as the invariance scale. The scattering operator $\mathcal{S}$ performs filtering of the scalogram:

$$\mathcal{S}x[n,\lambda] = \mathcal{U}x[n,\lambda] * \varphi[n], \;\; \lambda \in \Lambda, \tag{5.5}$$

For computational efficiency, the critically down-sampled scattering coefficients are instead calculated as

$$\mathcal{S}_{\downarrow r}x[n,\lambda] = \left(\mathcal{U}_{\downarrow d}x[n,\lambda] * \varphi_{\downarrow d}[n]\right)_{\downarrow r}\,, \;\; \lambda \in \Lambda, \tag{5.6}$$

with $r = \left\lfloor \frac{f_s}{4TdV} \right\rfloor$ and V indicating the oversampling factor, since T further band-limits the content when chosen such that it is smaller than the time support of $\psi_{\lambda_{\max}}$.

The discretisation of $\varphi(t)$ is identical to that of $\psi_\lambda(t)$. The down-sampling factor r down-samples to within 2 standard deviations of $\Phi(\omega)$, as is done with ψ. The resulting sampling frequency of the scattering coefficients is approximately $\frac{4}{T}$, which may not be exact due to the compounding down-sampling steps of r and d.

To implement the filter bank as defined by Λ, we use the MATLAB Parallel Processing toolbox, with all convolutions carried out as FFT convolutions on the graphics processing unit (GPU). We pre-compute the FFTs of the filter bank and the low-pass filter for fast execution, thereby only requiring a FFT on the input signal. This limits the input signal to a specified length, set to the segment size of L. Shorter signals are zero-padded and stripped after convolution.

5.3.2 Scattering Transform

A scattering transform is performed by F scattering filter banks defined by $\{\Lambda_1, ..., \Lambda_F\}$. A scattering transform to the i'th level is denoted by the operator $\mathcal{S}_i$. We suppress notation indicating time dependence for clarity. The scattering transform performed on a signal x is given as follows:

$$\mathcal{S}_0 x(t) = x * \varphi \tag{5.7}$$

$$\mathcal{S}_1 x(t, \lambda_1) = |x * \psi_{\lambda_1}| * \varphi,\ \lambda_1 \in \Lambda_1 \tag{5.8}$$

$$\mathcal{S}_2 x(t, \lambda_1, \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \varphi,\ (\lambda_1, \lambda_2) \in \Lambda_1 \times \Lambda_2 \tag{5.9}$$

$$\mathcal{S}_i x(t, \lambda_1, ..., \lambda_i) = ||...|x * \psi_{\lambda_1}| * ...| * \psi_{\lambda_i}| * \varphi,\ (\lambda_1, ...,\ \lambda_i) \in \Lambda_1 \times ... \times \Lambda_i \tag{5.10}$$

Each scattering filter bank defined by Λ_i is constructed by the user. In this paper, we neglect the 0'th level scattering transform $\mathcal{S}_0$ since signals of interest are contained in the 20-120 Hz band, thereby not requiring any low-frequency information. We additionally only utilise up to the second level scattering $\mathcal{S}_2$ for feature extraction. The down-sampling steps detailed by equation (5.6) are only performed by $\mathcal{S}_1$.

The scattering computations used in this study are given by

$$\mathcal{S}_1 x[n, \lambda_1] = (\mathcal{U}_{\downarrow d} x * \varphi_{\downarrow d})_{\downarrow r},\ \lambda_1 \in \Lambda_1; \tag{5.11}$$

$$\mathcal{S}_2 x[n, \lambda_1, \lambda_2] = \left(\left|\mathcal{U}_{\downarrow d} x * (\psi_{\lambda_2})_{\downarrow d}\right| * \varphi_{\downarrow d}\right)_{\downarrow r},\ (\lambda_1, \lambda_2) \in \Lambda_1 \times \Lambda_2. \tag{5.12}$$

Refer to Figure 5.3 for a visualisation of the successive filtering steps to perform a scattering computation.

For computational efficiency, it is not required to to calculate all paths for λ_2, since the lower frequency filters defined by Λ_1 have small bandwidths, implying their demodulated coefficients may not have energy contained in all the paths defined by Λ_2. In this study, we only compute the $\mathcal{S}_2$ scattering coefficients when

$$\lambda_2 - \frac{\lambda_2}{Q_2} < \frac{\lambda_1}{Q_1}, \tag{5.13}$$

where Q_1 and Q_2 are the number of filters per octave for each filter bank defined by Λ_1 and Λ_2. A similar optimisation is also performed by Andén and Mallat [6]. The filter-banks used in this paper is shown in Figure 5.4.

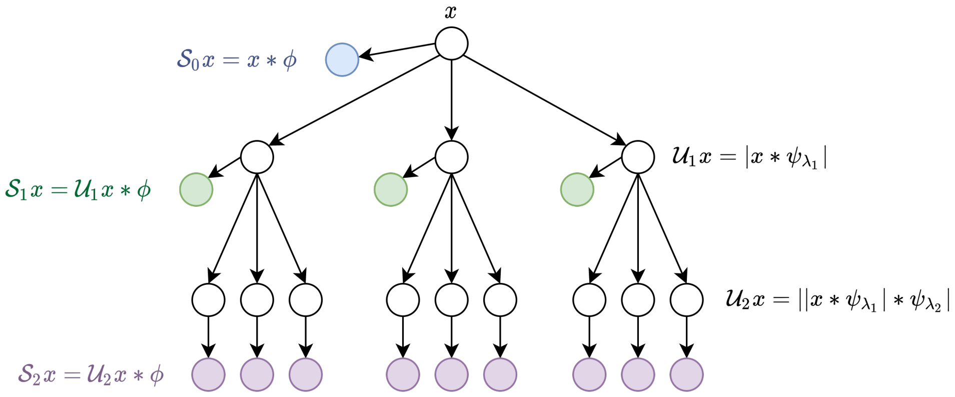
x
$\mathcal{S}_0 x = x * \phi$
$\mathcal{S}_1 x = \mathcal{U}_1 x * \phi$
$\mathcal{U}_1 x = |x * \psi_{\lambda_1}|$
$\mathcal{U}_2 x = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}|$
$\mathcal{S}_2 x = \mathcal{U}_2 x * \phi$

Figure 5.3: An illustration of the successive filtering operations performed by a scattering transform. Nodes in colour represent the extracted scattering coefficients, which is downsampled according to an invariance scale specified by the user.

The second filter banks in equations (5.11, 5.12) is sampled at a rate of $f_{ss} = \frac{f_s}{d}$. This implementation differs from the fast computation suggested by Andén and Mallat [6]. Typically, lower frequency filters are arranged so that the maximum time support does not exceed the time support of φ. Therefore, lower frequency filters are linearly spaced, whereas higher frequency filters are exponentially spaced when their time support become smaller than T. In the implementation for this paper, all filters are exponentially spaced, regardless of their time support exceeding φ. The resulting information leakage is largely irrelevant in the classification step for this paper, as later detailed in Section 5.5. For models dependent on time information, such as HMMs or LSTM networks, information leakage may affect results and the restriction of the time support of ψ specified by Andén and Mallat [6] should rather be used.

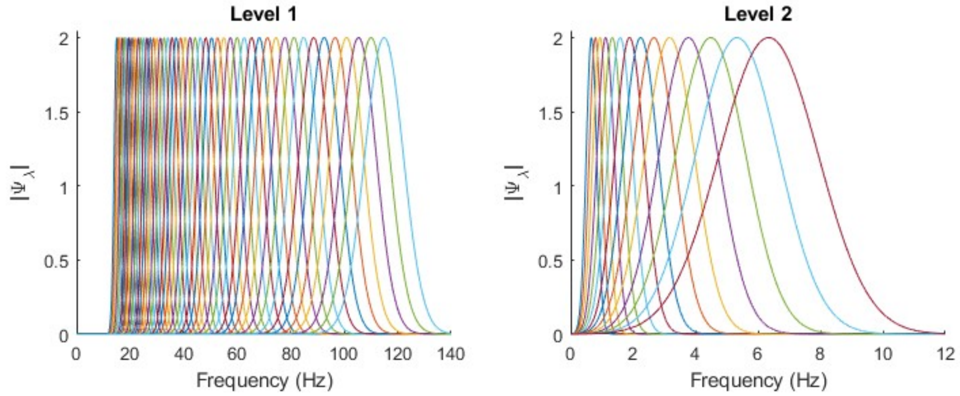

Figure 5.4: The filter banks used in this study, with $Q_1 = 16$ and $Q_2 = 4$. The filter bank used to calculate $\mathcal{U}_1$ is shown on the left and the $\mathcal{U}_2$ filter bank on the right. Note that the maximum magnitude of 2 accounts for the analytic nature of the wavelet, thereby capturing the true frequency amplitudes in a signal.

5.4 Improved Spectral Entropy Detector

The SE detector described by Rademan *et al.* [60] is improved in this section. Rademan shows empirically that the CWT yields improved results compared to the STFT for practical SNRs. As such, we utilise the first level scattering coefficients $\mathcal{S}_1$, since it is equivalent to a down sampled version of the CWT. We improve on the detector posed by Rademan by introducing true probabilistic output and spectral noise whitening.

5.4.1 Spectral Entropy

Given a discrete-time signal $x[n]$, the first level scattering coefficients are expressed as

$$X[k,m] = \mathcal{S}_1 x[m,\lambda_k],\ \ \lambda_k \in \Lambda_1$$

where k is the frequency index and m is sample index. This notation is adopted to emphasise that X is a TF decomposition similar to a STFT.

The power scalogram $S[k,m]$, which is proportional to the equivalent sinusoidal power per frequency band, may be derived from as $S[k,m] = X^2[k,m]$. S is used to construct a time-varying entropy measure $H[m]$, based on Shannon's entropy:

$$P[k,m] = \frac{S[k,m]}{\sum_{j=0}^{|\Lambda_1|-1} S[j,m]}; \tag{5.14}$$

$$H[m] = -\sum_{k=0}^{|\Lambda_1|-1} P[k,m]\log(P[k,m]). \tag{5.15}$$

5.4.2 Improving SE Robustness

Pre-whitening the power spectrum is an effective method of increasing the robustness between signal and noise entropy measures, and serves to stabilize the noise entropy level by enforcing the assumption of stable noise power spectral density (PSD). To accomplish the whitening, a robust noise PSD estimator which can function with and without signal presence is required.

A rolling median filter operator RM_v^B acting on a signal $y[v,w]$ with respect to the index v, is defined as

$$\mathrm{RM}_v^B y[v,w] = \underset{z}{\mathrm{median}}(y[z,w]),\ z \in \left\{v - \frac{B-1}{2}, ..., v, ..., v + \frac{B-1}{2}\right\}, \tag{5.16}$$

where B is an odd number describing the length of the filtering window. Similarly, we define a rolling average operator RA_v^B as

$$\text{RA}_v^B y[v,w] = \frac{1}{B}\sum_z y[z,w],\ z \in \left\{v - \frac{B-1}{2}, ..., v, ..., v + \frac{B-1}{2}\right\}. \tag{5.17}$$

Under the signal and noise assumptions specified in subsection 5.4.1, the coefficients of potential signals are first suppressed by median filtering over the frequency indices k, followed by another rolling median estimate over time indices m to suppress any transient signals. Finally, a rolling average calculates the time-varying noise PSD. We utilise the amplitude spectra in this estimation, as we find this works better in practise as squaring increases the noise variance and amplifies outliers.

$$\bar{S}[k,m] = \text{RA}_m^F\left(\text{RM}_m^D\left(\text{RM}_k^B\sqrt{S[k,m]}\right)\right)^2. \tag{5.18}$$

In this paper, we set $B = 9$ frequency bins, D to a value corresponding to 5 seconds, and F to 60 seconds.

A whitened power decomposition is calculated:

$$S_w[k,m] = \frac{S[k,m]}{\bar{S}[k,m]}. \tag{5.19}$$

S_w is used instead of S to produce the whitened spectral distribution P_w and SE H_w with equations (5.14) and (5.15) respectively. As in the related study by Rademan *et al.* [60], a rolling median filter is used as an effective method to stabilize the SE measure:

$$H'_w[m] = \text{RM}_m^R(H_w[m]). \tag{5.20}$$

5.4.3 GMM Automatic Threshold Selection

PAM system recordings which span multiple days are bound to be subjected to many noise conditions when tides, weather and the presence of wildlife change. As such, automatic threshold selection for small audio segments which take into account the specific conditions of the segment is paramount, since a fixed threshold will not necessarily yield good results over the entire database.

In the previous study by Rademan *et al.* [60], the k-means algorithm [38] is used to find an adaptive estimate for the low and high levels of an audio segment and generate a pseudo-probabilistic output of signal presence. This pseudo-probability is a first step in normalizing the entropy, but fails to account for the variance of the low and high levels respectively, since the k-means algorithm assumes spherical (equal variance) clusters without accounting for the prior probability of a cluster (i.e., the proportions of signal and noise within the audio segment).

Although this allows for normalized measure between 0 and 1, there is still some uncertainty on the threshold selection, and it requires a hyper-parameter to “squeeze" the measures around the equi-probability point (0.5) with a specific steepness.

To improve on threshold selection, this paper models the whitened and median filtered SE measure H'_w with the probability distribution $p(H'_w)$ as a mixture of Gaussians. A GMM is fitted using the expectation maximization (EM) algorithm. The GMM model accounts for each SE level's variance and the prior probability of each level occurring, assuming independence across all time indices m:

$$H'_w|\mathcal{C} = C_n \sim \mathcal{N}(\mu_n, \sigma_n^2) \tag{5.21}$$

$$H'_w|\mathcal{C} = C_s \sim \mathcal{N}(\mu_s, \sigma_s^2) \tag{5.22}$$

The observable probability density of H'_w is found through marginalization

$$p(H'_w) = p(C_n)\mathcal{N}(H'_w|\mu_n, \sigma_n^2) + p(C_s)\mathcal{N}(H'_w|\mu_s, \sigma_s^2), \tag{5.23}$$

where (μ_n, μ_s) and (σ_n, σ_s) are the noise and signal entropy means and variances respectively. $\mathcal{C} \in \{C_n, C_s\}$ represents the noise and signal levels of H'_w. The class prior probabilities $p(\mathcal{C})$ are given by the mixture weights of the GMM. The notation H'_w is used to express a random variable, over which all $H'_w[m]$ are independent observations indexed by m. The restriction $\mu_n > \mu_s$ is utilised, since the noise entropy measure is always greater than the signal entropy measure under the signal/noise assumptions.

To find the probability of an entropy measure belonging to the signal class C_s, Bayes' rule is employed:

$$p(C_s|H'_w) = \frac{p(H'_w|C_s)p(C_s)}{p(H'_w)}. \tag{5.24}$$

5.4.4 Problems with the Posterior

Equation (5.24) has complex behaviour with respect to its parameters determined by the GMM. It is known a priori that a higher entropy corresponds to a larger probability of noise, thus requiring equation (5.24) to be a monotonically decreasing function over the entire observed entropy range of audio segment of interest. To achieve this, Rademan *et al.* [60] uses a sigmoid function which is shaped by the k-means algorithm and a hyper-parameter. Most audio segments from the tested database do not satisfy the condition of monotonocity of the posterior.

This paper proposes that the posterior is clamped after reaching the monotonicity limits. Refer to appendix 5.A for the derivation of these limits. With this modification made, it is guaranteed that the posterior is a piece-wise function which is either decreasing or remains constant.

$$p_{\text{clamp}}(C_s|H'_w) = \begin{cases} p_{\max}, & \sigma_s^2 > \sigma_n^2 \ \wedge \ x < \theta \\ p_{\min}, & \sigma_s^2 < \sigma_n^2 \ \wedge \ x > \theta \\ p(C_s|H'_w), & \text{otherwise} \end{cases}, \tag{5.25}$$

where

$$\theta = \frac{\sigma_s^2 \mu_n - \sigma_n^2 \mu_s}{\sigma_s^2 - \sigma_n^2}, \tag{5.26}$$

with $p_{\max}$ and $p_{\min}$ as the maximum and minimum values of the posterior function that is observed within the range of H'_w. $p_{\max}$ and $p_{\min}$ are numerically approximated[1]. For an example of the output of the entropy detector, refer to Figure 5.5.

5.4.5 Method

We measure the performance of the detectors via receiver operating characteristic (ROC) curves, by testing each annotated sample in the power scalogram S. No annotation cleaning is performed and all audio files are included. Annotations which have bounding boxes that do not intersect with the 15-120 Hz band are neglected, since the detectors do not operate outside this bandwidth.

Three SE detectors are evaluated: GMM fitted on the non-whitened $\mathcal{S}_1$ coefficients; GMM on the whitened $\mathcal{S}_1$ coefficients; and k-means fitted on the non-whitened $\mathcal{S}_1$ coefficients as

[1] The maximum and minimum entropy values in the posterior distribution is calculated by sampling 1000 regularly spaced points in the observed entropy range and finding the maximum and minimum of the sampling points. For a more efficient approach, Newton's method may be used.

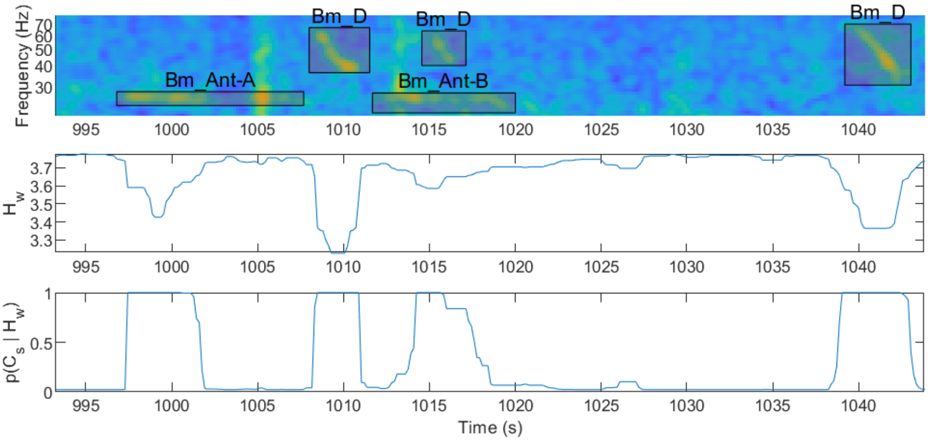

Figure 5.5: An example of the output of the detector taken from a densely populated vocalisation region of the dataset. The annotation bounding boxes are also indicated (top). In this example, the detector successfully detects all vocalisations. The whitened SE measure is shown (middle) with the posterior probability of signal (bottom).

performed by Rademan *et al.* [60]. Comparing these detectors will validate that modelling the entropy with a GMM and applying whitening yields incremental improvements from the original k-means implementation. As a baseline, we also compare the results to a BLED which sums over the whitened $\mathcal{S}_1$ coefficients. This form of BLED is more akin to the detector implementations present in PAM software, as it indirectly adapts to noise conditions during the whitening process. The whitened BLED additionally has its energy median filtered in the same manner as the SE detectors, such that a fair comparison is obtained with respect to whitening and filtering.

To perform unmodified energy detection with wavelet scattering, the average energy E of each time slice calculated using

$$E[m] = \frac{1}{|\Lambda_1|} \sum_{\lambda_1 \in \Lambda_1} \left(\frac{\mathcal{S}_1 x[m, \lambda_1]}{\sqrt{\lambda_1}} \right)^2, \tag{5.27}$$

which converts the wavelet L1 normalisation from Equation (5.1) to L2.

Two methods of evaluation are employed. We perform a sample-by-sample evaluation to produce a ROC curve. This method will measure the proportion of samples which has been detected and therefore penalise sporadic detections. However, this is not an optimal method of measuring the detector. The dataset does not contain accurate annotation bounding boxes, thereby unfairly penalising the detectors based on annotation endpoint accuracy. Additionally, we count the proportion of annotated calls that have at least 1 detected time sample, and plot it against the proportion of noise/un-annotated samples which have been detected. The true efficacy of the detection is therefore determined, however, sporadic detections are not penalised.

5.4.6 Results

Figure 5.6a shows the ROC curve for detection on a sample-by-sample basis. Incremental performance improvement is demonstrated by adding a GMM and enhancing the detector with whitening. In Figure 5.6b, we show that the proposed detector is able to detect the majority calls while detecting fewer noise. As illustrated in many previous studies, SE greatly outperforms a traditional BLED detector [19]. Whitening is also shown to greatly improve the capabilities of the BLED detector.

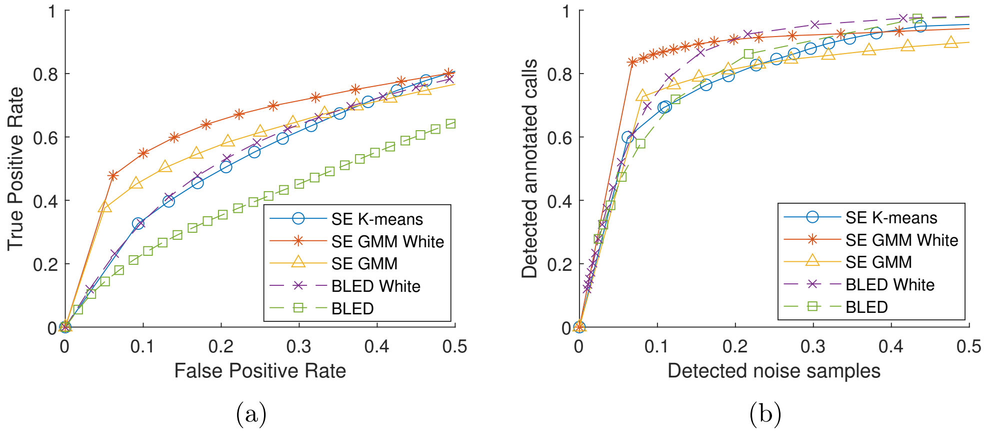

Figure 5.6: Results of the evaluated detectors. SE detectors are indicated as a solid line, while BLED is dashed. (a) The performance of the detectors when compared on a sample-by-sample bases. (b) The performance curve of the detectors expressed as the percentage of detected noise samples versus the percentage of detected calls. The whitened SE with GMM detector significantly outperforms all other detectors.

Table 5.3: The number of missed calls as a result of a non-converging GMM fit.

Vocalisation	Missed vocalisations
SE GMM	196 (5.8%)
Whitened SE GMM	43 (1.2%)

5.4.7 Discussion

Results for STFT-based BLED has been reported for the Casey Islands 2017 dataset in [46] using the PAMGuard Ishmael energy sum detector. These results are not directly comparable to this paper, since the detectors have additional parameters controlling false positives, such as time length and distance between detections and adaptive thresholding. These detections are measured on a per run-length basis, i.e., over the total run of the detector exceeding a threshold, instead of a per-sample basis. Therefore, this form of measurement may not accurately represent endpoint detection capabilities.

As seen in Figures 5.6a and 5.6b, the proposed SE detector outperforms the previous method [60]. Typical BLED with a fixed threshold performs very poorly due to varying noise conditions. BLED has repeatedly been shown to not be a reliable POI detector [79, 19].

Interestingly, adding whitening to BLED results in similar performance to the method proposed by Rademan *et al.* [60]. Spectral whitening and GMM fitting shows incremental improvements to the k-means based SE detector. The final proposed detector outperforms the other detectors both in terms of endpoint accuracy (figure 5.6a) and the number of calls versus noise detected (figure 5.6b). This paper later discusses the implication of annotation inaccuracies in section 5.5.4. As such, the indicated results may not be entirely accurate, and could be better than indicated. From the false positives estimated in Table 5.1, all of the detectors likely detect 5-10% more calls for realistic thresholds than indicated.

5.5 Classification

We combine the output of the SE detector with WS features and utilise linear discriminant analysis (LDA) to classify the detected calls. Our method shares many similarities with studies that also employ WS without a NN, but differs in allowing for variable window sizes produced by a detector [6, 35].

5.5.1 Feature Extraction

Detections are performed with a posterior signal probability threshold of 0.3, thereby corresponding to 91% of vocalisations detected, with a significant number of unlabeled noise detections. Refer to Table 5.4 for the number of detected calls.

After detection with with a whitened SE GMM detector, we threshold the posterior probability given by Equation (5.24). We ensure that an entire vocalisation is captured by applying a rolling maximum filter over Equation (5.24), thereby extending the boundaries of the detections with a length of G. We set the filter length to 6 seconds, thus extending detections by 3 seconds on either side. Note that boundary extension may add additional calls to the detected vocalisations compared to the results from section 5.4.1.

Labels are obtained by assigning each extended detection to an annotation that which occurs in the same timeframe, with the noise class being assigned if there are no corresponding annotations. Annotations have their endpoints shrunk by 5% to account for endpoint inaccuracies. Detections having multiple annotation assignments which do not belong to the same class are discarded in this study. Noise detections may erroneously overlap with annotations, which could negatively impact the classification results. This is not addressed in this paper.

We dispose of the time dimension across scattering features for each detection by taking the maximum over the consecutive detected indices given by the set $\mathcal{M}$ and normalising power:

$$\bar{P} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{\lambda_1 \in \Lambda_1} \left(\frac{\mathcal{S}_1[m, \lambda_1]}{\sqrt{\lambda_1}} \right)^2; \tag{5.28}$$

$$f_1[\lambda_1] = \frac{1}{\bar{P}} \max_{m \in \mathcal{M}} \mathcal{S}_1[m, \lambda_1]; \tag{5.29}$$

$$f_2[\lambda_1, \lambda_2] = \frac{1}{\bar{P}} \max_{m \in \mathcal{M}} \mathcal{S}_2[m, \lambda_1, \lambda_2]. \tag{5.30}$$

Equation (5.28) is the average power over detection indices which is calculated using equation (5.27).

Cosine-log scattering (CLS) features are used, as detailed by Andén and Mallat [6]. First-level CLS features are directly comparable to MFCC, with the addition of the the second level scattering features. The transform is applied to equations (5.29) and (5.30) as follows:

$$F_1[\lambda_1] = \mathrm{DCT}_{\lambda_1}\left(\log f_1[\lambda_1]\right); \tag{5.31}$$

$$F_2[\lambda_1, \lambda_2] = \mathrm{DCT}_{\lambda_2}\left(\log f_2[\lambda_1, \lambda_2]\right), \tag{5.32}$$

where DCT_x denotes the discreet cosine transform over the variable x. The feature vector v is constructed using using F_1 and F_2 over all significant scattering paths. This dataset is normalised to zero-mean and unit variance.

5.5.2 Linear Disciminant Analysis

We train a multiple "one versus the rest" LDA classifiers using a balanced covariance estimate. There are 2 classes (vocalisation and noise) indexed by $i \in \{1, 2\}$, each with K_i training observations. Each model considers only a single vocalisation class and regards all the other classes as noise.

Due to the similarity between calls of the *Bm_ Ant* classes, models for these classes should not be penalised for misclassifying other *Bm_ Ant* call types. As such, *Bm_ Ant* models have all similar classes removed from their training and test sets. Many fin whale calls do not reach the classification stage of this paper due to their presence in the chorus sections of the audio and overlap. As such, we combine the fin whale vocalisation classes (*Bp_ 20Hz*, *Bp_ 20Plus*) into a single class (*Bp*) due to their similarity.

We utilise regularised LDA with prior class probabilities $p(\mathcal{C})$. LDA assumes that each class is normally distributed, with a single class covariance matrix $\Sigma_\gamma = (1-\gamma)\hat{\Sigma} + \gamma\mathrm{diag}\left(\hat{\Sigma}\right)$, where $\hat{\Sigma}$ is the sample covariance estimate shared between the classes and γ is the covariance regularisation hyperparameter. Given an observation x as a column vector, the log posterior probability of x belonging to class $\mathcal{C}_i$ is

$$\log p\left(\mathcal{C} = \mathcal{C}_i | x\right) = -\frac{1}{2}\Sigma_\gamma^{-1}\hat{\mu}_i x - \frac{1}{2}\hat{\mu}_i^T\Sigma_\gamma^{-1}\hat{\mu}_i + \log p(\mathcal{C}_i) + \beta, \tag{5.33}$$

where β is a normalising constant that may be neglected to assign scores to each class.

To combat class imbalance, we use the following estimates:

$$\hat{\mu}_i = \frac{1}{K_i} \sum_{n \in \mathcal{N}_i} x_n; \tag{5.34}$$

$$\hat{\Sigma}_i = \frac{1}{K_i} \sum_{n \in \mathcal{N}_i} (x - \hat{\mu}_i)^T (x - \hat{\mu}_i); \tag{5.35}$$

$$\hat{\Sigma} = \frac{1}{2} \left(\hat{\Sigma}_1 + \hat{\Sigma}_2 \right), \tag{5.36}$$

where $\mathcal{N}_i$ is the set containing the indices of x_n belonging to class $\mathcal{C}_i$. Equation (5.36) approximates the covariance by assuming equal class weights, although this is typically weighted by the observation proportion in other applications. Equal weighting in shared covariance estimation ensures that all classes are able to significantly contribute to the covariance estimate, which combats the high proportion of noise samples in this study.

The prior probabilities $p(\mathcal{C})$ are typically empirically estimated from the training data. In this study, we employ the following model for prior probabilities to combat the high number of noise observations:

$$p(\mathcal{C}) = \begin{cases} \alpha, \; \mathcal{C} = \mathcal{C}_{\text{noise}} \\ 1 - \alpha, \; \mathcal{C} = \mathcal{C}_{\text{vocalisation}} \end{cases}. \tag{5.37}$$

The hyperparameter $\alpha \in (0, 1)$ sets the sensitivity of the classifier to noise detections.

5.5.3 Results

To evaluate our model, we split the training data into training and test sets proportional to the number of items belonging to that class. To demonstrate the practicality of our proposed technique, we perform experiments for different train-test splits with chi-square feature selection. An 80%-20% train-test split without feature selection is also used as a baseline. Results are shown in Table 5.5. Feature selection is performed with a chi-square test of independence (noise versus signal classes). We select features with a statistical significance greater than 0.05, by binning each feature into 10 uniform bins. Feature selection is performed using only the training data.

Classification accuracy is measured using the observations not included in the test set. Unless specified otherwise, training is repeated 20 times and the results presented are the average accuracy and its standard deviation. We additionally perform a coarse grid search to optimise the hyperparameters α and γ and present the results for the best hyperparameters. Hyperparameters are optimised by maximising the cost function $A_{\text{Vocalisation}} + 2A_{\text{Noise}}$, with A the accuracy of the test set, thereby enforcing a stricter policy on false detections.

Table 5.4: Number of labeled detections for all vocalisations of interest. The detections are produced by the whitened GMM SE detector. A significant number of detections are discarded in the training data due to overlapping.

Vocalisation	Number of detections
Bm_Ant_A	1514
Bm_Ant_B	474
Bm_Ant_Z	91
Bm_D	328
Bp	50
Noise	25638
Overlapping	150 [1]

5.5.4 Discussion

High noise rejection accuracies can be observed for both full and small training sets. As discussed in section 5.2, the dataset contains a significant portion of unlabeled and mislabeled vocalisations. There are many vocalisations labelled with very low SNR, thereby not allowing a certain positive identification. Additionally, many low frequency blue whale calls (*Bm_Ant*) are unlabeled, and therefore incorrectly labeled as noise. As such, the models often detect calls which are not annotated, thereby penalising the noise rejection capabilities. Erroneous annotation assignments as a result of boundary extension and inaccurate annotation bounding boxes are also suspected to negatively contribute the results. The high variance compared to the reported median accuracy reflects on the quality of dataset annotations.

Using Table 5.1, we estimate that the true accuracies of the models' vocalisation classification capabilities may be between 1-17% higher than indicated[2], depending on the class. As discussed in section 5.4.6, adjusting for the noise accuracy of the models is out of the

[2] This range is determined by considering that if our models correctly classified all erroneous annotations as noise, then their true accuracy is scaled by between $\frac{1}{1-\text{LB}}$ and $\frac{1}{1-\text{UB}}$, where LB and UB are the lower and upper bounds of erroneous annotation proportions. This calculation stems from the fact that the number of annotations are reduced, weheras the correct number vocalisation classifications remain the same. This range will change significantly based on whether the model classifies erroneous annotations as a vocalisation, which cannot be determined unless similar methods to Miller *et al.* [45] are employed . The estimation does not account for the SE detector's capability to rule out false positives.

Table 5.5: Average model classification accuracies for varying train-test splits with chi-squared feature selection. An 80% training split without chi-square feature selection is included as a baseline. 20 trials are performed, whereas 5% splits use 100 trials. The median accuracy is displayed.

Vocalisation	**Vocalisation Accuracy (%)**	**Noise Accuracy (%)**	α	γ
80% Train (No Feature Selection)				
Bm_Ant_A	82.2 ± 1.5	88.6 ± 1.2	0.66	0.71
Bm_Ant_B	83.3 ± 6.2	92.3 ± 1.1	0.54	0.43
Bm_Ant_Z	92.5 ± 10.6	98.2 ± 0.1	0.05	0.43
Bm_D	74.1 ± 6.6	90.3 ± 2.0	0.66	0.86
Bp	96.0 ± 14.6	99.0 ± 0.1	0.17	0.57
80% Train (With Feature Selection)				
Bm_Ant_A	80.8 ± 1.5	89.9 ± 1.6	0.66	0.29
Bm_Ant_B	84.5 ± 6.1	91.6 ± 3.0	0.54	0.43
Bm_Ant_Z	96.4 ± 10.5	98.0 ± 0.5	0.41	0.43
Bm_D	73.6 ± 5.3	90.7 ± 4.0	0.66	0.71
Bp	97.5 ± 9.2	98.0 ± 0.3	0.54	0.57
20% Train (With Feature Selection)				
Bm_Ant_A	80.3 ± 4.9	89.0 ± 8.8	0.66	0.43
Bm_Ant_B	79.1 ± 15.2	92.1 ± 4.3	0.54	0.57
Bm_Ant_Z	91.8 ± 12.4	97.9 ± 0.4	0.17	0.29
Bm_D	71.4 ± 12.9	90.6 ± 4.7	0.66	0.86
Bp	92.5 ± 13.8	97.9 ± 0.2	0.41	0.71
5% Train (With Feature Selection)				
Bm_Ant_A	73.5 ± 6.1	89.6 ± 8.5	0.66	0.57
Bm_Ant_B	64.7 ± 19.5	93.2 ± 12.3	0.54	0.71
Bm_Ant_Z	73.5 ± 18.7	95.7 ± 0.5	0.29	1.00
Bm_D	64.4 ± 18.5	90.7 ± 8.8	0.41	0.86
Bp	81.2 ± 19.9	95.2 ± 12.7	0.54	0.86
5% Train (With Feature Selection), Lowest 5% SNR Calls Removed				
Bm_D	74.2 ± 15.7	89.1 ± 6.1	0.05	0.86

scope of this paper, since it must also account for the SE detector's capability of removing false postives. Our estimated proportion of false positives correspond well with the discrepency in model accuracies per class: a higher proportion of estimated false positive annotations yields a clear lower classification performance compared to other models.

The use of feature selection greatly increases performance for smaller datasets, as the accuracies for a 20% training size with feature selection is nearly identical to that of the 80% training size without feature selection. Feature selection is therefore crucial when attempting to use smaller datasets.

A lower accuracy for *Bm_ D* is to be expected, since frequency-modulated downsweeps require temporal information and adjacent frequency information to be extracted. Furthermore, noise impulse sources may also resemble downsweeps when temporal information is discarded, as is performed in this study. There are especially many *Bm_ D* annotations in the dataset with a very low SNR, which are most likely noise patterns that happen to resemble a downsweep. This is suspected to have a large impact on *Bm_ D* noise rejection accuracy. This study may also be extended to use higher order wavelet scattering features, such as deeper levels and joint-TF scattering, which results in classification that is more reminiscent of CNN operating on spectrogram [4]. Joint-TF scattering will greatly improve the *Bm_ D* noise rejection.

The use of the annotations as the “ground truth" is questionable and is well documented and discussed by Miller et. al. [45]. Miller utilises the same dataset that is used in this study to train a *Bm_ D* classifier, and notes that annotations are significantly inaccurate across the dataset. They use an additional dataset for model evaluation, gathered in 2019 at the same location. Miller notes that more than 15% of the analyst annotations in the evaluation dataset are false positives, and that there are nearly 30% more calls than initially indicated by the analyst. After adjustment procedures, their measured model accuracy increased from 75% (ground truth analyst notation) to 92% (adjusted). This adjusted accuracy also corresponds well to the estimated adjustment required for our *Bm_ D* model (8-17%), assuming our model did not detect any erroneous annotations and that the entropy detector did not significantly alter the proportion of false positives. This effect is clearly seen when discarding low-SNR samples from training and testing, where the *Bm_ D* accuracy increases by 10% compared to when all samples are included.

The fin whale call (*Bp*) indicates particularly effective classification, despite utilising only 5 training samples in the most extreme case. However, the accuracy varies dramatically over the randomised training trails, due to the small sample size, making the model sensitive to the choice of training data. Erroneous and low-SNR annotations are suspected to contribute to the large variation when the training samples are small (< 100).

For the *Bm_ Ant_ A*, *Bm_ Ant_ B* and *Bm_ D* classes, our noise rejection may not be sufficient, which currently results in > 2000 false positives per class, a portion of which may be annotations missed by the analyst. Manual inspection of some audio segments indicate that there are large portions of unannotated calls from these classes. Further investigation is required to determine the practical acceptability of optimised noise rejection capabilities. The *Bm_ Ant_ A* and *Bm_ Ant_ B* models detect many unannotated vocalisations of all *Bm_ Ant* types due to their similarity, thus explaining their lower noise accuracy compared to the *Bm_ Ant_ Z* and *Bp* models, which tend to have higher SNR compared to the other classes. Additionally, segments in which blue whale choruses are present, which are typically left unannotated, may negatively impact the *Bm_ Ant* models' results.

Since we use independent models for classification, some calls will be detected by multiple models, especially for similar calls such as the *Bm_Ant* classes. It is not possible to implement multiclass classification in this paper due to large proportion of analyst misclassifications in the *Bm_Ant* classes. However, the posterior class probability may be used determine the correct class when multiple models have made a non-noise prediction in further studies.

Our disposal of the time dimension by using the maximum naturally lends our classification model being replaced with a HMM to incorporate time-dependent relationships similar to that of Ogundile *et al.* [51] and Usman and Versfeld [75]. However, instead of classifying on a frame-by-frame basis such as in Sattar [65], Andén and Mallat [6], Liu *et al.* [35], we utilise the scattering coefficients to first detect a variable-length segment of audio which may contain whale vocalisations, thereby reducing the amount of data in the noise class. This provides a distinct advantage over other methods, in that it already incorporates the extracted feature coefficients in detection prior to classification. As such, large portions of noise are discarded, the amount of which is tuned by the probability threshold of the proposed SE detector. This contrasts to Miller *et al.* [45], which contains over 3 million noise class samples for 5000 vocalisation samples using the sliding window approach. This corresponds to a ratio of 1:600 for vocalisation to noise classes, whereas our method, which includes a detector, reduces this ratio to approximately 1:7 (all classes), at the cost of rejecting very low SNR vocalisations.

Despite the problems posed by the dataset used for evaluation, this study is the first to show the effectiveness of non-neural network models combined with scattering features. A detector is also utilised to remove noise segments, which assists in reducing class imbalance and reduces training time. The proposed method is particularly effective for a very small and unbalanced dataset, while allowing for a variety of calls to be detected.

5.6 Conclusion

In this study, we introduce WS into the field of PAM for whale detection and classification and suggest further improvements of the traditional SE detector. We combine the detector with WS and demonstrate the practicality of the proposed technique on a real, large dataset for blue and fin whale calls. We evaluate classification accuracy for many types of calls, demonstrating the applicability of the proposed technique on many whale species.

The classical SE detector is not equipped for adaptivity to a changing noise profile, which this study improves upon by applying a noise whitening procedure and fitting the SE output with a GMM. The detector provides true probabilistic output, which accounts for varying noise and signal conditions, with the added benefit of interpretability for the user of this detector. The true-probabilistic extension and noise-whitening of the classical SE detector in this paper may be employed as a method of modifying and improving other existing detectors in terms of noise/signal adapativity and detector output interpretability.

The proposed system uses first level WS coefficients, thereby allowing an efficient detection and feature extraction using a single TF decomposition. Using a sensitive detector prior to classification allows for the elimination of large portions of noise, which assists in reducing the noise/signal class imbalance.

We use the SE detector's output is used to find POIs – identifying windows of an arbitrary length to classify with CLS features with LDA. For this study, we ignore the time-dependent nature of the features in the windows by taking the maximum of the features over the detected time-frame, and reduce the number of features using chi-square feature selection. Feature selection proves to be effective in improving the performance for smaller datasets.

Further work includes incorporating time information into the classification stage, such as utilising models like HMMs. Frequency modulated calls, like the blue whale downsweep, is not well-suited to the proposed technique which may be addressed by modifying the proposed method with joint-TF features, thereby more closely resembling the front-end of a 2D CNN operating on a spectrogram. This modification will allow for relationships between frequency bins to also be taken into account, whereas the proposed WS feature extraction method does not. More advanced and/or optimised feature selection methods may also be investigated to improve classification performance for smaller datasets.

The distinct advantage of the LDA classifier is the ability to use very small datasets consisting of 10s of samples per class to find other similar calls in an otherwise large audio database. Although not providing SOTA classification accuracy, the proposed technique enables conservation researchers to search for similar sounds without requiring many training examples, allowing researchers to correctly classify more than 65% of similar vocalisations detected by the SE detector when tuned for high noise rejection ($> 90\%$) with few training examples. The presented classification system will still yield some false positive detections, but will significantly reduce the effort required during annotation.

This paper provides the groundwork required to implement WS-based classification without the need of NNs and large datasets, and provides a method to combine classical signal detectors and classifiers in an efficient way, thus bridging the gap between POI detectors and classification models.

Appendices

5.A Conditions of Monotonicity of the GMM Posterior

Recall the normal distribution and its derivative:

$$\frac{d}{dx}\mathcal{N}(x|\mu,\sigma^2) = 2\frac{\mu - x}{\sigma^2}\mathcal{N}(x|\mu,\sigma^2). \tag{5.38}$$

We want to find the regions in which the posterior is decreasing:

$$\frac{d}{dx}p(C_s|x) < 0, \tag{5.39}$$

where $x = H'_w$. $H'_w \in [0, \log(N)]$ follows trivially from equation (5.15) and from the uniform distribution producing maximum entropy. N is the number of points in the spectral distribution.

The following restrictions are given:

1. Finite support on the interval: $x_{\text{max}} > x > x_{\text{min}}$.
2. The EM algorithm will converge to a solution where $x_{\text{max}} > \mu_n > \mu_s > 0$.

Let $\varphi_i = p(C_i)\mathcal{N}(x|\mu_i,\sigma_i^2)$, implying $\frac{d}{dx}\varphi_i = \varphi'_i = k_i\mathcal{N}(x|\mu_i,\sigma_i^2)$, where $k_i = 2\frac{\mu_i - x}{\sigma_i^2}$.

We calculate the derivative required by equation (5.39):

$$\frac{d}{dx}p(C_s|x) = \frac{\varphi_n\varphi_s(k_s - k_n)}{(\varphi_n + \varphi_s)^2}. \tag{5.40}$$

Since φ_n and φ_n are always positive and nonzero, equation (5.39) reduces to $k_s - k_n < 0$:

$$(\sigma_s^2 - \sigma_n^2)x - \sigma_s^2\mu_n + \sigma_n^2\mu_s < 0, \tag{5.41}$$

which is linear with respect x, showing that the curvature only changes direction once when assuming all other variables are constants.

Case 1: $\sigma_s^2 - \sigma_n^2 > 0$: $x < \frac{\sigma_s^2\mu_n - \sigma_n^2\mu_s}{\sigma_s^2 - \sigma_n^2}$.

Case 2: $\sigma_s^2 - \sigma_n^2 < 0$: $x > \frac{\sigma_s^2\mu_n - \sigma_n^2\mu_s}{\sigma_s^2 - \sigma_n^2}$.

Case 3: $\sigma_s^2 = \sigma_n^2$: $\mu_s < \mu_n$, which is always true given the entropy assumptions.

Case 3 is always true under the restrictions stated for this problem, implying that case 3 will always result in a monotonically decreasing function. To evaluate whether a given GMM model with $x \in [x_{min}, x_{max}]$ satisfies monotonicity conditions, we substitute x for x_{max} and x_{min} for cases 1 and 2 respectively.

5.B Summary of parameter values

Table 5.6: Summary of parameter values used in this study. Chosen values are indicated in boldface text, whereas non-bold parameters result from the chosen parameters.

Parameter	Description	Value
T	WS invariance scale	**1.5 seconds**
Q_1	Wavelets per octave in $\mathcal{S}_1$	**16**
Q_2	Wavelets per octave in $\mathcal{S}_2$	**4**
$f_{\min}$	Minimum filterbank frequency in $\mathcal{S}_1$	**15 Hz**
$f_{\max}$	Maximum filterbank frequency in $\mathcal{S}_1$	**120 Hz**
$\|\Lambda_1\|$	Resulting number of filters in $\mathcal{S}_1$	48
$\|\Lambda_2\|$	Resulting number of filters in $\mathcal{S}_2$	14
V	Wavelet scattering upsample factor	**2**
r	Downsampling factor of $\mathcal{U}_1$ due to maximum bandwidth	8
d	Downsampling factor of $\mathcal{S}_1$	5
f_{ss}	Resulting scattering transform sample frequency	6.25 Hz
B	Median filter length for noise coefficient suppression	**9**
D	Median filter length for transient coefficient suppression	**5 seconds**
F	Average filter length for noise coefficient estimation	**60 seconds**
R	Median filter length for stabilizing entropy	**1.5 seconds**
G	Maximum filter length for boundary extension	**6 seconds**
p_t	Probability threshold for the whitened SE GMM detector	**0.3**
k	Dimensions in the classification feature vector v	521

6 Wavelet Scattering in Higher Dimensions

In chapter 4, we reviewed a generalised definition of a multi-dimensional TF decomposition X, which is processed by a set of invariance-enforcing filters Φ_i. Wavelet scattering in 1 dimension utilising a single invariance-enforcing filter, ϕ, as discussed in chapter 5.

Wavelet scattering filterbanks have been extended to multiple dimensions using various different wavelets and filterbank construction techniques. This chapter provides an overview of the existing filterbanks and filters and their properties. We then identify a gap in the literature, which gives rise to a generalisation of cascasded 1D scattering transforms – separable wavelet scattering.

6.1 Types of Scattering Transforms

Existing scattering transforms are listed below convenience:

1. 1D scattering [2, 7]. This is the first-introduced scattering transform, which serves as a generalisation of MFCCs. It improves on MFCCs by recovering the information lost due to invariance filter (averaging) operations. Morlet wavelets are used, with cascaded scattering levels having its own number of filter per octave.

2. 2D scattering [12]. Gabor filters modified to 0-mean, are used to construct the filterbank. The filterbank consists of exponentially placed filters in one axis, which is then rotated by equally spaced angles to cover the frequency plane. This filterbank only has a single filter per octave ($Q = 1$). In particular, in polar coordinates, filters are placed exponentially along r, but linearly along θ.

3. 3D scattering – harmonic scattering [18]. The spherical harmonics are used as filters which consitute an orthogonal basis in 3 dimensions. These filters have “lobes” that extend in a multitude of dimensions, depending on the harmonic, in the frequency domain.

4. 3D scattering – extension of 2D scattering [71]. The filterbank is constructed similarly to 2D scattering, where Gabor filters are spaced with linearly varying angles (in spherical coordinates) in 3 dimensions.

5. Joint time-frequency [4, 5] – extending 1D to 2D. A first level of scattering in performed in 1D to produce a scalogram ($X[k, m]$). 2D scattering is then performed on subsequent levels, directly on the scalogram. The 2D wavelets are separable - a product of 1D wavelets. This allows 1D scattering on audio to have features that are also dependent on adjascent frequency bins. It is conceptually similar to a CNN operating on a STFT spectrogram.

6.2 A Detailed view of Joint Time-Frequency Scattering

6.3 The Benefits of Generalising Joint Time-Frequency

7 Separable Wavelet Scattering

M.W. Rademan, D.J.J Verself, J.A. Du Preez

Accepted in IEEE Signal Processing Letters on 15 May 2024

doi: doi

Abstract

Wavelet scattering is a widely used feature extraction method due to it efficacy in extracting invariant features, while retaining any lost high frequency information resulting from averaging to obtain the desired amount of invariance. By generalising a 1-dimensional scattering transform, we extend its definition to an arbitrary number of independent dimensions. We show that, in a modern deep learning setting, separable wavelet scattering performs similarly to its non-separable counterparts with the MNIST hand-written digit dataset. We additionally demonstrate state-of-the-art results for a subset of the MedMNIST3D datasets.

7.1 Introduction

WS has seen widespread use in classification applications as a powerful feature extraction method. It is an especially effective method for small datasets, since the feature filters are reminiscent of a CNN front-end, while requiring no learning. The resulting features are invariant across all averaged dimensions, and exhibit separable class subspaces, allowing linear classifiers to be used with high efficacy [12].

Various forms of WS exist, with 1D scattering first proposed by Anden and Mallat [2, 7], which was later extended to 2 [12] and 3 dimensions [71, 18]. Additional and more specialised forms for WS include rotation-invariant scattering [40] for the classification of textures and joint-TF scattering [4, 5]. Joint-TF scattering utilises a separable 2D filter that operates on the first level scattering scalogram, denoted by the operator $\mathcal{U}_1$, which is the only apparent usage of separable filters in the current scattering literature.

In a deep-learning setting, sensible initialisation of filters prior to learning can significantly improve performance and interpretability [61]. The greatest advantage of separable filters is their computation speed when computing convolutions directly, as is performed in CNNs [41]. Additionally, separable filters reduce the number of parameters of a CNN if the filters are configured to be learnable. Learnable filters can typically improve performance compared to their fixed-filters counterpart [44].

At first glance, a WS decomposition seems to be prohibitively expensive, but fast algorithms are possible due to the limited bandwidth of filter output. Fast algorithms utilise downsampling to take advantage of the demodulation of band-limited signals that result from the convolution of analytic wavelets [12, 71, 2]. Path pruning is also used to ignore filter combinations which have negligible energy. These algorithms are the standard implementation on many platforms, such as the Kymatio python package [3] and MATLAB [42].

The Morlet wavelet is the most widely used wavelet filter for a scattering filter bank implementation. Since the Morlet does not have compact support, implementations vary when considering the bandwidth/time support of the Morlet. This affects how Morlet finite impulse response (FIR) filters are discretised and truncated, how scattering paths with negligible energy are pruned, and how filters overlap in the frequency domain.

In this work, we generalise 1D scattering to an arbitrary number of dimensions, which we refer to as the separable scattering transform. We propose flexible definitions for the Morlet bandwidth and filter overlap, which allows for an exact definition of nearoptimal convolution computations with downsampling. Our implementation of separable scattering is GPU accelerated, similar to the Kymatio implementations [3].

We demonstrate using the MNIST handwritten digit dataset [32] that separable wavelet filters still perform adequately in a NN classification setting compared to the conventional 2D scattering filters. We demonstrate the arbitrarily scalable dimensionality of the separable scattering transform with three-dimensional medical datasets from MedMNIST3D [83]. Separable scattering coefficients show SOTA results for some of the MedMNIST3D datasets when combined with a simple NN, while performing comparably to CNN benchmarks on datasets which do not indicate SOTA results.

7.2 Separable Morlet Filterbank

7.2.1 Morlet

We define a 1D zero-mean Gaussian as

$$\theta_{\sigma_t}(t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{1}{2}\left(\frac{t}{\sigma_t}\right)^2}, \tag{7.1}$$

which has a Fourier transform transform $\theta_{\sigma_t}(t) \overset{\mathcal{F}}{\longleftrightarrow} \hat{\theta}_{\sigma_\omega}(\omega)$, where $\sigma_\omega = \frac{1}{\sigma_t}$.

For an arbitrary bandwidth definition, we define the bandwidth-to-σ ratio $\beta \in \mathbb{R}^+$, such that the one-sided bandwidth of a zero-mean Gaussian is $\beta\sigma_\omega$. For reference, Kymatio's implementation of 2D scattering indirectly defines $\beta \approx 2.5$ [3].

A 1D Morlet ψ has energy concentrated around 1 rad/s, with most of its energy contained in the interval $\omega \in [1 - \frac{1}{Q}, 1 + \frac{1}{Q}]$. $Q \in \mathbb{R}^+$ is defined as the number of wavelets per octave. The mother wavelet is given as

$$\psi(t) = \theta_{\sigma_t}(t)\left(e^{jt} - \theta_0\right), \tag{7.2}$$

where $\theta_0 = \frac{\theta_{\sigma_t}(-1)}{\theta_{\sigma_t}(0)}$ to ensure zero mean: $\hat{\psi}(0) = 0$.

For reasons that will become apparent in section 7.2.2, we defined a 1D-wavelet dilated by a factor λ as

$$\psi_\lambda(t) = \begin{cases} \psi(\lambda t), & \lambda \neq 0 \\ \phi_j(t), & \lambda = 0 \end{cases}, \tag{7.3}$$

where ϕ represents the 1D LPF utilised in scattering computations for the current dimension (j).

A n-dimensional separable wavelet may be constructed similarly, with

$$\psi_{\boldsymbol{\lambda}}(\mathbf{u}) = \prod_{i=0}^{n-1} \psi_{\lambda_i}(u_i), \tag{7.4}$$

where $\mathbf{u} = (u_1, ..., u_n)^T$ is the vector containing the dimensions of interest.

A filterbank is constructed by dilating the mother wavelet with a set of dilation factors $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)^T \in \Lambda_1 \times ... \times \Lambda_n$, with Λ_i the set of dilation factors for the variable u_i. The dilation factor λ_i is also the centre frequency of the dilated wavelet in rad/s.

7.2.2 1D Filterbank Construction

Suppose a 1D Morlet has a centre frequency λ_0. The following filter is placed at $\lambda_0 2^{\frac{1}{Q}}$. To define the amount of overlap between filters, it is useful express the corresponding filter's frequency standard deviation (SD) σ_ω in terms of the distance between filters in the frequency domain. If the wavelet at λ_0 requires a decay equivalent to $\alpha\sigma_\omega$ SDs at the next wavelet at $\lambda_0 2^{\frac{1}{Q}}$. It follows that $\sigma_\omega = \frac{1}{\sigma_t} = \frac{1}{\alpha}\left(2^{\frac{1}{Q}} - 1\right)$.

We refer to α as the clearance factor. More overlap will result in more redundancy in the frequency representation, but also shorter filter impulse responses.

The provided definitions of filter clearance (α) and bandwidth (β) may also be extended to non-separable wavelets, such as the rotational construction employed in [12]. However, a conversion is required to find the bandwidth in each axis for a specific wavelet for rotationally constructed filterbanks, in order to employ the downsampling strategies proposed in this paper.

All filterbank descriptions in this paper utilise normalised sampling frequency, i.e., the sample frequency is $f_s = 1$.

For some audio applications, for example speech recognition, low-frequency content is negligible, motivating filterbank construction starting at a specified frequency ω_0. Otherwise, the entire frequency domain may be covered by setting ω_0 to start at a position specified by ϕ and α.

The LPF ϕ is constant across all layers in a 1D scattering transform, and is chosen as $\phi(t) = \theta_{\sigma_{\phi,t}}(t)$, where $\sigma_{\phi,t} = \dfrac{d}{\pi\beta}$.

The scattering transform requires that the time support of all filters do not exceed the time support of ϕ, ensuring no filters contain time information exceeding the invariance scale d. d is equivalently defined as the total downsample factor of the scattering tranform, and is unique for each dimension. Equivalently, the frequency SD of the filters may not exceed the SD of $\phi(\omega)$. For brevity, we denote the frequency SDs as $\sigma_{\lambda,\omega}$ and $\sigma_{\phi,\omega}$ for the Morlets and the LPF respectively.

Since the maximum time support (minimum bandwidth) is a function of d, all dilated wavelets have their Gaussian envelopes restricted to a maximum time support of $\sigma_{\phi,t}$. This requires linearly spaced filters until the dilated bandwidth is larger than the LPF bandwidth. To construct a set of positive dilation factors $\lambda \in \Lambda^+$, refer to algorithm 1.

Algorithm 1 Λ^+ construction.

```
  $\sigma_{\phi,\omega} \leftarrow \frac{\pi\beta}{d}$
  if frequency limited then
Ensure: $\omega_0 \geq \frac{\pi\alpha}{d}$
      $\lambda \leftarrow \omega_0$
  else
      $\lambda \leftarrow \frac{\pi\alpha}{d}$
  end if
  $\Lambda^+ \leftarrow \varnothing$
  $\sigma_\omega \leftarrow \frac{1}{\alpha}\left(2^{\frac{1}{Q}} - 1\right)$
  $\sigma_{\lambda,\omega} \leftarrow \lambda\sigma_\omega$
  while $\sigma_{\phi,\omega} > \sigma_{\lambda,\omega}$ and $\lambda < \pi$ do
      $\Lambda^+ \leftarrow \Lambda^+ \cup \{\lambda\}$
      $\lambda \leftarrow \lambda + \alpha\sigma_{\phi,\omega}$
      $\sigma_{\lambda,\omega} \leftarrow \lambda\sigma_\omega$
  end while
  while $\lambda < \pi$ do
      $\Lambda^+ \leftarrow \Lambda^+ \cup \{\lambda\}$
      $\lambda \leftarrow \lambda 2^{\frac{1}{Q}}$
      $\sigma_{\lambda,\omega} \leftarrow \lambda\sigma_\omega$
  end while
```

Only positive λ's have been defined thus far, which provides inadequate coverage of the frequency domain in multiple dimensions. For real input signals, it is only necessary to cover half of one of the dimensions (only positive λ's), whereas full coverage (both negative and positive λ's) is required for additional dimensions. On-axis coverage is also required, in which each ψ_λ must be multiplied with a Gaussian (zero-frequency wavelet), which the dilated wavelet definition in equation (7.3) defines as $\lambda = 0$. A similar construction procedure is followed in [5].

Given m 1D filterbanks, with $m \geq 2$, each having Morlet filters with a positive set of lambdas Λ_i^+ and invariance scales $\mathbf{d} \in \mathbb{N}^m$, with $i = 1, ..., m$ indexing the dimension, we construct the m-dimensional filterbank with

$$\mathbb{F} = \{\psi_{\boldsymbol{\lambda}}(\mathbf{u}) \mid \boldsymbol{\lambda} \in \mathbb{L} \setminus \{\mathbf{0}\}\}, \tag{7.5}$$

where $\mathbf{u}$ is the m-dimensional spatial and/or time variable in which the each 1D filterbank is defined. By definition,

$$\mathbb{L} = (\Lambda_1^+ \cup \{0\}) \times (\Lambda_2^+ \cup \Lambda_2^- \cup \{0\}) \times ... \times (\Lambda_m^+ \cup \Lambda_m^- \cup \{0\}); \tag{7.6}$$

$$\Lambda_i^- = \left\{-\lambda \mid \lambda \in \Lambda_i^+\right\}, \tag{7.7}$$

where $\cup$ indicates the set union operator and $\times$ the Cartesian product. For $m = 1$, the provided definitions result in a conventional 1D scattering transform [2].

The m-dimensional LPF is defined as

$$\phi(\mathbf{u}) = \psi_{\mathbf{0}}(\mathbf{u}). \tag{7.8}$$

7.3 Separable Scattering Transform

7.3.1 Transform

The scattering transform requires 2 steps to provide scattering coefficients. The scalogram operator $\mathcal{U}_j$ iteratively filters a discrete signal $x[\mathbf{n}]$ for a given set of filters $\mathbb{F}$, which is then averaged by the LPF. $\mathbf{n}$ represents a multidimensional index variable.

$$\mathcal{U}_j x[\mathbf{n}, \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_j] = \left|(\mathcal{U}_{j-1} \; ... \; \mathcal{U}_1 x) * \psi_{\boldsymbol{\lambda}_j}\right|, \; \forall \; \psi_{\boldsymbol{\lambda}_j} \in \mathbb{F}; \tag{7.9}$$

$$\mathcal{U}_1 x[\mathbf{n}, \boldsymbol{\lambda}_1] = |x * \psi_{\boldsymbol{\lambda}_1}|, \; \psi_{\boldsymbol{\lambda}_1} \in \mathbb{F}. \tag{7.10}$$

The scattering operator $\mathcal{S}_j$ provides the output coefficients at the j'th order of the scattering transform:

$$\mathcal{S}_j x[\mathbf{n}, \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_j] = \mathcal{U}_j x * \phi. \tag{7.11}$$

The LPF ϕ remains constant throughout the transform. Note that the modulus/magnitude operator $|\cdot|$ demodulates the output of the filters, effectively extracting the Hilbert envelope from a band-limited signal [31].

The j'th scattering order adds an additional axis of paths indexed by $\boldsymbol{\lambda}_j$. However, not all paths need to be evaluated, since some paths have smaller bandwidths, thereby requiring fewer filters to extract the information lost by averaging. In particular, we only evaluate paths in which the centre frequencies of all elements of the vector $\boldsymbol{\lambda}_j$ are smaller than their corresponding bandwidth of the previous path's filter $\psi_{\boldsymbol{\lambda}_{j-1}}$. Path pruning is therefore dependent on α and β.

7.3.2 Downsampling Strategy

Since each filter specified by $\boldsymbol{\lambda}$ has its own bandwidth, we can employ downsampling across all paths non-uniformly. However, some care is required to ensure the compounded downsampling steps across all paths result in a uniform sampling frequency of the output scattering coefficients.

Many applications are insensitive to small changes in d. As such, we propose a strategy to find an optimal d given a target and tolerance value. For some applications, choosing d such that the downsampling factor is a power of 2 is the simplest solution to achieve optimality.

Without prior knowledge of the filterbank configurations, given a target invariance scale of $\bar{d}$ samples and a tolerance ϵ, we can optimise $d \in \{\lfloor(1-\epsilon)\bar{d}\rfloor, \lceil(1+\epsilon)\bar{d}\rceil\}$ such that it results in largest number of supported downsampling configurations.

A downsample factor d which decomposes into a set of n prime factors $\{p_1, ..., p_n\}$ with a corresponding multiset $\mathbb{M} = \{m_1, ..., m_n\}$, where m_i is the multiplicity of the prime p_i. We can find an optimal d by maximising the sum $\sum\limits_{m \in M} m$.

Morlet filters in a 1D filterbank may be downsampled by a factor d_{ψ_1} prior to low-pass filtering, and then downsampling again by a factor d_{ϕ_1} after low-pass filtering. As such, the compounded effect of downsampling restricts $d = d_{\psi_1} \cdot d_{\phi_1}$.

In the second order of scattering, the process is repeated with an additional pre-low-pass downsampling factor of d_{ψ_2}. The second level application of ϕ then downsamples by a factor d_{ϕ_2}. To maintain a consistent output sampling frequency, it restricts $d = d_{\psi_2} \cdot d_{\psi_1} \cdot d_{\phi_2}$.

Continuing the downsampling chain, the i'th level of downsampling requires $d = d_{\psi_i} \cdot ... \cdot$

$d_{\psi_1} \cdot d_{\phi_i}$. The output of each operation of $\mathcal{U}_i$ and $\mathcal{S}_i$ must be downsampled as much as possible in order to make subsequent operations faster. To ensure that the application of all the downsampling steps are efficient, we require d to have as many prime factors as possible, including factor multiplicity, so that a larger variety of downsampling combinations may be supported.

Consider a 1D wavelet filterbank and a single wavelet filter $\psi_1[n]$, applied to a discrete-time signal x. The operations required to compute the scattering coefficients is then notated for simplicity as

$$u_1[n] = |x * \psi_1| ; \tag{7.12}$$

$$s_1[n] = y_1 * \phi. \tag{7.13}$$

The bandwidth of u_1 is the bandwidth of an arbitrary first filter ψ_1. This follows from the Hilbert envelope computed by the analytic wavelet filter ψ_λ and the modulus $|\cdot|$. The bandwidth of s_1 is the bandwidth of ϕ.

Critical downsampling of a wavelet ψ with a bandwidth of σ_ω is achieved by a factor of $d_\psi = \left\lfloor \frac{\pi}{\beta\sigma_\omega} \right\rfloor$. Critical downsampling of s_1, is achieved via a factor of d, by definition.

We can efficiently compute s_1 using compounded downsampling steps:

$$(s_1)_{\downarrow d} = \Big((u_1)_{\downarrow d_1} * (\phi)_{\downarrow d_1} \Big)_{\downarrow d_2}, \tag{7.14}$$

such that $d = d_1 \cdot d_2,\ d_1, d_2 \in \mathbb{N}^+$, with $d_1 | d$ and $d_1 \leq d_{\psi_1}$. In order to find d_1, we decrement d_{ψ_1} until it divides d evenly. Each scalogram $u_j[n]$ is not necessarily downsampled optimally, but has a downsampling factor which guarentees a consistent scattering coefficient output sample frequency.

A second order of scattering with a filter path of (ψ_1, ψ_2) is performed on the downsampled u_1:

$$u_2 = \left| (u_1)_{\downarrow d_1} * (\psi_2)_{\downarrow d_1} \right| ; \tag{7.15}$$

$$(s_2)_{\downarrow d} = \Big((u_2)_{\downarrow d_3} * (\phi)_{\downarrow d_3} \Big)_{\downarrow d_4}, \tag{7.16}$$

such that $d = d_1 \cdot d_3 \cdot d_4,\ d_1, d_3, d_4 \in \mathbb{N}^+, d_3 \leq d_{\psi_2}, d_3 | \frac{d}{d_1}$. The proposed downsampling scheme can be extended to an arbitrary number of scattering orders.

Table 7.1: MedMNIST classification results of compared to baseline NN approaches ([83])

Methods	Organ ($c = 11$)		Nodule ($c = 2$)		Fracture ($c = 3$)		Adrenal ($c = 2$)		Ves
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AU
ResNet-1810+2.5D	0.977	0.788	0.838	0.835	0.587	0.451	0.718	0.772	0.7
ResNet-1810+3D	0.996	0.907	0.863	0.844	0.712	0.508	0.827	0.721	0.8
ResNet-1810+ACS41	0.994	0.900	0.873	0.847	0.714	0.497	0.839	0.754	0.9
ResNet-5010+2.5D	0.974	0.769	0.835	0.848	0.552	0.397	0.732	0.763	0.7
ResNet-5010+3D	0.994	0.883	0.875	0.847	0.725	0.494	0.828	0.745	0.9
ResNet-5010+ACS41	0.994	0.889	0.886	0.841	**0.750**	**0.517**	0.828	0.758	0.9
auto-sklearn11	0.977	0.814	**0.914**	**0.874**	0.628	0.453	0.828	**0.802**	0.9
AutoKeras12	0.979	0.804	0.844	0.834	0.642	0.458	0.804	0.705	0.7
Separable WS + NN (Ours)	**0.998**	**0.941**	0.858	0.797	0.614	0.458	**0.875**	0.792	**0.9**

7.3.3 Convolutions

Optimal FFT convolutions can be achieved by performing downsampling in the frequency domain instead of the time domain. It is straightforward to verify that $|x * \psi|_{\downarrow r} = |(x * \psi)_{\downarrow r}|$, since the modulus is an element-wise operation, for some downsampling factor r. Given the signal and filter Fourier transforms $\hat{x}$ and $\hat{\psi}$, we then have

$$x \otimes \psi[n] \xleftrightarrow{\mathcal{FFT}} \hat{x} \cdot \hat{\psi}[k], \tag{7.17}$$

where the $\otimes$ operator represents a circular convolution and k is the frequency index.

Given that $r|N$, we can express (7.17) when downsampled as a periodised summation [31] in the frequency domain

$$(x \otimes \psi[n])_{\downarrow r} \xleftrightarrow{\mathcal{FFT}} \frac{1}{r} \sum_{i=0}^{r-1} \hat{x} \cdot \hat{\psi}[k + iN/r], \tag{7.18}$$

which can be efficiently implemented via shape manipulation of tensors in computational packages like MATLAB or PyTorch [57]. To compute valid convolutions, we must pad x and ψ to have a total length of $N = N_x + d + c$, where $c \in \mathbb{N}^+$ is a constant that ensures that $d|N$.

Table 7.2: MNIST classification error rate (%) of separable and 2D scattering coefficients using a NN classifier

2D WS + NN	$l = 1,\ J = 2$	0.64 ± 0.05
	$l = 2,\ J = 3$	0.50 ± 0.03
Separable WS + NN	$l = 1,\ d = (4, 4)$	0.63 ± 0.05
	$l = 2,\ d = (4, 4)$	0.52 ± 0.04

7.4 Results

7.4.1 MNIST

The MNIST dataset [32] contains 60000 training and 10000 test samples. Unlike in [12], which decorrelates scattering coefficients with a discrete cosine transform prior to classification, we perform classification on the scattering coefficients directly. Kymatio [3] is used to produce the 2D scattering coefficients. Our implementation of separable scattering is implemented similarly to Kymatio, with PyTorch [57] as a backend for FFT convolutions and NN models.

Unless specified otherwise, all experiments have $\beta = \alpha = 2.5$. Scattering features are normalised prior to classification, according to the mean and SD calculated on the training set. No data augmentation is performed.

Due to its separability and non-angularly spaced filters, separable scattering does not perform as well compared to 2D scattering when using simple classifiers, such as LDA [20]. To illustrate that this performance discrepancy is not of significant consequence in a NN setting, we test performance on the full dataset utilising a simple architecture. The neural network architecture used has an input layer with 256 neurons, followed by two hidden layers with 128 and 64 neurons respectively. The output layer has 10 neurons - one for each digit. Input and hidden layers are followed by a batch norm layer [24] and ReLU activation function [1]. The output layer is followed by a softmax function. The Adam optimiser [30] is used with cross-entropy loss, a batch size of 256 and learning rate of $3 \cdot 10^{-5}$. 5000 of the 60000 training samples are reserved for validation and removed from the training set. Training is stopped when validation loss starts to increase. The NN model is initialised with random weights, and the experiment is repeated 50 times. Different invariance scales were tested, and the best results are reported in table 7.2. Tests are repeated for $l \in \{1, 2\}$ levels of scattering.

7.4.2 MedMNIST3D

The MedMNIST3D datasets are a subset of the MedMNIST dataset group [83], where each 3D dataset contains $28 \times 28 \times 28$ images with 2, 3 or 11 classes. Train, test and validation data partitions are provided by the authors. All datasets have on the order of 1000 training samples. We compare the baseline NN results provided in [83] with separable scattering features combined with a simple NN classifier.

We use NN classifier with an input layer containing 1024 neurons, followed by two hidden layers with 512 and 256 neurons respectively. Input and hidden layers layers are each followed by a batch norm layer [24] and a ReLU non-linearity [1]. For datasets with two classes, the output layer is a single neuron followed by a sigmoid activation function. For datasets with more than two classes, the output layer has a size equal to the number of classes, followed by a softmax activation. Binary cross-entropy loss are used for datasets with two classes, otherwise cross-entropy loss is used. The Adam optimiser [30] with a learning rate of $1 \cdot 10^{-5}$ is used. All other configuration parameters are identical to the model used for the MNIST dataset.

A single level of scattering coefficients are computed, with $Q = (2, 2, 2)$ and $d = (4, 4, 4)$. Many MedMNIST3D datasets tend to be unbalanced, implying that area under curve (AUC) is a more reliable metric to measure model performance. The results are shown in table 7.1, in which SOTA AUC performance is achieved for the Organ, Adrenal and Vessel datasets. Table 7.1 indicates the number of classes (c) for each of the datasets, with accuracy (ACC) also shown for reference. The performance of our method on non-SOTA results are comparable to the other baseline NN approaches presented in [83]. It is likely that better results can be achieved by the proposed method if the filters are made learnable and/or scattering parameters are uniquely optimised for each dataset.

8 Conclusion

Bibliography

[1] Agarap, A.F. (2019). Deep Learning using Rectified Linear Units (ReLU). 1803.08375.

[2] Andén, J. and Mallat, S. (2011). Multiscale Scattering for Audio Classification. In: *ISMIR*, pp. 657–662. Miami, Florida.

[3] Andreux, M., Angles, T., Exarchakis, G., Leonarduzzi, R., Rochette, G., Thiry, L., Zarka, J., Mallat, S., Andén, J., Belilovsky, E., Bruna, J., Lostanlen, V., Chaudhary, M., Hirn, M.J., Oyallon, E., Zhang, S., Cella, C. and Eickenberg, M. (2020). Kymatio: Scattering Transforms in Python. *Journal of Machine Learning Research*, vol. 21, no. 60, pp. 1–6.
Available at: http://jmlr.org/papers/v21/19-047.html

[4] Andén, J., Lostanlen, V. and Mallat, S. (2015 12). Joint Time-Frequency Scattering for Audio Classification.
Available at: http://arxiv.org/abs/1512.02125http://dx.doi.org/10.1109/MLSP.2015.7324385

[5] Andén, J., Lostanlen, V. and Mallat, S. (2018 7). Joint Time-Frequency Scattering.
Available at: http://arxiv.org/abs/1807.08869http://dx.doi.org/10.1109/TSP.2019.2918992

[6] Andén, J. and Mallat, S. (2011). Multiscale Scattering for Audio Classification.
Available at: http://www.cmap.

[7] Andén, J. and Mallat, S. (2013 4). Deep Scattering Spectrum.
Available at: http://arxiv.org/abs/1304.6763http://dx.doi.org/10.1109/TSP.2014.2326991

[8] Arts, L.P.A. and van den Broek, E.L. (2022 Jan). The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time–frequency analysis. *Nature Computational Science*, vol. 2, no. 1, pp. 47–58. ISSN 2662-8457.
Available at: https://doi.org/10.1038/s43588-021-00183-z

[9] Au, W.W.L., Pack, A.A., Lammers, M.O., Herman, L.M., Deakos, M.H. and Andrews, K. (2006). Acoustic properties of humpback whale songs. *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 1103–1110. `https://doi.org/10.1121/1.2211547`.
Available at: `https://doi.org/10.1121/1.2211547`

[10] Aulich, M.G., McCauley, R.D., Miller, B.S., Samaran, F., Giorli, G., Saunders, B.J. and Erbe, C. (2022 5). Seasonal Distribution of the Fin Whale (Balaenoptera physalus) in Antarctic and Australian Waters Based on Passive Acoustics. *Frontiers in Marine Science*, vol. 9. ISSN 22967745.

[11] Branch, T.A., Matsuoka, K. and Miyashita, T. (2004). Evidence for increases in Antarctic blue whales based on Bayesian modelling. *Marine Mammal Science*, vol. 20, pp. 726–754. ISSN 08240469.

[12] Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886.

[13] Civera, M. and Surace, C. (2022). An Application of Instantaneous Spectral Entropy for the Condition Monitoring of Wind Turbines. *Applied Sciences*, vol. 12, no. 3. ISSN 2076-3417.
Available at: `https://www.mdpi.com/2076-3417/12/3/1059`

[14] Civera, M. and Surace, C. (2022). Instantaneous Spectral Entropy: An Application for the Online Monitoring of Multi-Storey Frame Structures. *Buildings*, vol. 12, no. 3. ISSN 2075-5309.
Available at: `https://www.mdpi.com/2075-5309/12/3/310`

[15] Conrad, K. (2004). Probability distributions and maximum entropy. *Entropy*, vol. 6, no. 452, p. 10.

[16] Cornell University (nd). Band-limited energy detector (BLED) diagnostics.
Available at: `https://ravensoundsoftware.com/knowledge-base/diagnostics-for-band-limited-energy-detector/`

[17] Cramer, J.S. (2002). The origins of logistic regression.

[18] Eickenberg, M., Exarchakis, G., Hirn, M. and Mallat, S. (2017). Solid Harmonic Wavelet Scattering: Predicting Quantum Molecular Energy from Invariant Descrip-

tors of 3D Electronic Densities. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc.
Available at: `https://proceedings.neurips.cc/paper_files/paper/2017/file/72b386224056bf940cd5b01341f65e9d-Paper.pdf`

[19] Erbe, C. and King, A.R. (2008). Automatic detection of marine mammals using information entropy. *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2833–2840. `https://doi.org/10.1121/1.2982368`.
Available at: `https://doi.org/10.1121/1.2982368`

[20] Fisher, R.A. (1936). THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188. `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x`.
Available at: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x`

[21] Frazer, L. and Mercado, E. (2000). A sonar model for humpback whale song. *IEEE Journal of Oceanic Engineering*, vol. 25, no. 1, pp. 160–182.

[22] Ganapathi, S., Kumar, S.M. and Deivasigamani, M. (2016). Noise reduction in underwater acoustic signals for tropical and subtropical coastal waters. In: *2016 IEEE/OES China Ocean Acoustics (COA)*, pp. 1–6.

[23] George, G., Oommen, R.M., Shelly, S., Philipose, S.S. and Varghese, A.M. (2018). A Survey on Various Median Filtering Techniques For Removal of Impulse Noise From Digital Image. In: *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pp. 235–238.

[24] Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. `1502.03167`.

[25] jia Jiang, J., ran Bu, L., jie Duan, F., quan Wang, X., Liu, W., bo Sun, Z. and yue Li, C. (2019). Whistle detection and classification for whales based on convolutional neural networks. *Applied Acoustics*, vol. 150, pp. 169–178. ISSN 0003-682X.
Available at: `https://www.sciencedirect.com/science/article/pii/S0003682X18311241`

[26] Jiang, J., Guo, J., Fan, W. and Chen, Q. (2010). An improved adaptive wavelet

denoising method based on neighboring coefficients. In: *2010 8th World Congress on Intelligent Control and Automation*, pp. 2894–2898.

[27] Jin, L. and Cheng, J. (2010). An Improved Speech Endpoint Detection Based on Spectral Subtraction and Adaptive Sub-band Spectral Entropy. In: *2010 International Conference on Intelligent Computation Technology and Automation*, vol. 1, pp. 591–594.

[28] Justusson, B. (1981). Median filtering: Statistical properties. *Two-Dimensional Digital Signal Prcessing II*, pp. 161–196.

[29] Khemani, V., Azarian, M.H. and Pecht, M.G. (2022 2). Learnable Wavelet Scattering Networks: Applications to Fault Diagnosis of Analog Circuits and Rotating Machinery. *Electronics (Switzerland)*, vol. 11. ISSN 20799292.

[30] Kingma, D.P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. `1412.6980`.

[31] Kovacevic, J. and Vetterli, M. (1995). *Wavelets and Subband Coding*. Prentice-Hall signal processing series. Prentice Hall PTR. ISBN 9780130970800.
Available at: `https://books.google.co.za/books?id=4Qt_QgAACAAJ`

[32] LeCun, Y., Cortes, C., Burges, C. *et al.* (2010). MNIST handwritten digit database.

[33] Li, K., Sidorovskaia, N.A. and Tiemann, C.O. (2020). Model-based unsupervised clustering for distinguishing Cuvier's and Gervais' beaked whales in acoustic data. *Ecological Informatics*, vol. 58, p. 101094. ISSN 1574-9541.
Available at: `https://www.sciencedirect.com/science/article/pii/S1574954120300443`

[34] Lilly, J.M. and Olhede, S.C. (2012). Generalized Morse wavelets as a superfamily of analytic wavelets. *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 6036–6041.

[35] Liu, Z., Yao, G., Zhang, Q., Zhang, J. and Zeng, X. (2020). Wavelet Scattering Transform for ECG Beat Classification. *Computational and Mathematical Methods in Medicine*, vol. 2020. ISSN 17486718.

[36] Lu, T., Han, B. and Yu, F. (2021). Detection and classification of marine mammal sounds using AlexNet with transfer learning. *Ecological Informatics*, vol. 62, p. 101277. ISSN 1574-9541.

Available at: https://www.sciencedirect.com/science/article/pii/S1574954121000686

[37] MacQueen, J. (1967). Classification and analysis of multivariate observations. In: *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297.

[38] MacQueen, J. *et al.* (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297. Oakland, CA, USA.

[39] Mallat, S. (2008). *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. 3rd edn. Academic Press, Inc., USA. ISBN 0123743702.

[40] Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398.

[41] Mao, Y., He, Z., Ma, Z., Tang, X. and Wang, Z. (2019). Efficient convolution neural networks for object tracking using separable convolution and filter pruning. *IEEE Access*, vol. 7, pp. 106466–106474.

[42] MathWorks Inc. (2022). MATLAB Wavelet Toolbox (R2024a).
Available at: https://www.mathworks.com/help/wavelet/

[43] MATLAB (nd). Continuous 1-D wavelet transform.
Available at: https://www.mathworks.com/help/wavelet/ref/cwt.html

[44] Michau, G., Frusque, G. and Fink, O. (2022 2). Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119. ISSN 10916490.

[45] Miller, B.S., Calderan, S., Leaper, R., Miller, E.J., Širović, A., Stafford, K.M., Bell, E. and Double, M.C. (2021 12). Source Level of Antarctic Blue and Fin Whale Sounds Recorded on Sonobuoys Deployed in the Deep-Ocean Off Antarctica. *Frontiers in Marine Science*, vol. 8. ISSN 22967745. Casey2019 DS.

[46] Miller, B.S., Stafford, K.M., Opzeeland, I.V., Harris, D., Samaran, F., Širović, A., Buchan, S., Findlay, K., Balcazar, N., Nieukirk, S., Leroy, E.C., Aulich, M., Shabangu, F.W., Dziak, R.P., Lee, W.S. and Hong, J.K. (2021 12). An open access dataset for developing automated detectors of Antarctic baleen whale sounds and

performance evaluation of two commonly used detectors. *Scientific Reports*, vol. 11. ISSN 20452322.

[47] Mohammad, B. and McHugh, R. (2011). Automatic Detection and Characterization of Dispersive North Atlantic Right Whale Upcalls Recorded in a Shallow-Water Environment Using a Region-Based Active Contour Model. *IEEE Journal of Oceanic Engineering*, vol. 36, no. 3, pp. 431–440.

[48] Mohebbi-Kalkhoran, H. and Ratilal, P. (2022). Automatic detection and classification of baleen and toothed whale calls via machine learning approaches over instantaneous wide areas in the Gulf of Maine received on a coherent hydrophone array. *The Journal of the Acoustical Society of America*, vol. 152. ISSN 0001-4966.

[49] Nahak, S., Pathak, A. and Saha, G. (2023 8). Fragment-level classification of ECG arrhythmia using wavelet scattering transform. *Expert Systems with Applications*, vol. 224. ISSN 09574174.

[50] Nourani, V., Tootoonchi, R. and Andaryani, S. (2021). Investigation of climate, land cover and lake level pattern changes and interactions using remotely sensed data and wavelet analysis. *Ecological Informatics*, vol. 64, p. 101330. ISSN 1574-9541. Available at: `https://www.sciencedirect.com/science/article/pii/S1574954121001217`

[51] Ogundile, O., Usman, A., Babalola, O. and Versfeld, D. (2020). A hidden Markov model with selective time domain feature extraction to detect inshore Bryde's whale short pulse calls. *Ecological Informatics*, vol. 57, p. 101087. ISSN 1574-9541. Available at: `https://www.sciencedirect.com/science/article/pii/S1574954120300376`

[52] Ogundile, O., Usman, A., Babalola, O. and Versfeld, D. (2021). Dynamic mode decomposition: A feature extraction technique based hidden Markov model for detection of Mysticetes' vocalisations. *Ecological Informatics*, vol. 63, p. 101306. ISSN 1574-9541. Available at: `https://www.sciencedirect.com/science/article/pii/S1574954121000972`

[53] Olhede, S. and Walden, A. (2002). Generalized Morse wavelets. *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2661–2670.

[54] Ortega, J., Almanza-Ortega, N., Vega-Villalobos, A., Pazos-Rangel, R., Zavala-Diaz,

J.C. and Martínez-Rebollar, A. (2019 04). *The K-Means Algorithm Evolution.* ISBN 978-1-83880-333-9.

[55] PAMGuard (nd). PAMGuard core plug-in modules.
Available at: `https://www.pamguard.org/11_PluginModules.html`

[56] Parihar, D., Ghosh, R., Akula, A., Kumar, S. and Sardana, H. (2021). Seismic signal analysis for the characterisation of elephant movements in a forest environment. *Ecological Informatics*, vol. 64, p. 101329. ISSN 1574-9541.
Available at: `https://www.sciencedirect.com/science/article/pii/S1574954121001205`

[57] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc.
Available at: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`

[58] Powell, G.E. and Percival, I.C. (1979 nov). A spectral entropy method for distinguishing regular and irregular motion of Hamiltonian systems. *Journal of Physics A: Mathematical and General*, vol. 12, no. 11, pp. 2053–2071.
Available at: `https://doi.org/10.1088/0305-4470/12/11/017`

[59] Qiao, G., Ma, T., Liu, S., Zheng, N., Babar, Z. and Yin, Y. (2019). Spectral Entropy Based Dolphin Whistle Detection Algorithm and Its Possible Application for Biologically Inspired Communication. In: *OCEANS 2019 - Marseille*, pp. 1–6.

[60] Rademan, M.W., Versfeld, D.J. and du Preez, J.A. (2023 5). Soft-output signal detection for cetacean vocalizations using spectral entropy, k-means clustering and the continuous wavelet transform. *Ecological Informatics*, vol. 74. ISSN 15749541.

[61] Ravanelli, M. and Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. In: *2018 IEEE spoken language technology workshop (SLT)*, pp. 1021–1028. IEEE.

[62] Reddy, M.K., Keerthana, Y.M. and Alku, P. (2022). End-to-end Pathological Speech

Detection using Wavelet Scattering Network. *IEEE Signal Processing Letters.* ISSN 15582361.

[63] Ruiz, M.R., Rossi, C. and Esteban, J. (2023 11). Fin whale pulse detection with deep neural networks. *Ecological Informatics*, vol. 77, p. 102243. ISSN 15749541. Available at: `https://linkinghub.elsevier.com/retrieve/pii/S1574954123002728`

[64] Sadaiappan, B., Balakrishnan, P., Vishal, C.R., Vijayan, N.T., Subramanian, M. and Gauns, M.U. (2023 5). Applications of Machine Learning in Chemical and Biological Oceanography.

[65] Sattar, F. (2023 3). A New Acoustical Autonomous Method for Identifying Endangered Whale Calls: A Case Study of Blue Whale and Fin Whale. *Sensors*, vol. 23. ISSN 14248220.

[66] Schall, E. and Parcerisas, C. (2022 12). A Robust Method to Automatically Detect Fin Whale Acoustic Presence in Large and Diverse Passive Acoustic Datasets. *Journal of Marine Science and Engineering*, vol. 10. ISSN 20771312.

[67] Serra, O., Martins, F. and Padovese, L. (2020). Active contour-based detection of estuarine dolphin whistles in spectrogram images. *Ecological Informatics*, vol. 55, p. 101036. ISSN 1574-9541. Available at: `https://www.sciencedirect.com/science/article/pii/S1574954119303462`

[68] Shannon, C.E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55.

[69] Souli, S. and Lachiri, Z. (2018 1). Audio sounds classification using scattering features and support vectors machines for medical surveillance. *Applied Acoustics*, vol. 130, pp. 270–282. ISSN 1872910X.

[70] Sun, X., Liu, P., He, Z., Han, Y. and Su, B. (2022). Automatic classification of electrocardiogram signals based on transfer learning and continuous wavelet transform. *Ecological Informatics*, vol. 69, p. 101628. ISSN 1574-9541. Available at: `https://www.sciencedirect.com/science/article/pii/S1574954122000772`

[71] Tang, Y.Y., Lu, Y. and Yuan, H. (2015 5). Hyperspectral image classification based

on three-dimensional scattering wavelet transform. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 2467–2480. ISSN 01962892.

[72] Thaseen, I.S. and Kumar, C.A. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472.

[73] Trawicki, M.B. (2023). Automatic age recognition, call-type classification, and speaker identification of Zebra Finches (Taeniopygia guttata) using hidden Markov models (HMMs). *International Journal of Speech Technology*. ISSN 15728110.

[74] Urazghildiiev, I.R., Clark, C.W., Krein, T.P. and Parks, S.E. (2009). Detection and Recognition of North Atlantic Right Whale Contact Calls in the Presence of Ambient Noise. *IEEE Journal of Oceanic Engineering*, vol. 34, no. 3, pp. 358–368.

[75] Usman, A. and Versfeld, D. (2022). Detection of baleen whale species using kernel dynamic mode decomposition-based feature extraction with a hidden Markov model. *Ecological Informatics*, vol. 71, p. 101766. ISSN 1574-9541.
Available at: `https://www.sciencedirect.com/science/article/pii/S1574954122002163`

[76] Van Opzeeland, I., Samaran, F., Stafford, K.M., Findlay, K., Gedamke, J., Harris, D. and Miller, B.S. (2014). Towards collective circum-Antarctic passive acoustic monitoring: The Southern Ocean hydrophone network (SOHN). *Polarforschung*, vol. 83, no. 2, pp. 47–61.

[77] Van Trees, H., Tian, Z. and Bell, K. (2013). *Detection Estimation and Modulation Theory, Part I: Detection, Estimation, and Filtering Theory*. Detection Estimation and Modulation Theory. Wiley. ISBN 9781118539705.
Available at: `https://books.google.co.za/books?id=dnvaxqHDkbQC`

[78] Van Uffelen, L.J., Roth, E.H., Howe, B.M., Oleson, E.M. and Barkley, Y. (2017). A Seaglider-Integrated Digital Monitor for Bioacoustic Sensing. *IEEE Journal of Oceanic Engineering*, vol. 42, no. 4, pp. 800–807.

[79] van Wyk, J., du Preez, J. and Versfeld, J. (2022). Temporal separation of whale vocalizations from background oceanic noise using a power calculation. *Ecological Informatics*, vol. 69, p. 101627. ISSN 1574-9541.
Available at: `https://www.sciencedirect.com/science/article/pii/S1574954122000760`

[80] van Wyk, J., Versfeld, J. and du Preez, J. (2023 9). Detection of cetacean and fish sounds using normalized summation of harmonics and spectrogram masking. *JASA Express Letters*, vol. 3. ISSN 2691-1191.
Available at: `https://pubs.aip.org/jel/article/3/9/096002/2911548/Detection-of-cetacean-and-fish-sounds-using`

[81] Vickers, W., Milner, B., Risch, D. and Lee, R. (2021 6). Robust North Atlantic right whale detection using deep learning models for denoising. *The Journal of the Acoustical Society of America*, vol. 149, pp. 3797–3812. ISSN 0001-4966.

[82] Wu, M., Wang, D. and Brown, G. (2003). A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241.

[83] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H. and Ni, B. (2023). MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, vol. 10, no. 1, p. 41.