Wavelet Scattering Operators for Multiscale Processes: the Case Study of Marine Mammal Vocalizations

Alessandro Licciardi[1,2,*][0009-0006-8428-3408], Davide Carbone[1,2] [0000-0003-2859-6603], and Lamberto Rondoni[1,2] [0000-0002-4223-6279]

[1] Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[2] INFN, Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
{alessandro.licciardi, davide.carbone, lamberto.rondoni}@polito.it
(presenting author) (*corresponding author)

Abstract. Marine mammals vocalization pose challenges in understanding animal communication due to signal diversity and environmental factors. Researchers leverage machine learning (ML) to characterize vocalizations, monitor movements, and enhance comprehension of vocalizations. The Watkins Marine Mammal Sound Database, a crucial resource, spans decades but poses challenges for ML classification. Addressing interpretability issues in deep learning, we employ the Wavelet Scattering Transform (WST), offering invariance and stability. WST's application to this dataset contributes to understanding complex natural sounds. Our study presents a statistical analysis using WST, emphasizing class dispersion, demonstrating high accuracy compared to existing preprocessing methods.

Keywords: Signal Processing, Machine Learning, Invariant Representation Operators, Mammals Vocalization

1 Introduction

Marine mammals, encompassing species like whales, dolphins, and seals, are recognized for their intricate communication systems vital for survival and social interactions. However, understanding these systems is challenging due to the wide array of vocalizations, behaviors, and environmental factors involved [1][2]. Recently, researchers have turned to artificial intelligence (AI) and machine learning (ML) to analyze and interpret marine mammal communication patterns [3] [4]. Through AI and ML, researchers can classify vocalizations, monitor movements, and gain insights into behavior and social structures [5].

AI and ML play a crucial role in advancing our comprehension of whale vocalizations, involving tasks such as pattern recognition, diverse whale sound classification, and species-specific call identification. Automated detection and monitoring, combined with acoustic feature analysis, contribute to efficient tracking and population-level trend monitoring. Additionally, these technologies support ecological studies by correlating whale vocalizations with environmental factors, offering insights into behavioral patterns and social structures. Real-time monitoring provides early warning systems for conservation efforts, mitigating the impact of human activities on whale populations.

Integration with other data sources contributes to a comprehensive understanding of factors influencing whale communication. The collaborative synergy between marine biologists and data scientists holds promise for unraveling the complexities of whale vocalizations and supporting effective conservation strategies.
A pivotal resource for studying marine mammal communication is the Watkins Marine Mammal Sound Database (WMMD) [6]. This collection of recordings spans seven decades and encompasses various marine mammal species, holding immense historical and scientific value. The WMMD stands as a renowned reference dataset for studying vocalizations. However, it presents challenges for ML classification, including variability and complexity in vocalizations, environmental noise, and data scarcity for certain species.
State-of-the-art benchmarks rely on deep learning [7], yet the lack of interpretability in these architectures is a well-known issue. In this work, we address this challenge by employing the Wavelet Scattering Transform (WST) [8]. Viewed as the mathematical counterpart of Convolutional layers in deep networks, WST boasts invariance and stability properties concerning signal translation and deformation—qualities absent in standard preprocessing methods like the Short Time Fourier Transform [9]. Moreover, the scattering coefficient's structure proves valuable for providing a physical interpretation of multiscale processes, especially in the context of complex natural sounds [10].
The significance of the dataset extends beyond biology, representing a noteworthy example of a natural time series. Preprocessing and statistics of such objects pose a longstanding challenge in mathematical physics [11], from the advent of Fourier analysis to modern AI-based tools [12][13]. WST has been applied to various physical datasets, contributing to advancements in understanding multiscale and multifrequency processes challenging to address with standard Fourier techniques [14][15][16].
Summarizing the structure of this study:

- in Sec. **Theory** we present an extensive review of the theory of Mel spectrogram and Wavelet Scattering Transform (WST), two preprocessing methods for temporal series analysis. The result of this treatment is a summarized comparison between a standard tool in time-series analysis and WST. This could be potentially beneficial for the mathematical physics community, where the latter is still relatively unknown.
- in Sec. **Experiments and Results** we present a novel statistical analysis of the Watkins Marine Mammal Sound Database (WMMD) performed using WST, as opposed to Mel spectrogram. We firstly develop an original data preparation pipeline which is potentially usable for similar datasets. Then, we focus on intraclass dispersion after preprocessing, showing that the use of WST leads to less dispersed classes, critically outperforming the Mel spectrogram.

In order to ensure reproducibility of our study, the source code for our study is available at the public repository `https://github.com/alelicciardi99/marine_mammals`.

Related Works

Bioacoustics is a longstanding studied topic, positioned at the intersection between many fields, such as Biology, Zoology, Physics and Signal Processing [17]. The analysis of

the mammals vocalization is particularly important among all the possible species, as stressed in technical works such as [18], especially in the context of comparative studies [19]. We refer to the extended review [20] for a detailed analysis of the topic. For the concern of the present work, such dataset is mainly chosen as a qualitatively rich set of labelled temporal series; however, given the aforementioned premise, an extended analysis of such data is more than a toy example, as witnessed by the numerous recent works exploiting ML for classification tasks on WMMD dataset [7][21][22], and in the context of building new datasets [23]. Regarding Wavelet Scattering Transform [8]: its interpretability has been exploited in Physics Informed ML [24] and for neural network analysis [25]. The fields of application of such method are numerous; for the sake of the reader, we mention some regarding 1D signals: from music genre classification [26], to EEG [27] and ECG [28] in medicine, and in general to time-series analysis [29].

2 Theory

2.1 STFT and MEL Spectrogram

Spectrogram representation is one of the most common technique used in 1D signal representation theory, cfr. [30]. It provides information about the energy spectrum in the time-frequency domain (t, ω) and it is based on the *Short Time Fourier Transform* (STFT). Let us briefly recall the definition of STFT: we suppose that the time variable t is a positive real number, i.e. $t \in [0, +\infty)$. Let us fix a function $h(t)$ called *window function*, most common choices are *Hann window* or *Gaussian window*. Hann window, with support length $T > 0$, has the following form

$$h(t) = a \cos^2\left(\frac{\pi t}{T}\right) \mathbf{1}_{\{|t| \leq T/2\}}(t) \tag{1}$$

where $\mathbf{1}$ is the indicator function, while Gaussian window is a centered Gaussian function with amplitude a and spread σ, i.e.

$$h(t) = a \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{2}$$

As one can deduce from its name, a window function is typically selected to be localized in the time domain, and it can also have compact support, as shown in (1)

Definition 1 *For a given signal $x(t)$ and a fixed window function $h(t)$, the **Short Time Fourier Transform** is defined as*

$$\mathbf{STFT}\{x\}(t, \omega) = \int_{-\infty}^{\infty} x(\tau) h(\tau - t) e^{-i\omega\tau} \, d\tau. \tag{3}$$

Note that STFT is strictly related to the Fourier transform, due to the immediate relation

$$\mathbf{STFT}\{x\}(t, \omega) = \mathcal{F}\{x(\tau)h(\tau - t)\}(\omega) \tag{4}$$

i.e. the Fourier transform of the signal $x(\tau)$ multiplied by a moving window $h(\tau - t)$, for any $t > 0$. A trivial extension of the definition to the discrete time case is possible, by replacing the integral with an infinite summation. Given the STFT we recall the definition of spectrogram:

Definition 2 *For any $t > 0$ and $\omega > 0$, and for a chosen window $h(t)$ the **spectrogram** of a signal x is defined as the power spectrum of $x(\tau)h(\tau - t)$, i.e.*

$$|X(t,\omega)|^2 = |\mathbf{STFT}\{x\}(t,\omega)|^2\,, \tag{5}$$

The Mel spectrogram, a commonly utilized technique in audio signal processing [31], involves transforming the spectrogram introduced in Eq.(2) to a mel-frequency scale. This scale is designed to emulate the non-linear frequency perception of the human ear. Given a signal $x(t)$ and a chosen window function $h(t)$, the Mel spectrogram is defined as the power spectrum of the signal transformed to the mel-frequency scale. It offers a detailed representation of the signal's energy distribution across both time and mel-frequency variables. The initial step in computing the Mel spectrogram entails defining a set of triangular filters, commonly known as the Mel filter bank. These filters are spaced along the mel-frequency scale and overlap to capture the non-uniform nature of human hearing. This scaling choice is well-motivated for natural sounds and has been employed in preprocessing since its first application to the classification of labeled sounds [32].
Let N be the number of filters in the Mel filter bank, and $f(m)$ be the center frequency of the m-th filter. The Mel frequency m corresponding to a given frequency ω is computed using the formula:

$$m = 2595 \cdot \log_{10}\left(1 + \frac{\omega}{700}\right). \tag{6}$$

The center frequency $f(m)$ in Hertz corresponding to a mel frequency m is then given by:

$$f(m) = 700 \cdot (10^{m/2595} - 1). \tag{7}$$

Each triangular filter $H_m(\omega)$ is defined as

$$H_m(\omega) = \begin{cases} 0 & \text{if } \omega < f(m-1) \\ \frac{\omega - f(m-1)}{f(m) - f(m-1)} & \text{if } f(m-1) \le \omega \le f(m) \\ 1 - \frac{\omega - f(m)}{f(m+1) - f(m)} & \text{if } f(m) \le \omega \le f(m+1) \\ 0 & \text{if } \omega > f(m+1) \end{cases} \tag{8}$$

The Mel spectrogram is computed by summing the energy in each triangular filter bank applied to the magnitude of the Short Time Fourier Transform (STFT) of the signal:

$$\text{Mel Spectrogram}(t,m) = \sum_{k=0}^{N-1} |X(t,\omega_k)|^2 \cdot H_m(\omega_k), \tag{9}$$

where N is the number of frequency bins in the STFT, $X(t,\omega_k)$ is the STFT magnitude at time t and frequency bin ω_k, and $H_m(\omega_k)$ is the value of the m-th Mel filter at frequency bin ω_k. See Figure 1 for an example of Mel spectrogram.

2.2 Wavelet Scattering Transform

The Wavelet Scattering Transform (WST) [8] stands as a mathematical operator capable of yielding a stable and invariant representation for a given signal. Specifically,

when certain conditions are met [9], the resulting representation exhibits translation invariance, resistance to additive noise (i.e., it remains non-expansive), and stability to deformations. The latter property is formally expressed as Lipschitz-continuity under the influence of C^2-diffeomorphisms in its original derivation. Integrating a representation operator with these advantageous characteristics into a machine learning framework has the potential to significantly reduce the computational burden involved in training classification algorithms [33]. Since its derivation has been proposed very recently, in this section we provide an extended summary of definition and properties of WST for 1D signals (n.b. an extension to higher dimensions can be found for instance in [9]).
Let $\psi \in L^2(\mathbb{R}, dx)$ be a function, called *mother wavelet*, for a fixed scale factor $a > 1$ and for any $j \in \mathbb{Z}$, the $j-$th wavelet is defined as

$$\psi_{a^j}(t) = a^{-j}\psi(a^{-j}t) \tag{10}$$

Let $\lambda = a^j$ be the scaling-rotation operator, (10) can be redefined in terms of λ as

$$\psi_\lambda(t) = \lambda^{-1}\psi(\lambda^{-1}t)\,. \tag{11}$$

To build an intuitive connection with STFT, a is analogous to the width used for Hann or gaussian windows. About the choice of the mother wavelet, we refer in the following to Morlet wavelet [34]. In practice, in the usual definition of WST they define $Q \in \mathbb{N}$ such that $a = 2^{1/Q}$; this will play a role of hyperparameter.
In order to construct the wavelet scattering operator we fix the so-called depth $J \in \mathbb{N}$ and let $\Lambda_J = \{\lambda = a^j \,:\, |\lambda| = a^j \leq 2^J\}$ be the set of scattering indexes. Then, we define a low-pass filter using a gaussian $\phi_J = \mathcal{N}(0, \sigma)$ with $\sigma = 0.7$ and a path $p = (\lambda_1, \ldots . \lambda_m)$, $\lambda_i \in \Lambda_J$ which is any tuple of length m build using the scattering indexes; the wavelet scattering coefficient along a path p is defined as

$$S_J[p]x(u) = U[p]x \star \phi_J(t) = \int_{-\infty}^{\infty} U[p]x(\tau)\phi_J(t-\tau)d\tau\,, \tag{12}$$

where

$$U[p]x = U[\lambda_m]\ldots U[\lambda_1]x = |\ldots|x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|\ldots| \star \psi_{\lambda_m}|\,. \tag{13}$$

For the conducted experiments we couple *Morlet wavelets* with a Gaussian low-pass filter [34].
Before delving into the details of other beneficial properties of WST, let us clarify its definition in simpler terms. Through a straightforward combinatorial argument, it becomes evident that the longer the path, the greater the number of combinations of scattering indices. More precisely, this results in the characteristic tree structure illustrated in [26]. Each black dot in the diagram corresponds to a scattering coefficient, and it is customary to refer to the coefficient for a fixed m as the *m-order* scattering coefficient. Analogously to the spectrogram representation, it is common practice to visualize coefficients of the same order on a single heatmap, with time and j on the axes, as illustrated in Figure 1 as opposed to Mel spectrograms. It's important to note that J is another free hyperparameter, and its effect is to increase the cardinality of Λ_J, consequently influencing the number of coefficients per order. To practically infer the importance of the order, following [8] we introduce the path set up to length m,

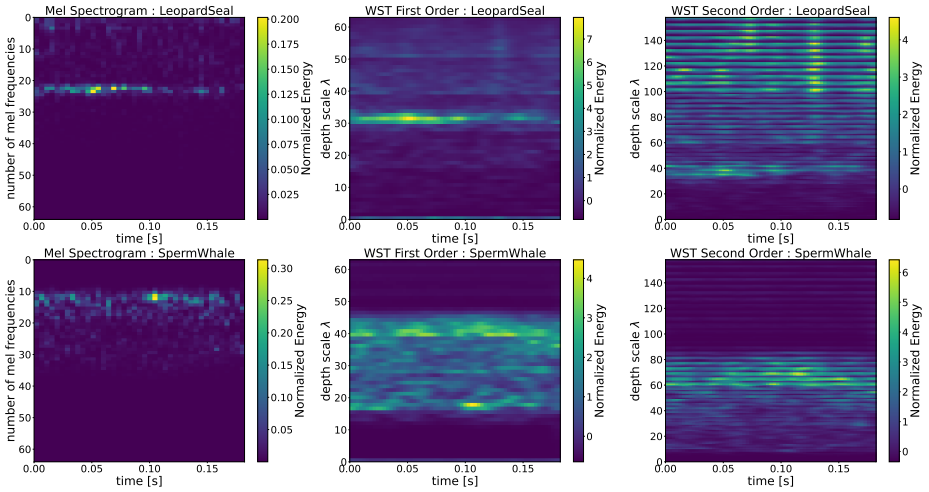

Fig. 1. Comparison of Mel spectrogram and WST of first and second order for vocalizations of two different species.

$\Lambda_J^m = \{(\lambda_1, \ldots, \lambda_m) : |\lambda_i| = a^j \leq 2^J\}$, it is possible to define the induced norm of the scattering operator over the set $\mathcal{P}_J = \bigcup \Lambda_J^m$, i.e.

$$\|S_J[\mathcal{P}_J]x\| = \sum_{p \in \mathcal{P}_J} \|S_J[p]x\| \tag{14}$$

where $\|\cdot\|$ stands for the L^2-norm. For fixed J and Q, and given the definition of Λ_J^m. One could be concerned about the depth requested in practice, but in different experiment [33] it has been showed that just 2 or 3 orders, also referred to as layers, of WST are sufficient to represent around 98% of the energy of the signal. Indeed the energy of each layer, i.e. $\|U[\Lambda_J^m]\|$, is empirically observed to rapidly converge to zero.

Invariance Properties

The principal target of the WST construction to define a non-linear operator which induces a metric that results **stable to additive noise** and **invariant to local deformations and translations**, in terms of *Lipschitz continuity*. We can use the induced norm (14) to verify this properties to hold for WST. The first is the stability to additive noise:

Proposition 1 *The norm induced by the WST is non-expansive, therefore the representation operator is stable to additive perturbation. Hence, for any signal $h \in L^2(\mathbb{R}, dx)$ and any perturbed version $h' = h + \epsilon$, it holds*

$$\|S_J[\mathcal{P}_J]x' - S_J[\mathcal{P}_J]x\| \leq \|x' - x\| = \|\epsilon\| . \tag{15}$$

Proposition 1 states that the wavelet scattering metric is stable to additive noise, therefore the error between two signals, say x and a perturbed version $x' = x + \epsilon$, can be controlled

in the transformed space, namely

$$\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \leq \|\epsilon\| .$$

Intuitively this property gains interest in many audio classification problems: small stochastic perturbations of the same signals do not affect much its representations in the scattering transformed space.
To then prove Lipschitz stability to deformations and translation invariance we need some preliminary result.

Proposition 2 *For any $x, x' \in L^2(\mathbb{R}, dx)$ and for any $J \in \mathbb{Z}$*

$$\|S_{J+1}[\mathcal{P}_{J+1}]x - S_{J+1}[\mathcal{P}_{J+1}]x'\| \leq \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\| \tag{16}$$

Here we provide a proof in the case $J \in \mathbb{Z}^+$, however the result can be easily generalized for any $J \in \mathbb{Z}$.

Proof. We recall the unpublished proof we presented in [35]. Let J be any positive integer. We recall that the set $\mathcal{P}_J = \bigcup_m \Lambda_J^m$ and $\mathcal{P}_{J+1} = \bigcup_m \Lambda_{J+1}^m$; since $\Lambda_J^m = \{(\lambda_1, \dots, \lambda_m) : |\lambda_i| \leq 2^J\}$ and $\Lambda_{J+1}^m = \{(\lambda_1, \dots, \lambda_m) : |\lambda_i| \leq 2^{J+1}\}$ it is easy to see that for any length m, it holds the inclusion $\Lambda_{J+1}^m \supset \Lambda_J^m$, and, moreover, it follows that $\mathcal{P}_{J+1} \supset \mathcal{P}_J$. Let us work on the first term of the inequality

$$\begin{aligned}
\|S_{J+1}[\mathcal{P}_{J+1}]x - S_{J+1}[\mathcal{P}_{J+1}]x'\|^2 &= \left\| \sum_{p \in \mathcal{P}_{J+1}} S_{J+1}[p]x - \sum_{p \in \mathcal{P}_{J+1}} S_{J+1}[p]x' \right\|^2 \\
&= \left\| \sum_{p \in \mathcal{P}_{J+1}} U[p]x \star \phi_{J+1} - \sum_{p \in \mathcal{P}_{J+1}} U[p]x' \star \phi_{J+1} \right\|^2 \\
&= \left\| \left(\sum_{p \in \mathcal{P}_{J+1}} U[p]x - U[p]x' \right) \star \phi_{J+1} \right\|^2 .
\end{aligned}$$

In order to complete the proof we need to state the following claim.

Claim. For any function $f \in L^2(\mathbb{R}, dx)$ and for any scale J the following holds true: $\|f \star \phi_{J+1}\| \leq \|f \star \phi_J\|$.

Such claim is quite immediate to prove, indeed, if we recall that $\phi_J(t) = 2^{-J}\phi(2^{-J}t)$, we easily see that for any ω its Fourier transform $\hat{\phi}_J(\omega) = \hat{\phi}(2^J\omega)$. Hence the scaling does not affect the amplitude of the low-pass filter in the frequency domain, but it just squeezes its support, indeed $Supp(\hat{\phi}_{J+1}) \subset Supp(\hat{\phi}_J)$, and it follows that $Supp(\hat{\phi}_{J+1}\hat{f}) \subset Supp(\hat{\phi}_J\hat{f})$. Since the norm of the convolution equals the norm of the product of the Fourier transforms, the previous inequality is proven.

Applying the claim, and the inclusion of the paths set we get

$$\begin{aligned}\left\|\left(\sum_{p\in\mathcal{P}_{J+1}} U[p]x - U[p]x'\right)\star\phi_{J+1}\right\|^2 &\leq \left\|\left(\sum_{p\in\mathcal{P}_{J}} U[p]x - U[p]x'\right)\star\phi_{J+1}\right\|^2 \\ &\leq \left\|\left(\sum_{p\in\mathcal{P}_{J}} U[p]x - U[p]x'\right)\star\phi_{J}\right\|^2 \\ &= \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\|^2 .\end{aligned}$$

And such result proves the thesis.

Let us point out that we just proved that the succession $\{\|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x'\|\}_J$ is decreasing, and, since its terms are non-negative, it converges to a non negative value l. We show that, provided that the wavelets are admissible, such limit is zero when x' is a translated version of x. Such result is properly stated by the following result.

Theorem 1 *Let us consider $x \in L^2(\mathbb{R}, dx)$ and a translated version $x_c(t) = x(t-c)$, for any constant $c \in \mathbb{R}$. Then for any choice of admissible wavelets we have*

$$\lim_{J\to\infty} \|S_J[\mathcal{P}_J]x - S_J[\mathcal{P}_J]x_c\| = 0 \tag{17}$$

This result is crucial; indeed, as the scale J increases, the distance between the scattering representation of x and the representation of any translation of it, x_c, approaches zero. It is worth emphasizing that this property of the wavelet scattering representation is shared with other classical tools, such as the Fourier modulus.
Now, let us delve into the concept of stability under the action of diffeomorphisms and displacements. In terms of preprocessing audio and image data, it is fundamental to construct a mathematical representation capable of extracting the most information, irrespective of shape deformations. For instance, in speech recognition, the signal of a spoken word may undergo significant changes when uttered by different individuals. However, from a mathematical standpoint, we can assert that the two signals are merely stretched versions of each other.
Let us recall the notation we previously introduced. We denote by $\tau : \mathbb{R} \to \mathbb{R}$ the displacement field having $\|\tau'\|_\infty < 1$, and by

$$\begin{aligned}L[\tau] : L^2(\mathbb{R}, dx) &\longrightarrow L^2(\mathbb{R}, dx) \\ x(t) &\longmapsto L[\tau]x(t) = x(t - \tau(t))\end{aligned}$$

the action of the diffeomorphism on a signal $x \in L^2(\mathbb{R}, dx)$. We are interested in bounding from above the quantity

$$\|S_J[\mathcal{P}_J]L[\tau]x - S_J[\mathcal{P}_J]x\| ,$$

hence we need to define the following auxiliary norm

$$\|U[\mathcal{P}_J]x\|_1 = \sum_{m\geq 0} \|U[\Lambda_J^m]x\| .$$

Theorem 2 *There exists a constant C such that every $x \in L^2(\mathbb{R}, dx)$ such that $\|U[\mathcal{P}_J]x\|_1 < \infty$ and every $\tau \in C^2(\mathbb{R})$ with $\|\tau'\|_\infty \leq 1/2$ satisfy*

$$\|S_J[\mathcal{P}_J]L[\tau]x - S_J[\mathcal{P}_J]x\| \leq C\|U[\mathcal{P}_J]x\|_1 K(\tau)\,, \tag{18}$$

where

$$K(\tau) := 2^{-J}\|\tau\|_\infty + \|\tau'\|_\infty \max\left(1, \log \frac{\sup_{t,t'} |\tau(t) - \tau(t')|}{\|\tau'\|_\infty}\right) + \|\tau''\|_\infty\,,$$

and for all $m \geq 0$, if $\mathcal{P}_{J,m} = \bigcup_{n<m} \Lambda_J^n$, then

$$\|S_J[\mathcal{P}_{J,m}]L[\tau]x - S_J[\mathcal{P}_{J,m}]x\| \leq Cm\|x\|K(\tau)\,. \tag{19}$$

This result shows that the distance between the representation of the displaced signal $L[\tau]x$ and x is controlled by a term proportional to $2^{-J}\|\tau\|_\infty$ and by another term proportional to the maximal elastic deformation $\|\tau'\|_\infty$. Moreover if x has a compact support the following result can be proven.

Corollary 1. *For any compact set $\Omega \subset \mathbb{R}$, there exists a constant C such that for any $x \in L^2(\mathbb{R}, dx)$ supported in Ω having $\|U[\mathcal{P}_J]x\|_1 < \infty$ and for all $\tau \in C^2(\mathbb{R})$ with $\|\tau'\|_\infty \leq 1/2$, then*

$$\|S_J[\mathcal{P}_{J,m}]L[\tau]x - S_J[\mathcal{P}_{J,m}]x\| \leq C\|U[\mathcal{P}_J]x\|_1(2^{-J}\|\tau\|_\infty + \|\tau'\|_\infty + \|\tau''\|_\infty) \tag{20}$$

The proofs of Theorem 2 and Corollary 1 can be found on [8] and [9]. Basically this corollary states that when the hypothesis are met, the windowed scattering transform metric is Lipschitz-continuous to the action of diffeomorphisms.

Example of Applications

For the reader's convenience, in this section, we elaborate further on the practical significance of the Wavelet Scattering Transform (WST) in contrast to the Mel spectrogram. As mentioned in the Related Works section, WST finds applications in diverse fields, ranging from music genre classification [26] to medical applications such as EEG [27] and ECG [28], as well as more broadly in time-series analysis [29].

A common descriptor of the signature property of WST is as follows: when two similar signals undergo transformation, their representations in the target space remain close. In contrast, any preprocessing based on the Short-Time Fourier Transform (STFT) does not exhibit such a property [26]. Furthermore, the sequence of modulus and convolution operations in WST mirrors the structure of convolutional neural networks, which have proven to be excellent tools in deep learning. Notably, the use of WST is not confined to 1D signals but can be extended to 2D (such as images [36]) or 3D data. In recent years, WST has gained popularity for extracting theoretical information about multiscale physical processes, including hydrodynamics [37] or field theory [38].

In the cited works and beyond, WST consistently outperforms STFT-based methods. On the theoretical side, its intrinsic multiscale nature introduces a novel approach to renormalization [16], while practical applications such as classification, clustering, or data analysis are significantly simplified, particularly when symmetries and invariance are crucial features of the studied process. In conclusion, the use of WST appears to be groundbreaking in various data science domains, particularly in computational physics, where symmetries and structure play a fundamental role.

3 Experiments and Results

In this section, our objective is to conduct an extensive comparison of data analysis between the Mel spectrogram and the Wavelet Scattering Transform (WST). It is important to emphasize that the pipeline for this comparison is entirely general and can potentially be extended to any temporal series. Notably, the application of WST as a theoretical tool is already prevalent in diverse fields such as cosmology [39] and field theory [40].
We posit that the utilization of WST as a substitute for Fourier analysis could be particularly valuable in applications involving both stochastic and deterministic chaotic dynamical systems. Natural datasets, such as the one under analysis, serve as robust benchmarks for evaluating the performance of theoretical models. The WST has the potential to assess the efficacy of predictions beyond the capabilities of Fourier analysis, making it a promising tool for gaining deeper insights into complex dynamical systems.

3.1 Data Preparation

The original dataset comprises 15,554 samples spanning a period of 70 years, meticulously collected by the Woods Hole Oceanic Institution, as documented by [6]. These samples encapsulate sounds professionally identified as originating from 51 marine mammal species. However, the dataset presents various challenges, including data heterogeneity due to collection with different sensors and class-wise imbalance, given that not all species are easily recorded. Following the methodology outlined in [21], we opted to exclude classes with fewer than 50 samples from the analysis, resulting in a refined dataset focused on 32 species.
Moreover, a thorough examination of the dataset uncovered the presence of over 300 repeated samples, some of which bore different labels. Consequently, we undertook the removal of duplicate signals, yielding a total of 14,767 signals for subsequent analysis. To mitigate issues arising from variations in signal lengths, we implemented a methodology to align and center them. We standardized the number of time stamps at 8,000, preserving the central points for time series longer than 8,000 and padding an equal number of zeros on both sides for shorter time series, as presented in Algorithm 1. In time series analysis, aligning signals ensures that corresponding observations across different signals represent the same time intervals, facilitating comparisons and analyses. Cutting involves removing excess data points beyond a specified length, while padding involves adding additional data points to reach the desired length. These steps are essential in order to consider signals of the same temporal length, removing biases and enabling a more robust and reliable analysis.

Algorithm 1 Align the signal with padding or cutting

Input: signal $x \in \mathbb{R}^K$, output signal length T

if $T \geq K$ **then** ▷ cut the original signal around its central time
  $t_c \leftarrow \lfloor K/2 \rfloor$
  $t_l \leftarrow T - t_c$
  $t_r \leftarrow T - t_l$
  $x' \leftarrow x[t_c - t_l : t_c + t_r]$ ▷ Slicing of the original signal
else ▷ center the original signal and then add zeros on both sides
  $\Delta \leftarrow K - T$
  $t_l \leftarrow \lfloor \Delta/2 \rfloor$
  $t_r \leftarrow \Delta - t_l$
  $x' \leftarrow (\mathbf{0}_{t_l}, x, \mathbf{0}_{t_r})$
end if

Output: transformed signal $x' \in \mathbb{R}^T$

Following the data preparation process, each signal underwent standardization ensuring a zero sample mean and unitary variance for each time series. The detailed steps for computing the standardized signal is described in Algorithm 2. Standardization mitigates issues related to varying scales and distributions in the input data. Standardizing signals before applying representation operators, e.g. WST or Mel spectrogram, does ensure consistent and accurate analysis of time series data. This operation brings signals to a common scale, thus enhancing the effectiveness of feature extraction techniques and improving the stability of resulting representations. Therefore representation operators are more able to capture significant underlying patterns more accurately.
We would like to stress that the described data preparation and preprocessing pipeline is robust and can easily handle the introduction of new audio samples. Moreover it could consistently be applied across different time series data types, regardless of their specific domain or characteristics.

Algorithm 2 Standardize Signal

Input: original signal $x \in \mathbb{R}^T$

$\hat{\mu} \leftarrow \dfrac{\sum_{t=1}^{T} x(t)}{T}$ ▷ Compute the sample mean

$\hat{\sigma}^2 \leftarrow \dfrac{\sum_{t=1}^{T} (x(t) - \hat{\mu})^2}{T-1}$ ▷ Compute the sample variance

$x'(t) \leftarrow \dfrac{x(t) - \hat{\mu}}{\hat{\sigma}}$ for all $t = 1, \ldots, T$ ▷ Standardization

Output: standardized signal $x' \in \mathbb{R}^T$

Algorithm 3 Data Preparation and Preprocessing

Input: original signal $x_i \in \mathbb{R}^K$, target signal length T, representation operator Φ, i.e. WST or Mel spectrogram

$x_i \leftarrow \text{Align}(x_i, T)$ ▷ use Algorithm 1 to center and cut, or pad, the signal up to length T

$x_i \leftarrow \text{Standardize}(x_i)$ ▷ use Algorithm 2 to standardize the signal

$\phi_i \leftarrow \Phi[x_i] \in \mathbb{R}^\Theta$ ▷ compute WST or Mel spectrogram

Output: transformed signal $\phi_i \in \mathbb{R}^\Theta$

order	representation	metric	CV
odontocetes	WST	L2	0.265
	WST	MSE	0.551
	Mel	L2	3.431
	Mel	MSE	18.532
mysticetes	WST	L2	0.281
	WST	MSE	0.571
	Mel	L2	2.495
	Mel	MSE	7.328
sirenians	WST	L2	0.700
	WST	MSE	1.479
	Mel	L2	2.109
	Mel	MSE	4.787
pinnipeds	WST	L2	0.303
	WST	MSE	0.648
	Mel	L2	2.428
	Mel	MSE	9.267

Table 1. Comparison of Coefficient of Variation (CV) for both preprocessing methods, and for L2 and MSE distance metrics. The 32 classes in the dataset are grouped in 4 *orders*, cfr. [6]. The reported CVs in the table are separately computed for each order with respect to the average representative. For every choice of metric (MSE or L2 distances) the CVs are smaller if computed using WST, that is the classes in the target space are less disperse.

3.2 Dispersion Analysis

The 32 species in the dataset belong to four different marine mammal orders, namely *mysticetes*, *odontocetes*, *pinnipeds* and *sirenians* [6]. In this analysis we aim to compare the WST and Mel spectrogram, that is the state-of-the art representation technique, by measuring the average dispersion of each marine mammal order that is present in the dataset.

Let $\mathcal{D} = \{(x_i(t), y_i),\ t = 1, \dots, T\}_{i=1}^N$ represent the data points— after applying the data-preparation process described in Algorithms 1 and 2— with labels $y_i \in \mathcal{I}_4$, the

corresponding order. Let

$$\tilde{\mathcal{D}} = \{(\phi_i(\theta), y_i),\ \theta = 1, \dots, \Theta\}_{i=1}^{N} \tag{21}$$

where $\phi_i = \Phi[x_i]$ for some preprocessing operator Φ, e.g. WST or Mel Spectrogram. Once Algorithm 3 had been performed, we estimated the dispersion of the representation computing the *coefficient of variation* (CV) of the pair-wise distance for each ϕ_i^j and the class-representant $\hat{\phi}^j$ of each class $j = 1, 2, 3, 4$. The j-th class representant is defined as

$$\hat{\phi}^j(\theta) = \frac{1}{n_j} \sum_{i=1}^{n_j} \phi_i^j(\theta) \tag{22}$$

for $\theta = 1, \dots, \Theta$, which is just the average of the element in the class. For a given distance metric d, in our experiment the Euclidean L^2 norm and the Mean Squared Error (MSE), the pair-wise distance for each element of class j

$$d_i^j = d(\phi_i^j, \hat{\phi}^j) \tag{23}$$

was computed for $i = 1, \dots, n_j$; then, the CV could be easily computed [41] as

$$CV^j = \frac{\hat{\sigma}^j}{|\hat{\mu}^j|} \tag{24}$$

where $\hat{\sigma}^j$ and $\hat{\mu}^j$ are the sample standard deviation and sample mean of $\{d_i^j\}_{i=1}^{n_j}$. A

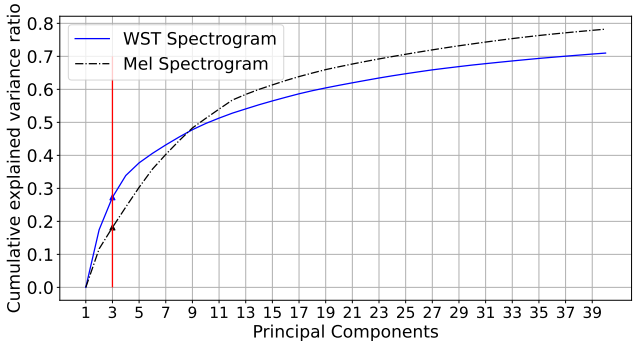

Fig. 2. Cumulative explained variance ratio for the two representations (WST in blue, Mel in black). For smaller number of principal components WST representation can capture more relevant information with respect to Mel spectrogram.

larger CV signifies a broader class dispersion, whereas a smaller CV indicates higher intra-cluster cohesion.

The second analysis we propose involves Principal Component Analysis (PCA) for dimensionality reduction [42]. Let $\Lambda : \mathbb{R}^{\Theta} \to \mathbb{R}^{p}$ denote the projection matrix obtained

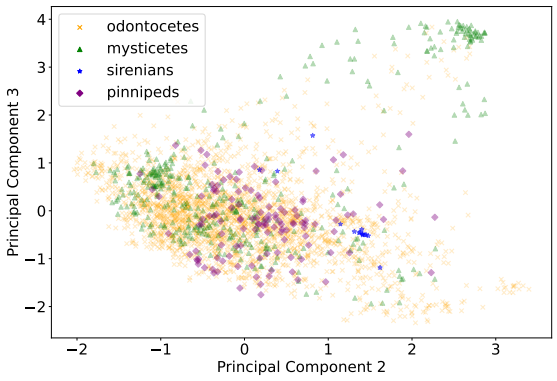

Fig. 3. Scatter plot for 20% of the data points in the dataset after WST computation and projection on first two principal components. We grouped the 32 classes in 4 macrogroups corresponding to *orders*. Each axis is normalized using a standard scaler.

with PCA over the entire preprocessed dataset $\tilde{\mathcal{D}}$, and $p \ll \Theta$ the number of principal components. The matrix Λ is obtained applying Singular Value Decomposition (SVD) to the covariance matrix of the feature vectors [43]. In a nutshell, PCA projection technique allows to transform the high-dimensional feature space $\mathbb{R}^{\Theta}$ into a low dimensional space $\mathbb{R}^{p}$, identifying the p eigenvectors, or principal components, that capture the most significant variations in the data. This approach simplifies the visualization and interpretation of similarities and dissimilarities within the high-dimensional feature space. The total explained variance ratio quantifies the proportion of variance in the original data explained by the p principal components, and PCA method ensures that the low-dimensional representation retains the most relevant information while reducing the dimensionality of the data. This quantity is strictly monotonically increasing with respect to the number of selected principal components p.

3.3 Discussion

Table 1 presents the first results of our analysis. We show the comparison between CV computed after preprocessing with WST and the Mel spectrogram: the use of WST causes a significant reduction in dispersion with respect to the representative element of each class, reaching nearly a 2-order-of-magnitude difference. This is a noteworthy result for statistical analysis, particularly in the identification of outliers: in practical applications the removal of potential outlying points is often performed using distance metrics from the mean and standard deviation [44]. Since WST tends to lead to less dispersed clusters in the representation space, the identification of points as outliers will be inherently less frequent. Indeed, in scenarios where the data preparation and measurement processes are presumed to be precise, such as sound recording, the elimination of an excessive amount of outliers would be contradictory. Moreover, for classes

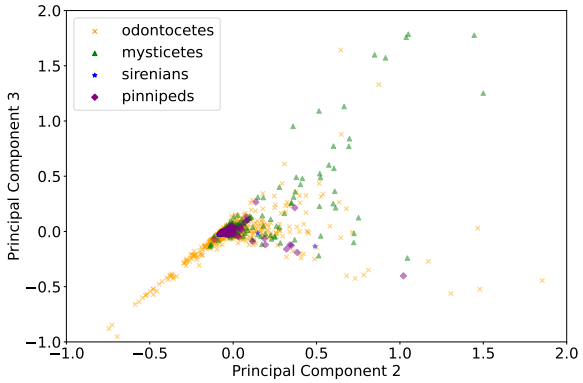

Fig. 4. Scatter plot for 20% of the data points in the dataset after Mel spectrogram computation and projection on first two principal components. We grouped the 32 classes in 4 macrogroups corresponding to *orders*. Each axis is normalized using a standard scaler.

with a limited number of examples, removing too many data points could reduce their size up to statistical irrelevance. Concluding, WST critically contributes to reduce these aforementioned issues.
Regarding PCA analysis, Figure 2 illustrates the cumulative explained variance per principal component for both preprocessing methods after PCA. Notably, for low-dimensional projections, WST outperforms the Mel spectrogram, almost doubling the explained variance around a 3-dimensional projection. This observation underscores the efficacy of WST in capturing relevant information in lower-dimensional spaces compared to the Mel spectrogram. Figures 3 and 4 show the projected points on two components, where each color correspond to a different order of mammals. In this case we focus on a qualitative consideration: the clouds of points appears to be very different. The plots are compatible with the quantitative analysis in Table 1: the Mel representatives overlap more in the core, but there is also the presence of many far isolated points which could be potentially misclassified as outliers. Moreover, Mel data points seem to be aligned along some direction, whereas WST seems to display a usual cloud-like plot. For further and quantitative consideration, a ML classification analysis is the following natural and necessary step. We will defer to a follow-up on this matter.

4 Conclusions

In this proceeding, we introduced a comprehensive pipeline for the analysis of 1D signals, with a particular focus on contrasting the standard Short-Time Fourier Transform (STFT) with the Wavelet Scattering Transform (WST). Theoretical comparison and review of these two procedures constitute our first contribution. It can serve as a simple instructional guide for a side-by-side comparison. However, the primary original result is the application of both methods to a significant dataset of mammal vocalizations,

showcasing the versatility of the WST in this domain. In comparison to the Short-Time Fourier Transform (STFT), the use of WST appears to be clearly beneficial and incurs no extra computational cost, as expected from previous applications in other contexts.
Beyond the originality of employing WST for analyzing mammal vocalizations, our objective was also to emphasize the potential of alternative preprocessing methods. The stability properties inherent in WST make it a preferable choice compared to standard methods, as quantitatively demonstrated in our dispersion analysis. In conclusion, this work strongly advocates for a possible wider application of WST in Bioacoustics and time-series analysis.
Regarding the possible limitations of the present work, our paper did not delve into a detailed grid search of the hyperparameters for WST. We acknowledge the importance of addressing this aspect in future research, even if the present configuration is sufficient to critically outperform the analysis based on Mel spectrogram. This avenue for future research would contribute to refining the application of Mel spectrum and WST in signal analysis, ensuring their efficacy across diverse datasets and signal characteristics.
Concerning data preparation before preprocessing (e.g., cutting, normalization, etc.), the pipeline is empirically motivated and necessary due to the vast heterogeneity of the dataset. While fairly standard, a more systematic analysis of its impact could be beneficial for extending the present work. However, any state-of-the-art analysis of WMMD dataset adopts a customized data preparation pipeline; the diversity of the dataset (in length of signals, amplitudes, sampling rate) makes an ad-hoc elaboration of the data points unavoidable.
We believe that natural follow-ups of the present work could focus on other similar dataset of animal vocalization, as for instance other marine mammals [45], primates [46], or birds [47]. In this sense, our contribution moves in the direction of utilizing modern data science techniques to enhance the understanding of animal vocalization and, consequently, the animal world. Back to Machine Learning applications, another possible future extension regards the use to our preprocessing and data preparation to improve state-of-the-art performances obtained in multiclass classification tasks. As discussed in the body, a representation leading to less disperse classes in the target space reduces the issue of outlier identification in the dataset; moreover, it possibly boosts separability of species, hence helping classification.

Acknowledgments

A.L., D.C. and L.R. worked under the auspices of Italian National Group of Mathematical Physics (GNFM) of INdAM. A.L. is part of the project PNRR-NGEU which has received funding from the MUR – DM 117/2023. D.C. was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

1. William A Watkins and Douglas Wartzok. Sensory biophysics of marine mammals. Marine Mammal Science, 1(3):219–260, 1985

2. Kathleen M Dudzinski, Jeanette A Thomas, and Justin D Gregg. Communication in marine mammals. In Encyclopedia of marine mammals, pages 260–269. Elsevier, 2009.
3. Suleman Mazhar, Tamaki Ura, and Rajendar Bahl. Vocalization based individual classification of humpback whales using support vector machine. In OCEANS 2007, pages 1–9. IEEE, 2007.
4. Peter C Bermant, Michael M Bronstein, Robert J Wood, Shane Gero, and David F Gruber. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Scientific reports, 9(1):12588, 2019.
5. Tom Mustill. How to Speak Whale: The Power and Wonder of Listening to Animals. Hachette UK, 2022.
6. Laela Sayigh, Mary Ann Daher, Julie Allen, Helen Gordon, Katherine Joyce, Claire Stuhlmann, and Peter Tyack. The watkins marine mammal sound database: an online, freely accessible resource. In Proceedings of Meetings on Acoustics, volume 27. AIP Publishing, 2016.
7. Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Scientific Reports, 13(1):22876, 2023.
8. Stéphane Mallat. Group invariant scattering. Communications on Pure and Applied Mathematics, 65(10):1331–1398, 2012.
9. Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. IEEE transactions on pattern analysis and machine intelligence, 35(8):1872–1886, 2013.
10. Fatemeh Khatami, Markus Wöhr, Heather L Read, and Monty A Escabí. Origins of scale invariance in vocalization sequences and speech. PLoS computational biology, 14(4):e1005996, 2018.
11. Santo Banerjee and Asit Saha. Nonlinear Dynamics and Applications: Proceedings of the ICNDA 2022. Springer Nature, 2022.
12. Tak-chung Fu. A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1):164–181, 2011.
13. Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. Information systems, 53:16–38, 2015.
14. Sihao Cheng, Yuan-Sen Ting, Brice Ménard, and Joan Bruna. A new approach to observational cosmology using the scattering transform. Monthly Notices of the Royal Astronomical Society, 499(4):5902–5914, 2020.
15. Joan Bruna and Stéphane Mallat. Multiscale sparse microcanonical models. Mathematical Statistics and Learning, 1(3):257–315, 2019.
16. Michael E Glinsky, Thomas W Moore, William E Lewis, Matthew R Weis, Christopher A Jennings, David J Ampleford, Patrick F Knapp, Eric C Harding, Matthew R Gomez, and Adam J Harvey-Thompson. Quantification of maglif morphology using the mallat scattering transformation. Physics of Plasmas, 27(11), 2020.
17. Elodie Ey and Julia Fischer. The "acoustic adaptation hypothesis"—a review of the evidence from birds, anurans and mammals. Bioacoustics, 19(1-2):21–48, 2009.
18. Rickye S Heffner and Henry E Heffner. Evolution of sound localization in mammals. In The evolutionary biology of hearing, pages 691–715. Springer, 1992.
19. William C Stebbins. The evolution of hearing in the mammals. In Comparative studies of hearing in vertebrates, pages 421–436. Springer, 1980.
20. Uwe Jürgens. The neural control of vocalization in mammals: a review. Journal of Voice, 23(1):1–10, 2009.
21. Nhat Hoang Bach, Le Ha Vu, Van Duc Nguyen, and Duy Phong Pham. Classifying marine mammals signal using cubic splines interpolation combining with triple loss variational autoencoder. Scientific Reports, 13(1):19984, 2023.
22. Jifeng Zhu, Wenyu Cai, Meiyan Zhang, and Yong Yang. Self-supervised denoising model based on deep audio prior using single noisy marine mammal sound sample. Applied Intelligence, 53(21):25697–25714, 2023.

23. Miles JG Parsons, Tzu-Hao Lin, T Aran Mooney, Christine Erbe, Francis Juanes, Marc Lammers, Songhai Li, Simon Linke, Audrey Looby, Sophie L Nedelec, et al. Sounding the call for a global library of underwater biological sounds. Frontiers in Ecology and Evolution, 10:39, 2022.
24. George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. Nature Reviews Physics, 3(6):422–440, 2021.
25. Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(5):726–742, 2021.
26. Joakim Andén and Stéphane Mallat. Deep scattering spectrum. IEEE Transactions on Signal Processing, 62(16):4114–4128, 2014.
27. Muhammad Zubair Ahmad, Awais Mehmood Kamboh, Sajid Saleem, and Amir Ali Khan. Mallat's scattering transform based anomaly sensing for detection of seizures in scalp eeg. IEEE Access, 5:16919–16929, 2017.
28. Zhishuai Liu, Guihua Yao, Qing Zhang, Junpu Zhang, Xueying Zeng, et al. Wavelet scattering transform for ecg beat classification. Computational and mathematical methods in medicine, 2020, 2020.
29. María B Arouxet, Verónica E Pastor, and Victoria Vampa. Using the wavelet transform for time series analysis. In Applications of Wavelet Multiresolution Analysis, pages 59–74. Springer, 2021.
30. Richard A Roberts and Clifford T Mullis. Digital signal processing. Addison-Wesley Longman Publishing Co., Inc., 1987.
31. Lawrence Rabiner and Ronald Schafer. Theory and applications of digital speech processing. Prentice Hall Press, 2010.
32. Chang-Hsing Lee, Chih-Hsun Chou, Chin-Chuan Han, and Ren-Zhuang Huang. Automatic recognition of animal vocalizations using averaged mfcc and linear discriminant analysis. pattern recognition letters, 27(2):93–101, 2006.
33. Joan Bruna. Scattering Representations for Recognition. Theses, Ecole Polytechnique X, February 2013.
34. Stéephane Mallat. A wavelet tour of signal processing. Elsevier, 1999.
35. Alessandro Licciardi. Wavelet scattering transform. mathematical analysis and applications to virgo gravitational waves data. Master's thesis, Politecnico di Torino, 2023.
36. Georgios Valogiannis and Cora Dvorkin. Going beyond the galaxy power spectrum: An analysis of boss data with wavelet scattering transforms. Physical Review D, 106(10):103509, 2022.
37. Andrew K Saydjari, Stephen KN Portillo, Zachary Slepian, Sule Kahraman, Blakesley Burkhart, and Douglas P Finkbeiner. Classification of magnetohydrodynamic simulations using wavelet scattering transforms. The Astrophysical Journal, 910(2):122, 2021.
38. Rudy Morel, Gaspar Rochette, Roberto Leonarduzzi, Jean-Philippe Bouchaud, and Stéphane Mallat. Scale dependencies and self-similar models with wavelet scattering spectra. Available at SSRN 4516767, 2023.
39. Georgios Valogiannis and Cora Dvorkin. Towards an optimal estimation of cosmological parameters with the wavelet scattering transform. Physical Review D, 105(10):103534, 2022.
40. Tanguy Marchand, Misaki Ozawa, Giulio Biroli, and Stéphane Mallat. Wavelet conditional renormalization group. arXiv preprint arXiv:2207.04941, 2022.
41. Hervé Abdi. Coefficient of variation. Encyclopedia of research design, 1(5), 2010.
42. Stéphane Tufféry. Data mining and statistics for decision making. John Wiley & Sons, 2011.
43. Gilbert Strang. Linear algebra and learning from data. SIAM, 2019.
44. Ihab F Ilyas and Xu Chu. Data cleaning. Morgan & Claypool, 2019.
45. Ayinde M Usman, Olayinka O Ogundile, and Daniel JJ Versfeld. Review of automatic detection and classification techniques for cetacean vocalization. Ieee Access, 8:105181–105206, 2020.

46. Daniel Romero-Mujalli, Tjard Bergmann, Axel Zimmermann, and Marina Scheumann. Utilizing deepsqueak for automatic detection and classification of mammalian vocalizations: a case study on primate vocalizations. Scientific reports, 11(1):24463, 2021.
47. Jiri Stastny, Michal Munk, and Lubos Juranek. Automatic bird species recognition based on birds vocalization. EURASIP Journal on Audio, Speech, and Music Processing, 2018(1):1–7, 2018.