Improving Classification of Marine Mammal Vocalizations Using Vision Transformers and Phase-Related Features

Dimitris N. Makropoulos[1,2,4], Panagiotis P. Filntisis[1,3,4], Aristides Prospathopoulos[2], Dimitris Kassis[2], Antigoni Tsiami[1,3], Petros Maragos[1,3,4]

[1]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[2] Hellenic Centre for Marine Research (HCMR), Institute of Oceanography, Anavyssos, Greece
[3] Robotics Institute, Athena Research Center, 15125 Maroussi, Greece
[4] HERON - Center of Excellence in Robotics, Athens, Greece

dsmakropoulos@mail.ntua.gr, {pfilntisis, antsiami, petros.maragos}@athenarc.gr,{aprosp, dkassis}@hcmr.gr

***Abstract*—In this study, we investigate the relative performance of Vision Transformers (ViTs) compared to convolution-based neural networks in categorizing vocalizations from a medium-sized marine mammal dataset. Additionally, we evaluate whether phase information derived from Fourier decomposition can serve as a complementary source of useful information to magnitude for classification tasks. Our study focuses on bioacoustics, utilizing the publicly available Watkins Marine Mammal Sound Database, which contains sound clips identified as originating from 32 marine mammal species. In this framework, we first trained convolution-based networks (ResNet-101, MobileNetV3) and Transformer-based networks (ViT B-16, Swin Transformer V2) on log-magnitude spectrograms (baseline models). In a second set of experiments, we incorporated the derivative of unwrapped phase from the Fourier representation into the magnitude spectrograms. Our results show that (a) Shifted Window (Swin) Transformers outperform MobileNets and achieve performance similar to ResNets while maintaining lower computational complexity and (b) the inclusion of phase derivatives into spectrograms leads to (i) consistently improved performance metrics across all biosignal categories for Swin Transformers and (ii) enhanced classification ability for both convolution-based and self-attention-based networks, particularly for the narrow-band frequency modulated (FM) whistles emitted by delphinids.**

***Index Terms*—Bioacoustics, Vision Transformers, Phase Derivative.**

I. Introduction

The identification of marine mammal vocalizations is essential for studying population movements, and ultimately for protecting endangered species. The complexity of the recognition task is enhanced by the ability of mammals to modify the acoustic properties of their calls during social interactions [1] or in response to vessel noise and other anthropogenic activities [2]. The field of computational bioacoustics has so far been dominated by convolution-based architectures, primarily focused on feature extraction for recognition or detection tasks [3]. In particular, Residual Networks (ResNets) which incorporate shortcut connections between layers to allow identity mapping [4], and MobileNets, which use depthwise separable convolutions [5], have proven to be popular backbones for classification tasks. However, the recent success made by Vision Transformers (ViTs) has raised the question of whether self-attention-based models can effectively replace Convolutional Neural Networks (CNNs), particularly when working with limited data. Nevertheless, ViTs have not yet been widely explored in bioacoustic tasks, as they typically rely on large quantities of training data [6]. In fact, collecting and annotating calls from endangered species, is a particularly difficult task and few extensive bioacoustic databases are publicly available. In this study, we evaluate both typical and hierarchical Transformers, with a particular focus on **S**hifted **Win**dow (Swin) Transformers, introduced in [7]. Swin Transformers limit self-attention computation to non-overlapping local windows, while also introducing a window partitioning approach for cross-window connections [6]. In this

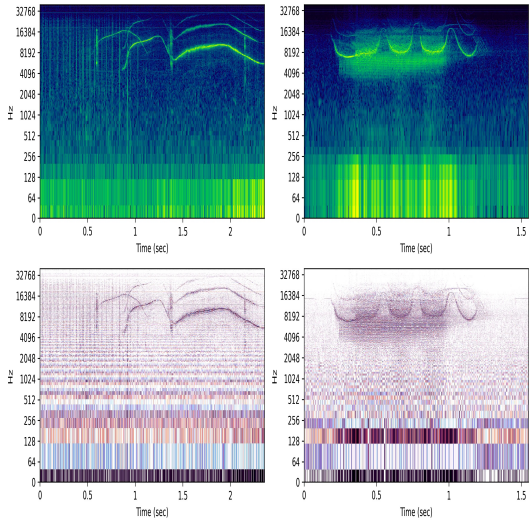


Fig. 1: Log-Magnitude spectrograms of Common dolphins' (*Delphinus Delphis*) calls (top-left) vs Atlantic spotted dolphins' (*Stenella frontalis*) vocalizations (top-right) and their respective representations incorporating phase derivative (bottom-left and bottom-right). A cyclic colormap ('twilight') is selected to represent phase derivative; see Section III.

Acknowledgment—"This project is funded by the European Union under Horizon Europe (grant No. 101136568 - project HERON)." Funded by the European Union

way they achieve both linear computational complexity and global information flow. Moreover, they construct hierarchical feature maps and have the flexibility to model at various scales [7], that enhances their learning capacity in deeper layers.

Additionally, this study evaluates whether incorporating phase information, alongside magnitude, into spectrograms improves feature extraction and enhances the recognition of marine mammal vocalizations. Typically, phase is often eliminated when, during an intermediate step of spectrogram visualization, windowed Fourier Transform is squared. However, we know that the phase array of the Discrete Fourier Transform (DFT) preserves important features of a signal [8]. Indeed, in various contexts, such as signals of speech and images, phase information from the Fourier representation is considered to carry necessary information for efficient reconstruction [8], [9]. Moreover, studies have shown that a phase-only-synthesis in speech or image signals -where a signal is reconstructed using the phase from its Fourier representation combined with a unity magnitude- can result in a high degree of intelligibility [8] as it efficiently preserves the location of patterns of the original signal. Motivated by the fact that narrow-band whistles produced by delphinids [10], [11], [12] contain nonlinear structures such as frequency modulation (FM), we further investigate whether incorporating phase derivative into spectrograms could improve the classification performance of FM bioacoustic patterns. This hypothesis is substantiated by the experimental results in this study. Another motivation for exploring the FM structure in marine mammal vocalizations is the fact that AM-FM structures have been discovered in human speech [13],[14]; the AM-FM information was captured using the Teaker-Kaiser energy operator (TKEO) [15]. Afterwards, the TKEO was used in [16] to detect marine biosounds.

II. Related Work

The raw source of information in bioacoustics consists of 1-D waveforms representing pressure over time. In most studies, these 1D-sequences are converted into log or mel-scaled magnitude spectrograms, and the recognition or detection of vocalizations is typically based on identifying intensity patterns determined by the magnitude of the DFT in the time-frequency plane [17], [18], [19]. The importance of phase derived from Fourier Transform in efficiently solving problems on image and speech reconstruction has been described and demonstrated in [8]. The idea to utilize time-frequency representations combining both magnitude and the derivative of the phase in the same plots comes from [20] and is based on the fact that the latter creates solid continuous lines in plots of constant-Q transforms (CQT) for harmonics of a consistent frequency. Regarding networks architectures, Transformer-based models, which perform global self-attention [21] to learn long-term time dependencies between input and output sequences [22], are used systematically for sound and music classification [23] or bioacoustic event detection [24]. For classification and detection tasks involving subsets of the specific Watkins Marine Mammals Sound Database (WMMSD) used in our study, several learning algorithms have been proposed: As demonstrated in [25], Support Vector Machines (SVMs) and VGGish networks trained on Mel Frequency Cepstral Coefficients (MFCCs) and power spectrograms, respectively, achieved accuracies of approximately 0.87 and 0.847 on a 31-class classification task using the 'Best of' cuts section, under resampling of recordings to 44.1 kHz. In [26], a ResNet is trained on single-channel spectrograms achieving an F1-score of 0.867 with an area under the curve (AUC) of 0.9281 outperforming a multi-channel implementation on a 32-class categorizing task. In [27], a network based on Efficient-B1 pre-trained on bird-song vocalizations, with its weights kept frozen, was adapted by training a single linear probe on various bioacoustic datasets; on a subset of WMMSD, this model achieved an accuracy of 0.83 and an AUC of 0.98. Finally, in [28], the authors use ResNets in parallel (WhaleNet) and explore the use of the Wavelet Scattering Transform and Mel-spectrograms to extract features, from a balanced 32-class subset of the 'all cuts' section of the WMMSD, reporting an accuracy of 97.61% and an average F1-score of 93.8%.

III. Materials and Methods

Origin of data and preprocessing

Our study utilizes the publicly available 'best of' cuts section of the Watkins Marine Mammal Sound Database, which contains approximately 1.700 sound clips, representing calls from 32 marine mammal species. These sounds were recorded over a span of 70 years using hydrophones of varying technologies, at different sampling rates, and in environments with diverse background noise levels. Each recording is associated with a metadata file providing a brief description and an evaluation of the bioacoustic events, present in both the background and foreground soundscapes. However, the dataset is imbalanced: some classes are underrepresented, with only a few short audio recordings (e.g., the Minke whale class contains 17 sound files, each lasting 1 to 2 seconds) while other classes, are relatively overrepresented (e.g., the Sperm whale class has over an hour of total recording time). For data preprocessing, all recordings were resampled to 44.1 kHz. Additionally, we excluded audio files containing overlapping calls, where individuals from different species vocalized simultaneously. Since audio clips vary widely in length -ranging from less than a second to several minutes- we split the recordings into 5-second sub-clips. For shorter audio files, that contained mostly repetitive whistle patterns, iterative padding was applied to extend them to 5 seconds, to preserve the temporal integrity of the sequence.

Visualization of Time-Frequency Representations

Spectrograms are obtained by framing and windowing the signal of pressure, then computing the DFT over each windowed segment. For the visualization of log-scaled magnitude spectrograms, we apply the windowed Fourier transform to each 5-seconds sequence, using a Hamming window of size 1024, with 50% overlap. The intensity of the power spectrum is represented by a perceptually uniform sequential colormap ('viridis'). In order to incorporate the

TABLE I: Main Results: Mean values and 95% confidence interval and state-of-the-art benchmarks (first four rows)

Inputs	Models (Parameters, GFLOPS)	Accuracy	Avg F1-score	Wgt. F1-score	Avg Precision	Wgt. Precision	Avg Recall	Wgt. Recall
MFCCs (mean, stdev, min, max)	SVM [25]	87%	-	-	-	-	-	-
Magnitude Spectrogram	ResNet [26]	85.43%	85.1%	-	85.99%	-	85.43%	-
Magnitude Spectrogram	EfficientNet-B1 (Perch) [27]	83%	-	-	-	-	-	-
Mel Spectrogram - Wavelet Scattering Transform	ResNets in Parallel (WhaleNet) [28]	97.60%	93.81%	97.61%	-	-	-	-
MFCCs (mean, stdev, min, max) + spectral (centroid, bandwidth and rolloff) + zero crossing rate, rms	SVM	95.8% ± 0.45%	91.6% ± 0.86%	95.7% ± 0.45%	94.0% ± 0.66%	95.9% ± 0.41%	90.5% ± 1.03%	95.7% ± 0.45%
Magnitude Spectrogram (baseline)	ResNet-101 ($44.5M, 7.8$ GFLOPS)	97.0% ± 0.39%	94.2% ± 0.87%	97.0% ± 0.40%	95.2% ± 0.71%	97.2% ± 0.40%	93.9% ± 0.89%	97.0% ± 0.39%
	MobileNetV3 ($5.5M, 0.22$ GFLOPS)	96.4% ± 0.28%	93.3% ± 0.70%	96.4% ± 0.28%	94.1% ± 0.77%	96.6% ± 0.28%	93.2% ± 0.72%	96.4% ± 0.28%
	ViT B16 ($86.6M, 17.56$ GFLOPS)	96.1% ± 0.32%	91.9% ± 0.92%	96.0% ± 0.33%	93.8% ± 0.73%	96.2% ± 0.29%	91.1% ± 0.92%	96.1% ± 0.32%
	Swin Transformer V2 ($28.4M, 5.94$ GFLOPS)	96.9% ± 0.32%	93.9% ± 0.93%	96.9% ± 0.33%	94.9% ± 0.50%	97.1% ± 0.30%	93.8% ± 1.02%	96.9% ± 0.33%
Magnitude and Phase Derivative Spectrogram	**ResNet-101** ($44.5M, 7.8$ **GFLOPS**)	**97.6% ± 0.33%**	**94.8% ± 1.12%**	**97.6% ± 0.36%**	**95.9% ± 1.09%**	**97.7% ± 0.33%**	**94.5% ± 1.07%**	**97.6% ± 0.31%**
	MobileNetV3 ($5.5M, 0.22$ GFLOPS)	96.9% ± 0.27%	93.3% ± 0.74%	96.9% ± 0.27%	94.4% ± 0.74%	97.0% ± 0.27%	93.0% ± 0.79%	96.9% ± 0.27%
	ViT B16 ($86.6M, 17.56$ GFLOPS)	97.4% ± 0.31%	94.0% ± 1.20%	97.3% ± 0.33%	95.0% ± 1.34%	97.4% ± 0.34%	93.6% ± 1.09%	97.4% ± 0.33%
	Swin Transformer V2 ($28.4M, 5.94$ **GFLOPS**)	**97.5% ± 0.43%**	**94.9% ± 1.09%**	**97.5% ± 0.43%**	**95.7% ± 0.91%**	**97.6% ± 0.41%**	**94.8% ± 1.01%**	**97.5% ± 0.44%**

phase derivative into the T-F representation, we decompose the complex-valued spectrogram F into the product of two matrices: the amplitude spectrum $|F(\omega)|$ and the phase spectrum $\theta(\omega)$, such that $F(\omega) = |F(\omega)|e^{j\theta(\omega)}$. A phase unwrapping operation follows to correct discontinuities between consecutive elements of the phase vector, as described in [29]. Subsequently, the time derivative of the unrolled phase angle is computed. Finally, we visualize both components of the energy distribution across frequencies in a single spectrogram : the log-magnitude of the power spectrum and the derivative of unrolled phase. The intensity of lines is proportional to the amplitude of power spectrum, while the phase derivative, defined within the range of values $[-\pi, \pi]$, is visually represented using a cyclic colormap ('twilight'). A cyclic colormap starts and ends on the same color, increasing monotonically from start to a symmetric point in the middle and inversely from middle to end. An image resolution of 224×224 is assumed for all spectrograms, which are generated using the Librosa Python Library. In Fig.1 we plot log-spectrograms alongside time-frequency representations combining power spectrum amplitude and phase derivative for calls of Common dolphins' (*Delphinus Delphis*) whistles or/and clicks (top-left, bottom-left) versus Atlantic spotted dolphins' (*Stenella frontalis*) whistles (top-right, bottom-right).

Design of Neural Networks

At the outset of this study, we confirmed that an SVM classifier based on MFCCs, is effective in categorizing biosignals, provided an exhaustive search is conducted across a broad range of SVM hyper-parameters (regularization parameter C and kernel coefficient gamma) of the radial basis function (RBF) kernel. For the convolution-based experiments in our study, we selected ResNet-101 ($\sim 44.5M$ parameters) and MobileNetV3 ($\sim 5M$ parameters) architectures. As an initial step to evaluate their efficiency, we used pretrained versions of these models on the ImageNet-1K dataset, removing the fully connected layer and replacing it with an SVM classifier. Pretrained-models produced embeddings that were then passed to the SVM classifier which was then trained on these extracted features from either the ResNet or the MobileNet, without any fine-tuning of the backbone networks. Our preliminary results (93.3% and 88.7% accuracy for a MobileNetV3-Large and a ResNet respectively) indicated that frozen, pretrained convolution-based networks on image datasets (such as ImageNet) encode transferable knowledge useful for recognizing patterns in audio-spectrograms. In subsequent experiments, we fine-tuned MobileNetV3 and ResNet-101 networks, by using a fully connected layer instead of an SVM to further optimize performance. For the Transformer-based experiments, we selected a ViT-B16 ($\sim 86.6M$ parameters), and a Swin Transformer V2-tiny architecture ($\sim 28.4M$ parameters), both pretrained on the ImageNet-1K dataset. This choice is motivated by the fact that the Swin V2 model, introduced in [30], adopts a scaled cosine attention mechanism along with a residual post-normalization technique, a method that significantly reduces the average feature variance in deeper layers, improving training stability and accuracy [30].

IV. Experiments, Results and Discussion

In our experimental setup, the dataset was divided into two subsets: 60% for training and 40% for testing. We used $K = 10$ different partitions of the dataset, conducting an equal number of experiments for each partition. The generalization ability of the classifier was evaluated on test sets comprising 1563 biosignals. Standard performance metrics, such as accuracy, precision, recall, and F1-score were calculated as the mean of K experiments, along with a 95% confidence interval. For the SVM model, we extracted the first 20 cepstral coefficients per frame from each audio clip and calculated four summary statistics -mean, standard deviation, min, and max- of each MFCC dimension over time, as in [25], along with mean values of rms, spectral bandwidth, spectral centroid, roll-off and zero-crossing rate. These features were averaged across all frames to construct an 85x1 feature vector for each vocalization. Results for the best-performing model are presented in Table 1. For all deep networks, models were trained for 100 epochs with a batch size of 64. An Adam optimizer was used with an initial learning rate of 10^{-3}, decaying by a factor of 0.1 every 30 epochs through a StepLR scheduler. Categorical cross entropy was employed as a loss function during the optimization process.

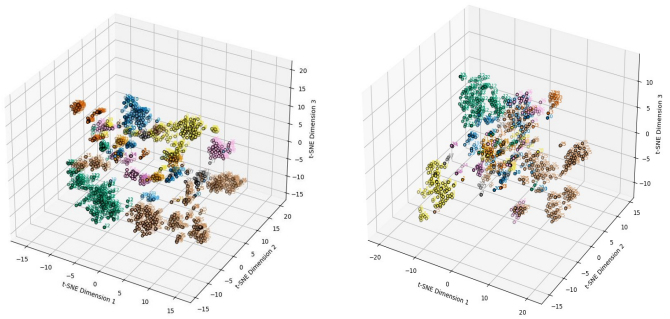
Table 1 presents the main performance metrics for all

models, with the best-performing networks emphasized. The following observations can be made: (a1) Swin Transformers and ResNets present similar performance within the margin of statistical error (94.9% $\pm$ 1.09% average F1 score versus 94.8% $\pm$ 1.13% respectively), while Swin Transformers offer an advantage in terms of computational complexity ($\sim$ 5.94 GFLOPS required for a forward pass for Swin Transformers vs $\sim$ 7.8 GFLOPS for ResNet101); (a2) Both models outperform MobileNets, which achieve a 93.3% average F1-score and 96.9% accuracy; (a3) Although ViT-B16 achieves comparable performance to hierarchical ViTs, its higher computational complexity ($\sim$ 17.56 GFLOPS) renders the model less attractive; (a4) The relative deviation between the average F1-score and its weighted metric is attributed to dataset imbalance. (b) Both convolution-based and Transformer-based algorithms outperform SVMs, although SVMs still demonstrate a remarkable speed-accuracy trade-off, a finding that is consistent with observations in [25]. (c) Incorporating the phase derivative into spectrograms systematically improves model performance: the average F1-score and accuracy of Swin Transformer increase by 1.0% and 0.6%, respectively; corresponding improvements are 0.6% and 0.6% for ResNet-101, 2.1% and 1.3% for ViT-B16, while MobileNet shows minor gains. (d) Finally, we include in Table 1 several state-of-the-art benchmarks (first four rows) reported on different subsets of this dataset, although direct comparisons with these works are not possible due to differences in evaluation protocols. Specifically, in [25], only 31 classes out of 32 classes were selected and in [28] a balanced subset of 32 classes from 'all cuts' was utilized whereas we use the full 'best-of' cut dataset and in [26], a lower sampling rate of 22.050Hz was used, and data augmentation was performed.

TABLE II: Average F1 score per vocalization class

Inputs	Models	FM Whistles and Clicks	Vocalizations from Whales	Vocalizations from Seals
Magnitude Spectrogram	ResNet-101	94.5%	95.0%	92.1%
	MobileNetV3	93.0%	95.0%	91.5%
	ViT B-16	92.4%	92.2%	89.4%
	Swin Transformer V2	93.9%	95.0%	92.2%
Magnitude and Phase Derivative Spectrogram	ResNet-101	95.4%	95.4%	92.1%
	MobileNetV3	94.4%	94.2%	89.1%
	ViT B-16	95.3%	94.7%	89.4%
	Swin Transformer V2	**95.1%**	**95.7%**	**92.9%**

Table 2 presents the mean evaluation average F1 scores for three broad categories of vocalizations in the same dataset: FM whistles and clicks from delphinids, and vocalizations from whales and seals. Sperm whale calls are excluded, as they consist of broadband impulsive signals (clicks), for which the incorporation of phase has no impact on classification performance. Key observations are: (a) The addition of the phase derivative improves classification metrics for FM whistles produced by delphinids across all networks, supporting the intuition that phase derivative adds useful information complementary to magnitude for recognition tasks. (b) Overall, Transformer-based models show a greater ability to leverage phase-related information compared to convolutional networks. In particular, Swin Transformers achieve better results across all categories of calls, followed by ViT-B16 which exhibit the highest positive deviations (+2% for delphinid vocalizations, +2.5% for calls from whales vocalizations). In contrast, while ResNet-101 also benefits from the inclusion of phase-related features, MobileNets show mixed results - achieving higher performance with log-power spectrograms for non-whistle classes.



(a) Swin's Transformer features (b) MobileNet's features

Fig. 2: t-SNE visualization of extracted feature vectors into a 3D feature space.

Finally, in Fig.2, we apply t-distributed Stochastic Neighbor Embedding (t-SNE), to project the high-dimensional feature space onto a lower-dimensional plane. For each architecture, we train both a Swin Transformer and a MobileNetV3-Large on the same training set. After training, the classification head is removed, and a forward pass is performed on the validation set to extract feature embeddings. For the Transformer, features of dimension 768 are reduced to 3, and for the MobileNet, features of dimension 1280 are similarly reduced to 3 using t-SNE. These features are then visualized in a Euclidean plane. Fig.2 illustrates the superior capacity of the Swin Transformer to partition the feature space: clusters corresponding to different species are more clearly separated and distinct compared to those produced by the MobileNet.

V. Conclusion

In this study, we show that hierarchical ViTs effectively address common challenges associated with Transformer architectures, such as high computational complexity and sensitivity to dataset size. Among all evaluated models, they achieved the most favorable speed-accuracy trade-off for classifying marine mammal vocalizations on a mid-sized dataset. Furthermore, our findings show that both convolution-based and self-attention-based networks benefit from the incorporation of phase derivatives into spectrogram representations, with more pronounced improvements observed in Transformer-based architectures, across all vocalization classes. These findings suggest that phase-augmented representations can enhance classification performance, especially for frequency-modulated (FM) sounds produced by delphinids. Future work could extend this analysis to other bioacoustic datasets and explore efficient methods for incorporating phase-related features to modern challenges such as vocalization detection and few-shot learning.

REFERENCES

[1] Z. Song, A. Mooney, L. Quakenbush, R. Hobbs, E. Gaglione, C. Goertz, and M. Castellote, "Variability of echolocation clicks in beluga whales (delphinapterus leucas) within shallow waters," *Aquatic Mammals*, 2023.

[2] E. K. Skarsoulis, G. S. Piperakis, E. Orfanakis, P. Papadakis, D. Pavlidi, M. A. Kalogerakis, P. Alexiadou, and A. Frantzis, "A real-time acoustic observatory for sperm-whale localization in the eastern mediterranean sea," *Frontiers in Marine Science*, vol. 9, 2022.

[3] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PEERJ*, vol. 10, 2022.

[4] H. Kaiming, X. Zhang, R. Shaoqing, and J. Sun, "Deep Residual Learning for Image Recognition," *"Proc. IEEE/CVF Conf. Computer Vision & Pattern Recognition (CVPR)"*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594692

[5] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *Computing Research Repository (CoRR)*, 2017.

[6] K. Han, Y. Wang, H. Chen, J. Chen, X.and Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 87–110, 2023.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[8] A. Oppenheim and J. Lim, "The Importance of Phase in Signals," *Proceedings of the IEEE*, vol. 69, pp. 529–541, 1981.

[9] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Prentice Hall, 2008.

[10] J. Oswald, S. Walmsley, C. Casey, S. Fregosi, B. Southall, and V. Janik, "Species Information in Whistle Frequency Modulation Patterns of Common Dolphins," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2021.

[11] A. Simonis, S. Baumann-Pickering, E. Oleson, M. Melcón, M. Gassmann, S. Wiggins, and J. Hildebrand, "High-frequency Modulated Signals of Killer Whales (Orcinus orca) in the North Pacific"," *J. Acoust. Soc. Am.*, Apr. 2012.

[12] A. Frankel and S. Yin, "A Description of Sounds Recorded from Melon-Headed Whales (*Peponocephala electra*) off Hawai," *J. Acoust. Soc. Am.*, 05 2010.

[13] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.

[14] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Am.*, 1996.

[15] D. Dimitriadis, A. Potamianos, and P. Maragos, "A comparison of the squared energy and teager-kaiser operators for short-term energy estimation in additive noise," *IEEE Transactions on Signal Processing*, 2009.

[16] V. Kandia and Y. Stylianou, "Detection of Sperm Whale Clicks Based on the Teager–Kaiser Energy Operator," *Applied Acoustics*, vol. 67, pp. 1144–1163, Nov. 2006.

[17] P. Bermant, M. Bronstein, R. Wood, S. Gero, and D. Gruber, "Deep Machine Learning Techniques for the Detection and Classification of Sperm Whale Bioacoustics," *Scientific Reports*, vol. 9, pp. 1–10, Aug. 2019.

[18] D. N. Makropoulos, A. Tsiami, A. Prospathopoulos, D. Kassis, A. Frantzis, E. Skarsoulis, G. Piperakis, and P. Maragos, "Convolutional Recurrent Neural Networks for the Classification of Cetacean Bioacoustic Patterns," in *Proc. IEEE ICASSP 2023*.

[19] A. Allen, M. Harvey, L. Harrell, A. Jansen, K. Merkens, C. Wall, J. Cattiau, and E. Oleson, "A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset," *Frontiers in Marine Science*, 2021.

[20] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders," in *Proc. of the 34th International Conference on Machine Learning (PMLR)*, 2017.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. ICLR*, 2021.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

[23] C. Garoufis, N. Zlatintsi, and P. Maragos, "Pre-training music classification models via music source separation," in *Proceedings of the 32nd European Signal Processing Conference (EUSIPCO)*. Lyon, France: IEEE, 2024.

[24] L. You, E. P. Coyotl, S. Gunturu, and M. Van Segbroeck, "Transformer-based bioacoustic sound event detection on few-shot learning tasks," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[25] M. Hagiwara, B. Hoffman, J.-Y. Liu, M. Cusimano, F. Effenberger, and K. Zacarian, "Beans: The Benchmark of Animal Sounds," in *Proc. IEEE ICASSP 2023*.

[26] D. Murphy, E. Ioup, M. Hoque, and M. Abdelguerfi, "Residual Learning for Marine Mammal Classification," *IEEE Access*, 2022.

[27] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global Birdsong Embeddings Enable Superior Transfer Learning for Bioacoustic Classification." *Scientific Reports*, vol. 9, pp. 1–10, Aug. 2023.

[28] A. Licciardi and D. Carbone, "WhaleNet: A novel deep learning architecture for marine mammals vocalizations on watkins marine mammal sound database," *IEEE Access*, vol. 12, 2024.

[29] A. Oppenheim and G. Verghese, *Signals, Systems and Inference*. Pearson Education, 2015.

[30] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin Transformer V2: Scaling Up Capacity and Resolution," in *Proc. IEEE/CVF Conf. Computer Vision & Pattern Recognition (CVPR)*, 2022.