TECNIACÚSTICA
CUENCA 2023

DETECCIÓN DE SILBIDOS DE MAMÍFEROS MARINOS EN ESPECTROGRAMAS UTILIZANDO YOLO-V5

Vicent Avaria Avaria[1]
Didac Diego-Tortosa[2]
Sergio Morell-Monzó[3*]
Carlos-Alberto Quiroz-Rangel[3]

[1]Pixelabs S.L., Salamanca, 6, 28020. Madrid (Madrid), España.
[2]Istituto Nazionale di Fisica Nucleare (INFN-LNS), Via S. Sofia 62, Catania, 95123 Italy.
[3]Universitat Politècnica de València (UPV-IGIC), Paranimf, 1, 46730. Gandia (València), España.

RESUMEN

La monitorización acústica del medio marino es esencial para el seguimiento de especies clave como los mamíferos marinos. Los sistemas de monitorización acústica pasiva contribuyen al estudio de estas especies generando una gran cantidad de datos que deben ser procesados y analizados. Los recientes avances en el campo del aprendizaje profundo permiten automatizar algunas de estas tareas, como la identificación de silbidos de mamíferos marinos en espectrogramas. Sin embargo, entrenar estos modelos desde cero requiere procesar y etiquetar una gran cantidad de datos. En este estudio realizamos un trabajo de aprendizaje por transferencia para el ajuste fino del modelo You-Only-Look-Once (YOLO-v5) que permite la detección de silbidos de mamíferos marinos en espectrogramas. Los resultados demuestran la viabilidad para detectar dichos silbidos con una precisión promedio (IoU ≥ 0,5) de 0,70. Se requiere una mayor investigación centrada en la mejora del procesamiento de los datos y del entrenamiento para generar un modelo más preciso y robusto.

ABSTRACT

Acoustic monitoring of the marine environment is essential for tracking key species, such as marine mammals. Passive acoustic monitoring systems contribute to monitoring these species by generating a large amount of data that must be processed and analyzed. Recent advances in the deep learning field make it possible to automate some of these tasks, such as the identification of marine mammal whistles in spectrograms. However, training these models from scratch requires processing and labeling a large amount of data. In this study, we perform a transfer learning exercise for fine-tuning the You-Only-Look-Once (YOLO-v5) model in detecting marine mammal whistles in spectrograms. The results demonstrate the feasibility to detect such whistles with an AP (IoU ≥ 0.5) of 0.70. Further research focused on improving data processing and training is required to generate a more accurate and robust model.

Palabras Clave— aprendizaje profundo, espectrograma, monitorización acústica pasiva, silbidos de mamíferos marinos.

1. INTRODUCCIÓN

La monitorización acústica del medio marino es un aspecto importante desde el punto de vista científico y medioambiental. El seguimiento de especies clave, como los mamíferos marinos, es esencial para su conservación y para un mayor conocimiento de dichas especies. En este sentido, los sistemas de monitorización acústica pasiva (PAMS, por sus siglas en inglés), cada vez más accesibles, contribuyen al estudio de estas especies generando una gran cantidad de datos que deben ser procesados y analizados. Generalmente, las grabaciones de audio registradas por los PAMS son analizadas de forma manual por biólogos marinos y oceanógrafos expertos a través de espectrogramas. Esta información permite identificar diferentes especies de mamíferos marinos [1], analizar el impacto de las actividades antrópicas sobre las poblaciones [2] y monitorizar la distribución espacial y temporal de las mismas [3].

* ***Autor de contacto****: sermomon@upv.es*

Los recientes avances en el campo del aprendizaje profundo (o *deep learning*) permiten automatizar algunas de estas tareas de identificación de silbidos de mamíferos marinos en espectrogamas [4, 5, 6], e incluso existen algunos trabajos cuyos modelos permiten clasificarlos en base a diferentes criterios [7]. Sin embargo, en la práctica, entrenar estos modelos requiere procesar y etiquetar una gran cantidad de datos, con el alto coste asociado en términos de tiempo y recursos. Además, a menudo es necesario reentrenar estos modelos a la hora de enfrentarlos a otros escenarios (lo que se conoce como cambio de dominio).

Actualmente, los grandes modelos de redes neuronales convolucionales (CNN, por sus siglas en inglés) permiten abordar estos problemas con menor cantidad de datos debido al alto conocimiento adquirido a través del entrenamiento con grandes bases de datos. Un ejemplo de estas CNN es You-Only-Look-Once (YOLO) [8]. YOLO-v5 fue pre-entrenado con cientos de miles de imágenes para la detección de objetos generalistas, con bases de datos como *COCO* o *Imagenet*. Una de las principales ventajas de este modelo es su velocidad de predicción (~60 FPS). En este estudio realizamos un trabajo de ajuste fino (o *fine tuning*) de YOLO-v5 para la detección de silbidos de mamíferos marinos en espectrogramas, generados por un sistema de implementación propia.

2. DATOS Y METODOLOGÍA

2.1. Adquisición de datos y generación del espectrograma

En este experimento se analizó una hora y veinte minutos de grabación registrada por un hidrófono estanco ubicado a gran profundidad (~2500 m) en el Mar Mediterráneo. Se trata de un hidrófono omnidireccional con una sensibilidad de recepción, o respuesta de voltaje en recepción (RVR, por sus siglas en inglés) de -165 dB re 1V/µPa en 5kHz. La frecuencia de muestreo es de 65,1 kHz, que permite el estudio de señales comprendidas entre los 5 y 30 kHz, que es la región de interés bioacústico.

Un espectrograma es la representación visual de una señal en su dominio tiempo-frecuencia (figura 1). Esta representación resulta útil para evidenciar la presencia de contenido frecuencial enmascarado en la señal temporal por un ruido más alto en frecuencias fuera del rango de interés. El espectrograma es una representación ampliamente utilizada para el estudio de silbidos de mamíferos marinos.

Se generó el espectrograma de la grabación registrada por el hidrófono utilizando la Transformada Rápida de Fourier (FFT), que discretiza la señal en tiempo en sus componentes de frecuencia. La FFT está condicionada por el número de muestras (nFFT) de la señal utilizada. Para realizar una FFT se necesitan al menos dos ciclos de la señal a exponer [9], lo que limita la frecuencia válida representable (f_{ok}) que dependerá de la frecuencia de muestreo (f_s) de la señal y del nFFT utilizado (ecuación 1).

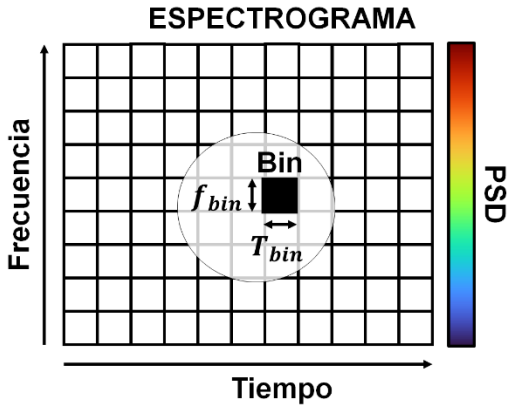

Figura 1. Representación gráfica de un espectrograma.

$$f_{ok} \geq \frac{2\, f_s}{nFFT} \tag{1}$$

Es común utilizar números nFFT en base dos para una FFT eficiente, dado que el tiempo de computación se reduce notablemente. Se realiza la FFT en una ventana de tiempo con nFFT muestras y se calcula su contenido frecuencial, este cálculo se realiza en intervalos de tiempo conocidos como *bins*, los cuáles marcarán la resolución temporal (anchura) y frecuencial (altura) del espectrograma (figura 1). La resolución frecuencial y temporal vienen delimitadas por las ecuaciones 2 y 3, respectivamente:

$$f_{bin} = \frac{f_s}{2\, nFFT} \tag{2}$$

$$T_{bin} = \frac{nFFT}{f_s} \cdot (1 - overlap) \tag{3}$$

Los valores de cada píxel del espectrograma corresponden a la densidad espectral de potencia (PSD) logarítmica, que tiene unidades de dB re A^2/Hz (donde A es la unidad de la amplitud de la señal, en nuestro caso µPa).

Con objeto de poder diferenciar varios tipos de silbidos, se prioriza un equilibrio entre resolución frecuencial-temporal. Para ello se utilizaron 1024 nFFT con un solape (*overlap*) del 50%, lo que supone un f_{bin} de 31,8 Hz y un T_{bin} de unos 7,86 ms. Finalmente, las señales de audio fueron recortadas en porciones de 3,54s para generar los espectrogramas. Este proceso permite generar una matriz (o imagen) cuadrada de 448 x 448 píxeles (*bins*). En la figura 2 se muestra el espectrograma entre 1,97 y 30,39 kHz (f_{ok} se sitúa en 127,15 Hz). Después, se limitó la escala de PSD entre los 20 y 100 dB, para evitar que un ruido mayor altere la escala de grises. Finalmente, los valores de PSD fueron transferidos a un espacio RGB con 256 niveles de gris en cada canal. En total se generaron 2873 imágenes 8-bit con un solape del 50% entre ellas.

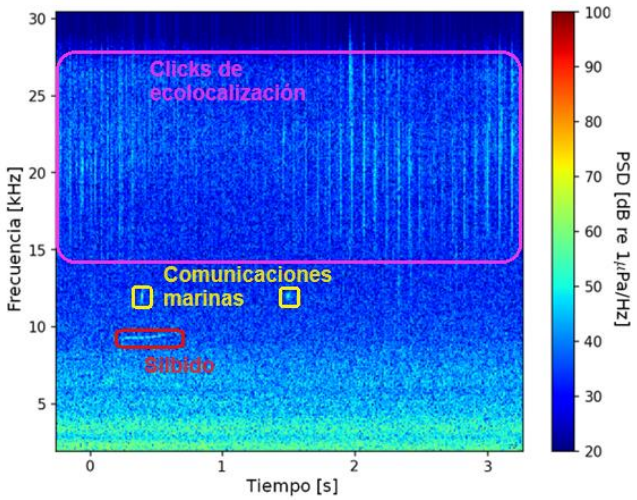

Figura 2. Espectrograma con diferentes sonidos identificados: clicks de ecolocalización, sonidos derivados de comunicaciones marítimas y silbidos emitidos por mamíferos marinos.

2.2. Etiquetado de espectrogramas

El objetivo del etiquetado es proporcionar ejemplos resueltos (*ground truth*) para el reentrenamiento del modelo en la tarea de identificación de silbidos. Este proceso de reentrenamiento para una tarea específica se conoce como *fine tuning*. Las 2873 imágenes generadas fueron etiquetadas manualmente por tres técnicos independientes, previamente entrenados para la identificación de los patrones que definen los silbidos de diferentes mamíferos marinos. En total se pudieron identificar 1586 silbidos en una hora y veinte minutos de grabación, lo que supone una ocurrencia de 19,82 silbidos/minuto. Solamente un tercio de las imágenes contenían al menos un silbido identificable. Las etiquetas fueron almacenadas en formato "*DarkNet*":

```
<object-class> <x_center> <y_center> <width> <height>
```

Durante el proceso de etiquetado se observó una gran variedad de tipos de silbidos tanto en forma, duración, intensidad y frecuencia. También se identificaron sonidos derivados de comunicaciones marinas a una frecuencia de 12 kHz de forma periódica y de corta duración como se puede observar en la figura 2. Estos sonidos también fueron etiquetados para el entrenamiento.

2.3. You Only Look Once v5

YOLO-v5 es la primera versión de código abierto desarrollada por la compañía Ultralitics para la detección de objetos en tiempo real [10]. Esta arquitectura de red neuronal convolucional se basa en el trabajo de Redmon y Farhadi (2018) [11]. La arquitectura de la red consta de tres partes principales:

- ***Backbone*:** La columna vertebral utiliza la estructura *CSP-Darknet53* [12], una modificación de la arquitectura *DarkNet* utilizada en versiones anteriores y que se utiliza como extractor de características.
- ***Neck:*** Esta parte conecta la columna vertebral y la cabeza. En YOLO-v5, se utilizan estructuras *SPPF* y *CSP-PAN* [12, 13].
- ***Head:*** Esta parte es responsable de generar la inferencia. YOLO-v5 utiliza el mismo cabezal YOLO-v3 [10].

2.4. Entrenamiento y validación

Con las imágenes completamente etiquetadas se reentrenaron las últimas capas de la red, donde se extraen las características más especializadas y se genera la inferencia. El resto de las capas permanecen congeladas. Este proceso permite especializar el modelo en una tarea concreta, en este caso, la detección de silbidos. En este trabajo, se utilizó la versión YOLO-v5m que corresponde a la arquitectura de tamaño mediano que proporciona un equilibrio entre precisión y tiempo de ejecución. Se utilizó un equipo con 16 GB RAM y una GPU NVIDIA RTX2060 6GB RAM en un entorno Ubuntu 22.04 LTS. Siguiendo las especificaciones de la compañía Ultralitics [11] el entrenamiento se realizó en la plataforma PyTorch 2.0 con CUDA 11 y CudNN 7.5.

El entrenamiento del modelo se realizó utilizando los siguientes hiperparámetros:

- **Array de entrada:** [448, 448, 3] que corresponde al alto y ancho de la imagen y número de canales RGB, respectivamente. Se eligió este tamaño de entrada para evitar deformaciones de la imagen.
- **Tamaño de lote (*batch size*):** 8, limitado por las capacidades de la GPU utilizada.
- **Número de épocas:** 300. El modelo fue entrenado hasta un máximo de 300 épocas, sin embargo, el modelo resultante (*checkpoint*) fue el de aquella época que obtuvo los mejores resultados.
- **Optimizador:** *Stochastic Gradient Descent* (SGD).
- **Ratio de aprendizaje:** 0,01.
- **Otros hiperparámetros:** el resto de hiperparámetros fueron ajustados por defecto.

Durante el entrenamiento del modelo se hace uso de la función de pérdida, proporcionada por el optimizador SGD, para visualizar si el modelo está aprendiendo correctamente.

Para validar el modelo y evaluar su rendimiento se realizó una validación cruzada de 5 iteraciones. En cada iteración se utilizaron el 90% de las imágenes para entrenamiento y el 10% para validación sin reemplazo. Finalmente, se promediaron los resultados obtenidos en todas las iteraciones. Se calcularon las siguientes medidas de exactitud:

- **Precisión:** se define como el número de verdaderos positivos entre la suma de verdaderos positivos y falsos positivos. Consideramos verdaderos positivos aquellas detecciones cuya intersección sobre la unión (IoU, por sus siglas en inglés) es ≥ 0,5. Esta medida informa sobre la

proporción de sibidos correctamente detectados respecto a las detecciones del modelo.

- **Recall (recuperación):** se define como el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos. Se consideran verdaderos positivos aquellas detecciones cuya IoU es ≥ 0,5. Esta medida informa sobre la proporción de silbidos correctamente detectados del total de silbidos presentes.
- **F1-score:** corresponde a la media harmónica entre precisión y recall.
- **mAP** y **AP_i:** la mAP se define como el promedio de área bajo la curva precisión-recall para todas las clases, mientras que el AP_i es el área bajo la curva precisión-recall de la clase *i*.

2.5. Experimentos

Se plantearon dos escenarios de detección. El primero de ellos (escenario 1) se centró en identificar únicamente los silbidos (*whistle*) procedentes de mamíferos marinos. Para ello, se utilizaron únicamente las etiquetas correspondientes a silbidos para entrenar el modelo. El segundo (escenario 2) se centró en detectar tanto los silbidos (*whistle*) como los sonidos derivados de comunicaciones marinas (*noise*). Para ello, se utilizaron tanto las etiquetas correspondientes a silbidos como las etiquetas correspondientes a sonidos de comunicaciones marinas para entrenar el modelo. Este segundo escenario se planteó con objeto ayudar al modelo a detectar mejor los silbidos sin confusión con los sonidos derivados de comunicaciones marinas.

3. RESULTADOS

3.2. Escenario 1: whistle

En este escenario se entrenó el modelo hasta un máximo de 300 épocas. La época con mejores resultados fue la 185. La tabla 1 muestra los resultados del modelo en el escenario 1 en la fase de validación. Según estos resultados el modelo es capaz de detectar los silbidos con una mAP (IoU ≥ 0,5) del 65%.

Tabla 1. Resultados de la validación del modelo en el escenario 1.

Precisión	0,63
Recall	0,72
F1-score	0,67
mAP (IoU ≥ 0,5)	0,65

En la figura 3 se muestra la curva precisión-recall y el F1-score a diferentes niveles de confianza. Se obtiene el valor máximo de F1-score a una confianza de 0,29.

3.2. Escenario 2: whistle + noise

En este escenario se entrenó el modelo hasta un máximo de 300 épocas. La época con mejores resultados fue la 248. La tabla 2 muestra los resultados del modelo en el escenario 2 en la fase de validación. Según estos resultados el modelo es capaz de detectar los silbidos con una mAP (IoU ≥ 0,5) del 76%.

Tabla 2. Resultados de la validación del modelo en el escenario 2.

Precisión	0,77
Recall	0,75
F1-score	0,76
mAP (IoU ≥ 0.5)	0,76

En la figura 4 se muestra la curva precisión-recall y el F1-score a diferentes niveles de confianza. Se obtiene un valor máximo de F1-score a una confianza de 0,42.

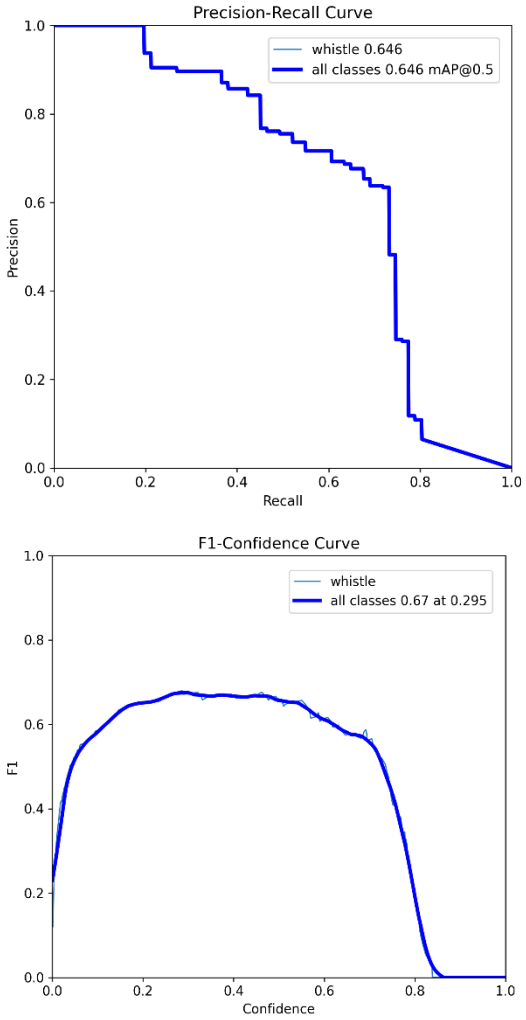

Figura 3. Curva precisión-recall y F1-score a diferentes niveles de confianza en el escenario 1.

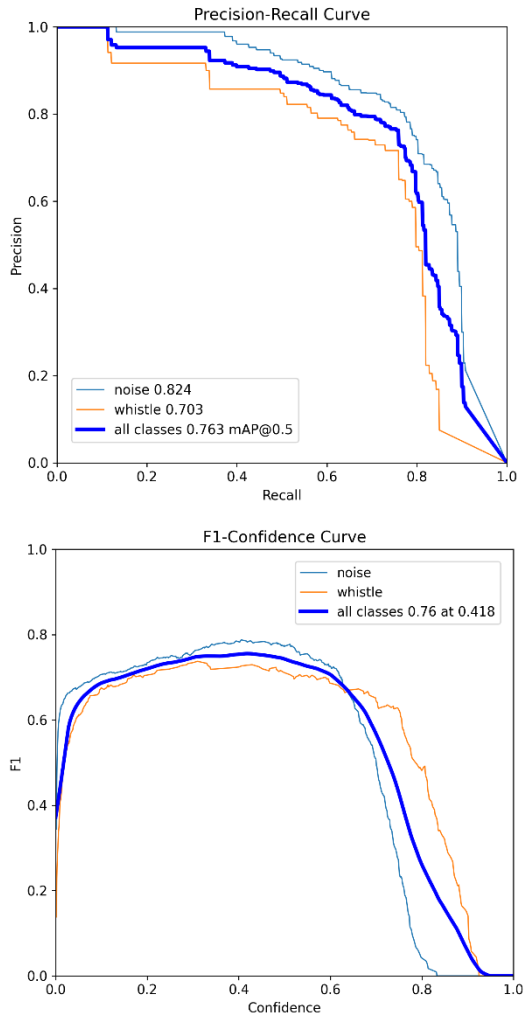

Figura 4. Curva precisión-recall y F1-score a diferentes niveles de confianza en el escenario 2.

Si analizamos el AP (IoU ≥ 0,5) de cada clase por separado observamos que el AP de la clase *noise* que corresponde a los sonidos derivados de comunicaciones marinas es de 0,82 mientras que el de la clase *whistle* que corresponde a los silbidos es de 0,70.

3. DISCUSIÓN

Este trabajo permitió entrenar un modelo YOLO-v5m para la detección de silbidos de mamíferos marinos en espectrogramas. En base al mejor de los dos escenarios de detección planteados se obtuvo un mAP (IoU ≥ 0,5) de 0,76 con un AP (IoU ≥ 0,5) de 0,70 para la detección de silbidos. Añadir las etiquetas de otros sonidos que puedan generar confusión mejoró la detección de silbidos con un incremento de AP (IoU ≥ 0,5) de 0,65 a 0,70. Por tanto, esta es una buena recomendación para que el modelo sea capaz de discernir entre ambas categorías.

El uso de un modelo preentrenado como YOLO-v5 presenta una ventaja con respecto a otros trabajos donde se entrenaron modelos desde cero por su menor requerimiento de datos. Además, el enfoque basado en detección de objetos es especialmente conveniente de cara a la cuantificación y caracterización de silbidos con respecto a otros enfoques como la clasificación de imágenes. Por otro lado, es posible que existan limitaciones para elevar la exactitud de YOLO-v5, incluso usando una mayor cantidad de datos y mejores cadenas de procesamiento, debido a que este fue entrenado con imágenes de objetos generalistas.

En base a la interpretación visual de las detecciones del modelo (ver figura 5) y la experiencia adquirida durante el proceso de etiquetado observamos que a menudo es difícil definir los límites de un silbido. Además, frecuentemente los silbidos están divididos en dos o más partes. Estas características provocan que una cantidad considerable de detecciones no sean consideradas verdaderos positivos debido a que no superan el umbral de IoU ≥ 0,5, a pesar de que el modelo fue capaz de identificar el silbido. En este sentido, incluso cuando los tres técnicos trataron de etiquetar las mismas muestras, el solape entre sus cuadros delimitadores fue relativamente bajo. Por este motivo, esta validación podría estar subestimando el rendimiento del modelo. De hecho, el modelo fue capaz de identificar silbidos en más del 80% de las imágenes que contenían uno o más silbidos. Una validación más exhaustiva debería calcular el rendimiento del modelo a diferentes umbrales IoU y compararlo con la capacidad de detección de un etiquetador humano.

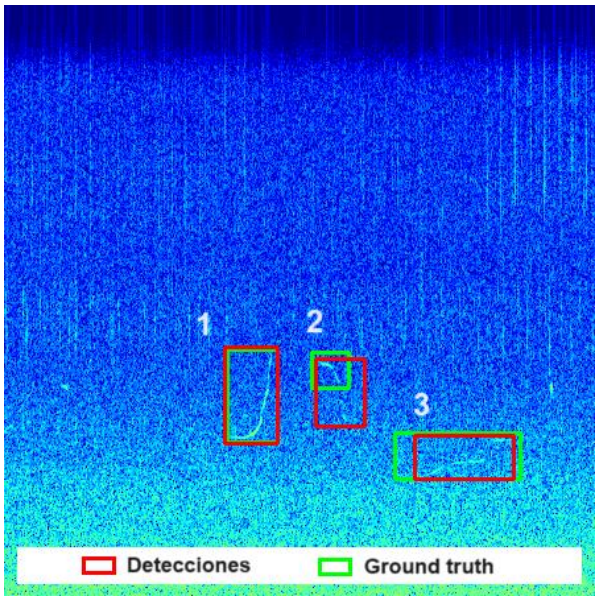

Figura 5. Ejemplo de detecciones del modelo en la imagen 301 del conjunto de datos de validación. El silbido 2 fue detectado, pero no se consideró verdadero positivo por no tener una IoU ≥ 0,5.

Una herramienta que incorpore este modelo podría ahorrar tiempo y costes a la hora de caracterizar y etiquetar grandes cantidades de datos. Esta herramienta podría agilizar la creación de muestras para entrenar nuevos modelos de detección y clasificación de silbidos de mamíferos marinos. En este sentido, es necesario realizar una validación más exhaustiva y definir diferentes umbrales IoU aceptables. También es necesario validar el modelo con un conjunto de datos de *test*, así como ponerlo a prueba en diferentes escenarios, como grabaciones de otros hidrófonos, en otras zonas geográficas y otros periodos de tiempo.

Por otra parte, se sabe que la relación señal a ruido (SNR, por sus siglas en ingés) afecta a la capacidad de reconocer visualmente los silbidos en espectrogramas. Por tanto, es asumible que una reducción de la SNR afectará a la capacidad de detección del modelo. Futura investigación debería tratar de cuantificar la sensibilidad del modelo al ruido para definir unos límites de operabilidad.

4. CONCLUSIONES

Este trabajo demuestra la viabilidad de entrenar un modelo YOLO-v5 para la detección de silbidos de mamíferos marinos en espectrogramas. So obtuvo un mAP (IoU $\geq$ 0.5) de 0.76 y un AP (IoU $\geq$ 0.5) para la detección de silbidos de 0.70. Entrenar el modelo también con etiquetas de otros sonidos que pueden confundirse con los silbidos permitió al modelo distinguir ambas categorías y mejorar la detección de silbidos.

Los resultados obtenidos son prometedores y justifican una mayor investigación en este campo. Es necesario realizar una validación más exhaustiva, así como mejorar el procesamiento de datos y el entrenamiento del modelo.

5. AGRADECIMIENTOS

Los autores agradecen la financiación y recursos proporcionados por el proyecto *Nostrum, Telescopios de neutrinos para la física fundamental y astronomía multi-mensajero* en la Universitat Politècnica de València (UPV): PID2021-124591NB-C42 a través de la Agencia Estatal de Investigación. Igualmente, los autores agradecen el apoyo de los doctores Víctor Espinosa Roselló y Miguel Ardid Ramírez del Instituto de Investigación para la Gestión Integrada de Zonas Costeras – IGIC, de la UPV.

6. REFERENCIAS

[1] Janik, V. M., King, S. L., Sayigh, L. S., & Wells, R. S. (2012). Identifying signature whistles from recordings of groups of unrestrained bottlenose dolphins (Tursiops truncatus). Marine Mammal Science, vol. 29-1, pp. 109–122. https://doi.org/10.1111/j.1748-7692.2011.00549.x

[2] Lara, G., Bou-Cabo, M., Llorens, S., Miralles, R., & Espinosa, V. (2023). Acoustical Behavior of Delphinid Whistles in the Presence of an Underwater Explosion Event in the Mediterranean Coastal Waters of Spain. Journal of Marine Science and Engineering, vol. 11-4, p. 780. https://doi.org/10.3390/jmse11040780

[3] Silva, T., Mooney, T., Sayigh, L., & Baumgartner, M. (2019). Temporal and spatial distributions of delphinid species in Massachusetts Bay (USA) using passive acoustics from ocean gliders. Marine Ecology Progress Series, vol. 631, pp. 1–17. https://doi.org/10.3354/meps13180

[4] Duan, D., Lü, L., Jiang, Y., Liu, Z., Yang, C., Guo, J., & Wang, X. (2022). Real-time identification of marine mammal calls based on convolutional neural networks. Applied Acoustics, vol. 192, p. 108755. https://doi.org/10.1016/j.apacoust.2022.108755

[5] Li, P., Liu, X., Palmer, K. J., Fleishman, E., Gillespie, D., Nosal, E.-M., Shiu, Y., Klinck, H., Cholewiak, D., Helble, T., & Roch, M. A. (2020). Learning Deep Models from Synthetic Data for Extracting Dolphin Whistle Contours. IEEE 2020 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/ijcnn48605.2020.9206992

[6] Nur Korkmaz, B., Diamant, R., Danino, G., & Testolin, A. (2023). Automated detection of dolphin whistles with convolutional networks and transfer learning. Frontiers in Artificial Intelligence, vol. 6. https://doi.org/10.3389/frai.2023.1099022

[7] Jiang, J., Bu, L., Duan, F., Wang, X., Liu, W., Sun, Z., & Li, C. (2019). Whistle detection and classification for whales based on convolutional neural networks. Applied Acoustics, vol. 150, pp. 169–178. https://doi.org/10.1016/j.apacoust.2019.02.007

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection (Version 5). arXiv. https://doi.org/10.48550/ARXIV.1506.02640

[9] G.F. Lara Martínez, Caracterización y modelado de la producción de sonidos de las ballenas beluga (Delphinapterus Leucas) basado en modelos de análisis/síntesis de voz, Tesis Doctoral, Universitat Politècnica de València (UPV), Oct, 2016. https://doi.org/10.4995/Thesis/10251/74645

[10] Ultralitics. YOLOv5. Github repository: https://github.com/ultralytics/yolov5

[11] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1804.02767

[12] Wang, C.-Y., Liao, H.-Y. M., Yeh, I.-H., Wu, Y.-H., Chen, P.-Y., & Hsieh, J.-W. (2019). CSPNet: A New Backbone that can Enhance Learning Capability of CNN (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1911.11929

[13] Jani, M., Fayyad, J., Al-Younes, Y., & Najjaran, H. (2023). Model Compression Methods for YOLOv5: A Review (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.11904