Identificación de sonotipos a partir de grabaciones captadas con hidrófonos

Detección automática de peces y embarcaciones mediante técnicas de inteligencia computacional

Manuela Ospina Giraldo

Proyecto de Investigación presentado como requisito parcial para optar al título de:
Bioingeniería

José David López Hincapié, PhD.

Universidad de Antioquia
Facultad de Ingeniería
Bioingeniería
Medellín, Antioquia
2023

Cita		Ospina Giraldo [1]
Referencia Estilo IEEE (2020)	[1]	M. Ospina Giraldo, "Identificación de sonotipos a partir de grabaciones captadas con hidrófonos: Detección automática de peces y embarcaciones mediante técnicas de inteligencia computacional", Bioingeniería, Pregrado, Universidad de Antioquia, Medellín, Antioquia, 2023.

Grupo de Investigación Sistemas Embebidos y Computacionales (SISTEMIC)

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación de Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano/Director: Julio César Saldarriaga Molina.

Jefe departamento: John Fredy Ochoa Gómez.

Dedicatoria

Quiero dedicar este trabajo a mis padres quienes fueron mi motivación constante, que me dieron su apoyo en todo momento, quienes creyeron en mí incansablemente dándome todo lo necesario para continuar con mis estudios, celebrando a mi lado cada logro y siendo parte de mi crecimiento como persona y como profesional, gracias por su amor, comprensión, paciencia y apoyo en cada paso que daba.

A mis hermanos que compartieron a mi lado risas, que apoyaban mi proceso y que fueron amigos y compañeros con los que compartía momentos de paz proporcionando una fuerza revitalizante en mi vida y aliento en los momentos más desafiantes.

Por último, quiero agradecer a Dios por cuidarme, porque incluso en mis momentos de mayor soledad estuvo a mi lado. Dedico este logro principalmente a Él quien me ha dado fuerza e inteligencia para lograr todas mis metas.

Agradecimientos

Quiero expresar mi más profundo agradecimiento a la profesora Claudia Victoria Isaza, PhD., al profesor José David López, PhD., y al profesor Eduardo Romero Vivas, PhD., por su orientación experta, paciencia y dedicación a lo largo de todo el proceso de investigación. Sus consejos y retroalimentación han sido fundamentales para el desarrollo de este trabajo y para mi crecimiento académico e investigativo buscando oportunidades para que pueda continuar investigando.

Mi gratitud también se extiende a la Universidad de Antioquia (UdeA) y al Centro de Investigaciones Biológicas del Noroeste (CIBNOR) por brindarme las oportunidades y recursos necesarios para llevar a cabo esta investigación. Asimismo, deseo agradecer al grupo de investigación SISTEMIC por darme la oportunidad de incursionar en este proceso investigativo proporcionando medios para lograr completar mi trabajo.

Agradezco especialmente a María José Guerrero, quien me ha acompañado y brindado su apoyo durante este recorrido. Sus discusiones y aportes han enriquecido mi trabajo y han sido un aliento valioso en los momentos de desafío.

Por último, quiero agradecer a Dios, mi familia y seres queridos por su amor incondicional, comprensión y apoyo en todo momento. Las palabras de aliento y cariño que me han brindado han sido un motor vital en este camino de crecimiento y aprendizaje.

TABLA DE CONTENIDO

TABLA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

CIBNOR	Centro de Investigaciones Biológicas del Noroeste
UdeA	Universidad de Antioquia
PhD	Philosophiae Doctor
PAM	Passive Acoustic Monitoring
PDI	Procesamiento Digital de Imágenes
CNN	Convolutional Neural Network
LFCC	Linear Frequency Cepstral Coefficients
LAMDA	Learning Algorithm for Multivariate Data Analysis
ANH	Acoustic Niche Hypothesis
AAH	Acoustic Adaptation Hypothesis
PPV	Positive-Predictive-Value
PSD	Potential Spectral Density
MFCC	Mel-Frequency Cepstral Coefficients
SSA	Spectral Subtraction Algorithm
AED	Acoustic Event Detection
STFT	Transformada de Fourier de Tiempo Corto
DFT	Transformada Discreta de Fourier
MAD	Grado de Adecuación Marginal
NIC	Clase No Informativa
GAD	Grado de Adecuación Global

RESUMEN

La bioacústica es un método de monitoreo pasivo de los ecosistemas que permite hacer seguimiento a las comunidades animales a través de los sonidos que emiten, los cuales son grabados y proporcionan información a los biólogos sobre el comportamiento y número de los seres acuáticos en una zona determinada de manera eficiente, sin causar daños adicionales al ecosistema. Este método de monitoreo es especialmente importante en hábitats marinos, que representan el 70% de la superficie terrestre y son de difícil acceso para un estudio profundo y prolongado.

En este estudio se propone la implementación de una metodología que permita realizar una identificación y detección de fonaciones de dos clases de peces objetivo nombradas inicialmente como Peces 1 y Peces 2. Como paso inicial se realiza una identificación de embarcaciones las cuales son fuente de ruido al momento de la detección de sonotipos. Para esto se aplicó una regresión logística a audios de un minuto de duración alcanza un *F1-Score* de 0.9942. Los audios clasificados como embarcación son descartados con el fin de mejorar el rendimiento del algoritmo de reconocimiento de fonaciones de peces.

Para la detección de las fonaciones objetivo se aplicó un algoritmo de inteligencia computacional no supervisada, específicamente el algoritmo LAMDA 3π, el cual es complementado con una red neuronal convolucional ResNet18 para la identificación y detección automática de sonidos producidos principalmente por los peces objetivo en grabaciones de audio captadas por investigadores del Centro de Investigaciones Biológicas del Noroeste con hidrófonos en la Ensenada de La Paz, México. Se aplicó *Acoustic Event Detection* para la segmentación de las fonaciones objetivo de este proyecto realizando modificaciones y parametrizando dicho proceso para dar un tratamiento diferente a segmentos de área extensa y área pequeña. Dichos segmentos son procesados mediante técnicas de PDI antes de ser clasificados por la red neuronal. Se obtuvo un *F1-Score* de 0.7805 y 0.8272 para Peces 1 y Peces 2 respectivamente.

Palabras clave **— Bioacústica, Peces, Embarcaciones, Identificación Automática de Fonaciones, Agrupamiento.**

ABSTRACT

Bioacoustics is a passive monitoring method of ecosystems that allows tracking animal communities through the sounds they emit, which are recorded and provide information to biologists about the behavior and number of aquatic beings in a specific area efficiently, without causing additional harm to the ecosystem. This monitoring method is especially important in marine habitats, which represent 70% of the Earth's surface and are difficult to access for in-depth and prolonged study.

In this study, the implementation of a methodology is proposed to identify and detect vocalizations of two target fish species initially named as Fish 1 and Fish 2. As an initial step, the identification of vessels, which are a source of noise during the detection of sound types, is conducted. For this purpose, logistic regression was applied to one-minute audio clips, achieving an F1-Score of 0.9942. The audio clips classified as vessels are discarded to improve the performance of the fish vocalization recognition algorithm.

For the detection of the target vocalizations, an unsupervised computational intelligence algorithm, specifically the LAMDA 3π algorithm, was applied, complemented with a Convolutional Neural Network (CNN) ResNet18 for the automatic identification and detection of sounds produced mainly by the target fish in audio recordings captured by researchers from the Center for Biological Research of the Northwest with hydrophones in La Paz Bay, Mexico. Acoustic Event Detection was applied to segment the target vocalizations of this project, making modifications, and parameterizing this process to treat large and small area segments differently. These segments are processed through image processing techniques before being classified by the CNN. An F1-Score of 0.7805 and 0.8272 was obtained for Fish 1 and Fish 2, respectively.

Keywords **— Bioacoustics, Fish, Vessels, Automatic Vocalization Identification, Clustering.**

1. INTRODUCCIÓN

El monitoreo ambiental se ha utilizado ampliamente para comprender el funcionamiento de los ecosistemas, identificar patrones de biodiversidad, posibles amenazas a las comunidades biológicas y análisis de variaciones en el tiempo y el espacio. El monitoreo acústico pasivo (PAM, por sus siglas en inglés) es una herramienta de bajo costo que permite recolectar datos de forma continua y menos invasiva. Los sonidos emitidos por un ecosistema conforman un paisaje acústico. Según Pijanowski [1], un paisaje acústico es "la colección de sonidos biológicos, geofísicos y antropogénicos que emanan de un paisaje y varían en el tiempo y espacio, reflejando procesos importantes en el ecosistema y actividades humanas". Dependiendo de la fuente del sonido, se pueden distinguir: (1) Biofonía, proveniente de fuentes biológicas; (2) Geofonía, generada por el ambiente geofísico que incluye viento, agua, truenos, movimiento de la tierra, etc.; y (3) Antropofonía, que representa los sonidos producidos por objetos construidos por humanos, estacionarios y móviles.

La bioacústica es una disciplina que se encarga del estudio de los sonidos producidos por animales [2]. Esta área de investigación ofrece la posibilidad de analizar el comportamiento de comunidades y la estructura de un ecosistema, así como sus variaciones en el tiempo y el espacio [3]. Además, la bioacústica proporciona una ventana para el monitoreo no invasivo de la biodiversidad, evitando interferencias en el comportamiento natural de los individuos facilitando la creación de programas de conservación biológica de tal manera que el avance del ser humano no represente una degradación ambiental [4].

En el caso de especies marinas, muchas dependen de la propagación de ondas de sonido para explorar su entorno puesto que, en general, la capacidad visual es limitada debido a materiales en suspensión y proliferaciones de plancton que disminuyen la claridad del agua, adicional a que la luz es absorbida de manera desigual. Es por esto que el paisaje acústico toma un papel fundamental en un ecosistema marino e influye en su comportamiento y en las dinámicas de las comunidades, reflejando la calidad de los biomas acuáticos [5].

Las principales teorías conocidas en ecología del paisaje acústico son la hipótesis del nicho acústico (ANH, por sus siglas en inglés) y la hipótesis de adaptación acústica (AAH,

por sus siglas en inglés). La primera se basa en observaciones empíricas y establece que las especies compiten por un nicho acústico ocupando una banda de frecuencias que las demás especies del ecosistema respetan. La segunda hipótesis propone que las especies buscan su nicho acústico y modulan los sonidos que producen para ocupar una banda de frecuencia disponible en el hábitat que ocupan [4]. Estas hipótesis son utilizadas por los biólogos para limitar su búsqueda de una especie a una banda de frecuencias objetivo y, por lo tanto, reducir la cantidad de operaciones realizadas por técnicas de preprocesamiento y algoritmos de aprendizaje automático.

En este proyecto se estudia el grupo taxonómico de peces. En este caso la evolución de los mecanismos de producción de sonido es poco conocida y parecen haber evolucionado independientemente entre las casi 30,000 especies de peces por lo que, mientras que los vertebrados terrestres poseen un órgano vocal principal, los peces no disponen de dicha estructura, y el conocimiento de los mecanismos de producción de sonidos se basa enteramente en el estudio de peces óseos actinopterigios modernos (clase Actinopterygii) [6]. Los peces pertenecientes a dicha clase son conocidos como peces con aletas radiadas debido a que estas son sostenidas por finas estructuras óseas llamadas radios.

Los mecanismos más importantes comprenden estructuras dedicadas exclusivamente a la producción de sonido como músculos sónicos encargados de producir vibración en la vejiga natatoria (*drumming*). Dichos músculos pueden estar unidos a la vejiga natatoria (Intrínseco), se pueden generar fuera de ella e insertarse en sus paredes (Extrínseco) o los músculos pueden hacer vibrar la vejiga indirectamente mediante tendones o platos óseos sin ninguna unión a ella (Extrínseco Indirecto) [6].

Otro mecanismo usa las aletas pectorales y cinturas para la producción de sonido. En este caso los músculos involucrados tienen funciones adicionales a la producción de sonido. Varias familias de los peces gato tienen un primer radio de la aleta pectoral mejorado (espina pectoral) que puede generar sonidos estridulatorios cuando se frota contra una ranura de la cintura escapular. En los gouramis croadores, dos tendones mejorados de las aletas pectorales se estiran y son arrancados, similar a las cuerdas de una guitarra, durante el aleteo rápido de las aletas. En los peces escorpiones, toda la cintura pectoral vibra por un músculo sónico [6]. Estos mecanismos se ilustran en la figura 1.

Existen otros mecanismos no relacionados con los dos anteriores pero que son más diversos y específicos entre las especies y géneros de peces.

La ecoacústica ha utilizado métodos de análisis computacional, tales como procesamiento de señales o aprendizaje automático [2]. Debido a la gran cantidad de datos que se recolectan y que deben ser analizados por parte de los expertos, el proceso se vuelve extenso y abrumador [7]. Por esta razón, las técnicas de inteligencia computacional ofrecen una alternativa de análisis a través de la identificación automática de sonotipos presentes en las grabaciones captadas. En los últimos años, el uso de técnicas no supervisadas para esta identificación ha ganado popularidad debido a su capacidad para agrupar datos con base a su similitud y la posibilidad de obtener grados de pertenencia mediante lógica difusa sin la necesidad de etiquetas evitando la etapa de etiquetado de los datos y aportando eficiencia en el proceso de identificación de sonotipos [8]. Sin embargo, las aproximaciones no supervisadas tienden a presentar gran cantidad de falsos positivos en ambientes marinos por lo que un modelo supervisado es complementario en estos casos permitiendo alcanzar mejores rendimientos en los modelos, similar a lo aplicado por Bergler, en donde se implementó un enfoque semi-supervisado obteniendo un PPV (Positive-Predictive-Value) de 93.2% [9].

Varios estudios se han realizado para la detección de fonaciones de peces, por ejemplo, en el estudio de Ibrahim se alcanza un *Accuracy* similar al presentado en este estudio igual a 0.827 con una base de datos conformada por tres conjuntos de datos de 9,999 archivos cada una [10]. Por otro lado, Amal obtiene un *Accuracy* de 0.84 usando grabaciones tomadas en un ambiente controlado (peceras) [11] al igual que en el estudio de Noda quien obtuvo un valor de 0.9558 en esta métrica [12]. Ozanich, alcanzó 0.68 de *Accuracy* aplicando métodos no supervisados [13], mientras que Harakawa exhibe un *F1-Score* de 0.865 con una base de datos de 39,888 audios [14]. En aplicaciones de *Deep Learning* se resalta el trabajo de Laplante quien obtuvo un *F1-Score* igual a 0.94 con un *dataset* de 446,764 grabaciones [15], y el estudio de Waddell alcanzando un *Accuracy* de 0.87 con una base de datos conformada por 91,387 segmentos [16].

En cuanto a enfoques no supervisados, destaca el rendimiento del algoritmo LAMDA-3pi en el reconocimiento de anuros [8], alcanzando precisiones entre el 99.38%

y el 100% en la identificación de seis especies de anuros, e incluso descubriendo dos especies adicionales no incluidas en la etapa de entrenamiento. Este algoritmo fue posteriormente modificado por Guerrero [3] para la detección de múltiples especies, utilizando 24 LFCC, frecuencia mínima, máxima y dominante como características, y logrando rendimientos entre el 75% y el 96% en predicciones de presencia-ausencia. Sin embargo, estas propuestas no han sido probadas en ambientes marinos.

En este estudio, se busca complementar la metodología y el algoritmo LAMDA-3pi, previamente aplicados para la identificación de múltiples especies terrestres por Guerrero [3], para la identificación de fonaciones marinas, específicamente de drumming y estridulación producidas por dos clases de peces nombrados como Peces 1 y Peces 2. Además, se propone una metodología de detección de embarcaciones con el fin de descartarlas, ya que su ruido afecta el reconocimiento de la metodología propuesta.

Inicialmente se disponen de cinco grabaciones tomadas en La Paz, México. Cada una se segmenta en audios de un minuto de duración, obteniendo una base de datos compuesta por 1,856 elementos. En primer lugar, se identifican los audios con presencia de embarcaciones, para lo cual se extraen 16 coeficientes cepstrales en frecuencia Mel (MFCC) y el promedio de la densidad espectral de potencia (PSD) para cada audio. Estas características se ingresan a un modelo de regresión logística. Los audios que contienen embarcaciones son descartados y no ingresan al algoritmo de *clustering*. A los elementos restantes se les extrae el espectrograma correspondiente, al cual se le aplica un filtro Gaussiano y un algoritmo de sustracción espectral (SSA) para la reducción de ruido. Luego de esto, se aplica umbralización de Otsu y se realizan operaciones morfológicas de apertura y cierre para obtener segmentos. Las características extraídas para el algoritmo de agrupamiento son 24 coeficientes cepstrales lineales en frecuencia (LFCC) y las frecuencias mínima, máxima y dominante, lo que resulta en un total de 27 variables por cada segmento. Posteriormente, estas variables se ingresan al algoritmo LAMDA-3pi, el cual genera los clústeres con base a los grados de adecuación global [3]. Una vez identificado el cluster que mejor representa la fonación objetivo se extraen los segmentos pertenecientes a dicho grupo para aplicar una binarización a partir de la umbralización de Otsu, operación de apertura y esqueletización. Dichos segmentos son clasificados por una red convolucional ResNet18.

3. OBJETIVOS

3.1. Objetivo general

Desarrollar una metodología para detectar e identificar coros de peces usando técnicas de inteligencia computacional en audios marinos.

3.2. Objetivos específicos

- Analizar las técnicas de preprocesamiento, características y algoritmos de identificación automática de fonaciones de peces más relevantes en el estado del arte.
- Determinar presencia de fuentes de ruido (embarcaciones) en audios para no tenerlos en cuenta al momento de detectar posibles fonaciones de peces con un algoritmo de inteligencia computacional.
- Diseñar un método para identificar fonaciones de peces usando una técnica no supervisada de inteligencia computacional.
- Definir una métrica para distinguir características y parámetros espectrales y temporales útiles para la identificación automática de fonaciones de peces.
- Implementar un aplicativo de tal manera que el usuario pueda elegir la sensibilidad del algoritmo de identificación de acuerdo con su necesidad.

4. MATERIALES Y MÉTODOS

4.1. Base de datos

Para el desarrollo del proyecto se hacen uso de cinco (5) grabaciones tomadas con hidrófonos por investigadores del Centro de Investigaciones Biológicas del Noroeste (CIBNOR) en la Ensenada de La Paz, México, en Julio de 2018 entre los días 17 y 18 con una frecuencia de muestreo de 44.1 kHz.

4.1.1. Descripción de las grabaciones originales

Los audios fueron captados con una frecuencia de muestreo de 44.1 kHz. De los cinco audios de los que se disponían para este estudio las primeras cuatro (4) grabaciones tienen una duración de 06:45:46 mientras que la quinta grabación tiene una duración de 03:52:44. Con el fin de tener un manejo más eficiente de la base de datos se segmentaron las grabaciones en audios de un (1) minuto de duración obteniendo un total de 1,856 elementos que alcanzan frecuencias de hasta 22.05 kHz en donde se encuentran silbidos, Burst, Clicks y coros de peces.

4.1.2. Casos de Estudio

4.1.2.1. Identificación de Embarcaciones

Se disponen de 39 audios de un minuto de duración que contienen embarcaciones. Se realizó un entrenamiento para dos modelos de regresión logística, el primero fue entrenado con datos balanceados, teniendo en cuenta que la falta de disponibilidad de datos de embarcaciones se entrenó con 39 audios con embarcaciones y 39 audios sin embarcaciones elegido aleatoriamente. Para el segundo modelo se entrenó con el 20% de audios de la base de datos, es decir, con 371 audios de los cuales 39 contienen embarcaciones y 332 son grabaciones sin embarcaciones. En la etapa de prueba se hizo uso de los 1,856 audios de un minuto de duración que conforman la base de datos disponible.

4.1.2.2. Identificación de Coro de Peces 1

La metodología propuesta detecta la presencia o ausencia de coros de peces, en este caso, Peces 1, en segmentos obtenidos a partir de una etapa de segmentación expuesta más adelante. Teniendo esto presente, se realizó el entrenamiento de la red neuronal ResNet18

con 2,510 segmentos, de los cuales 1,747 no contienen la clase Peces 1, mientras que 763 de estos segmentos tienen presencia de la clase. Dichos segmentos fueron obtenidos a partir de 46 audios seleccionados de tal manera que se pudiese obtener el mayor balance entre las clases posible. Sin embargo, el balanceo es complejo debido a que los segmentos en los audios tienden a contener mayor cantidad de "NoPeces1" que de "Peces1".

Para la prueba de este modelo se hizo uso de 301 segmentos de los cuales 198 no contienen la fonación de Peces 1, mientras que 103 contienen dicha fonación.

4.1.2.3. Identificación de Coro de Peces 2

En este caso también se detecta la presencia o ausencia de Peces 2 en segmentos obtenidos a partir de la etapa de segmentación. Se utilizaron 602 segmentos de los cuales 58 no contienen la clase Peces 2 mientras que 544 la contienen. Contrario al caso de Peces 1, los audios que contienen este sonotipo presentan una mayor cantidad de segmentos con la fonación de Peces 2 que sin ella haciendo complejo un balance de los datos.

Para la prueba de este modelo se hizo uso de 1,648segmentos de los cuales 291 no contienen la fonación de Peces 1, mientras que 1357 contienen dicha fonación.

4.2. Metodología Propuesta

En este estudio se busca identificar y detectar dos clases de sonotipos denominados Peces 1 y Peces 2, cuyas fonaciones se observan en la figura 1. La metodología propuesta para la detección e identificación de coros de peces mediante técnicas de inteligencia computacional en audios marinos se divide en dos etapas fundamentales: Entrenamiento y Reconocimiento. Para el entrenamiento se tienen diferentes etapas en donde un algoritmo de clustering se complementa con una red neuronal convolucional (CNN – Convolutional Neural Network). En un principio se realiza una división de los audios captados por los hidrófonos, luego se identifican los audios que contienen embarcaciones como fuente de ruido los cuales son descartados. Posterior a esto, se da una segmentación de los espectrogramas aplicando *Acoustic Event Detection* (AED). A partir de los segmentos obtenidos se extraen frecuencia mínima, máximo, dominante y 24 Coeficientes Cepstrales de Frecuencia Lineal (LFCC) para ser usadas como entrada al algoritmo de clustering, Learning Algorithm for Multivariate Data Analysis (LAMDA 3π). Dicha técnica propone

clusters y un experto analiza qué cluster corresponde al grupo de interés, en este caso, Peces 1 o Peces 2.

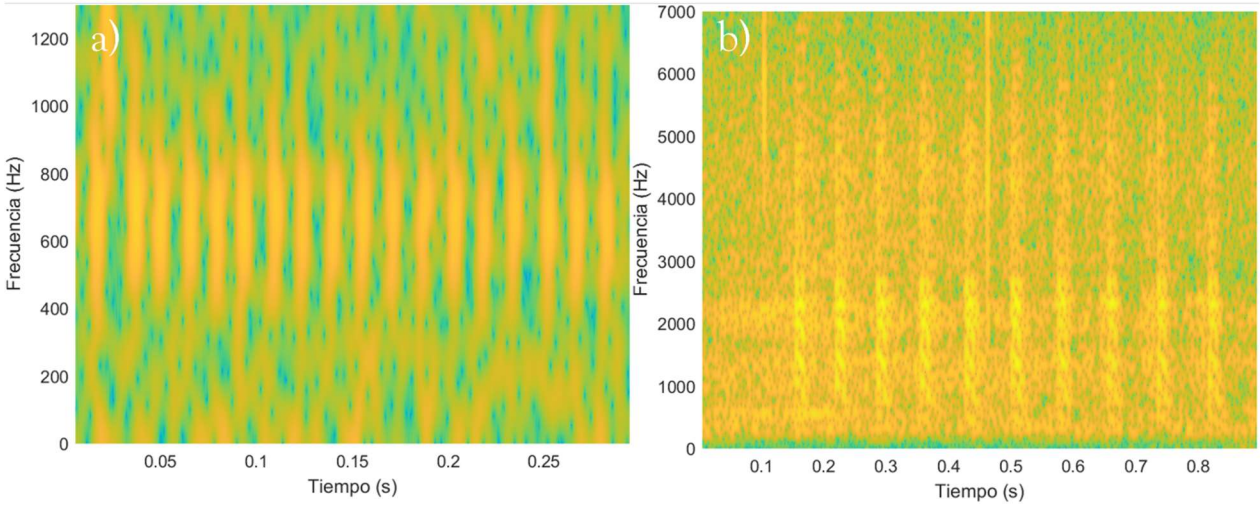

Fig. 1. Espectrograma de (a) segmento con fonaciones de Peces 1 y (b) segmento con fonaciones de Peces 2.

Los segmentos que fueron agrupados en el cluster seleccionado son procesados mediante técnicas de PDI, etiquetados e ingresados a una CNN ResNet18. Para el caso de reconocimiento se sigue el mismo flujo de trabajo con la excepción de que el cluster ya se encuentra identificado y la CNN se encuentra entrenada por lo que las etapas de identificación de cluster y etiquetado de imágenes son obviadas. Todo esto es aplicado en el software MATLAB R2020b.

4.2.1. Entrenamiento

Esta etapa es necesaria para la construcción del modelo que realizará la detección de las clases objetivo de este estudio (Peces 1 y Peces 2). Sin embargo, no se espera que sea utilizada por un usuario, sino que se busca entrenar el modelo de inteligencia computacional para su uso en la etapa de reconocimiento.

4.2.1.1. Recorte de Audios

En la recolección de audios con hidrófonos realizada por investigadores del CIBNOR se obtienen grabaciones con una duración mayor a seis horas lo que hace que el procesamiento requiera de una mayor capacidad computacional. Con el fin de aumentar la eficiencia en el algoritmo se dividen dichas grabaciones en audios de un minuto de duración.

4.2.1.2. Identificación de Embarcaciones

Las embarcaciones representan ruido responsable de la disminución del rendimiento obtenido por los algoritmos de inteligencia computacional, en este caso, el algoritmo de *clustering* y la CNN. Este ruido causado por embarcaciones se presenta en todas las frecuencias, tal y como se muestra en la figura 2. En aras de disminuir fuentes de ruido que confundan al algoritmo, se descartan las grabaciones que contengan embarcaciones y, para la identificación de señales de audio con presencia de estos sonotipos, se aplica una regresión logística con el fin de detectar su presencia o su ausencia.

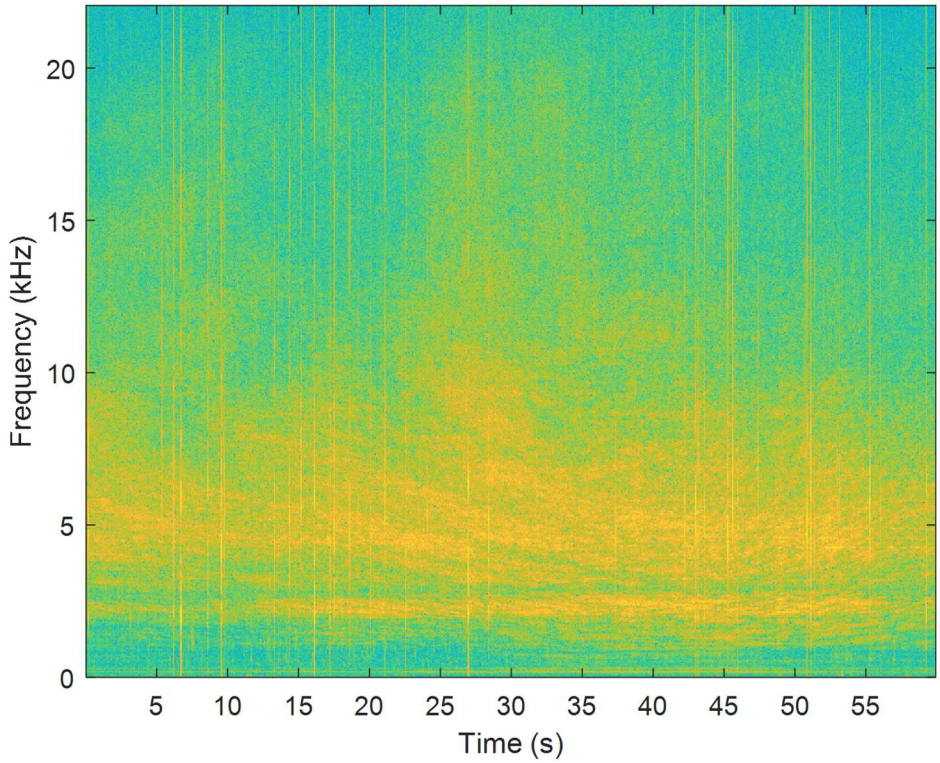

Fig. 2. Espectrograma de embarcación captada por hidrófonos.

Un modelo de regresión logística utiliza la función sigmoide con el fin de modelar la relación entre un conjunto de características o predictores y una etiqueta binaria. En el entrenamiento se encuentra una serie de coeficientes que aumenten la probabilidad de clase positiva o la disminuyan para cada predictor. El modelo se define como se muestra en la ecuación 1.

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \qquad Ecuación\ 1$$

Donde:

- $P(y=1)$ es la probabilidad de que sea una clase positiva
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes del modelo
- $x_1, x_2, \ldots, x_n$ son los predictores o características

Para dicha detección son extraídos 16 MFCC y la mediana de los datos de PSD a cada audio. Los MFCC son extraídos debido a que estas características también fueron utilizadas por Shi y Fan [17] quienes identificaron sonotipos como lluvia y vehículos blindados, sonidos similares a los producidos por las embarcaciones, obteniendo una tasa de reconocimiento de 95%. Por otro lado, como se observa en la figura 3, la distribución del PSD para ambos casos es diferente puesto que, ante una embarcación, la media del PSD aumenta alrededor de los 5 kHz, con respecto a los audios con ausencia de este fenómeno.

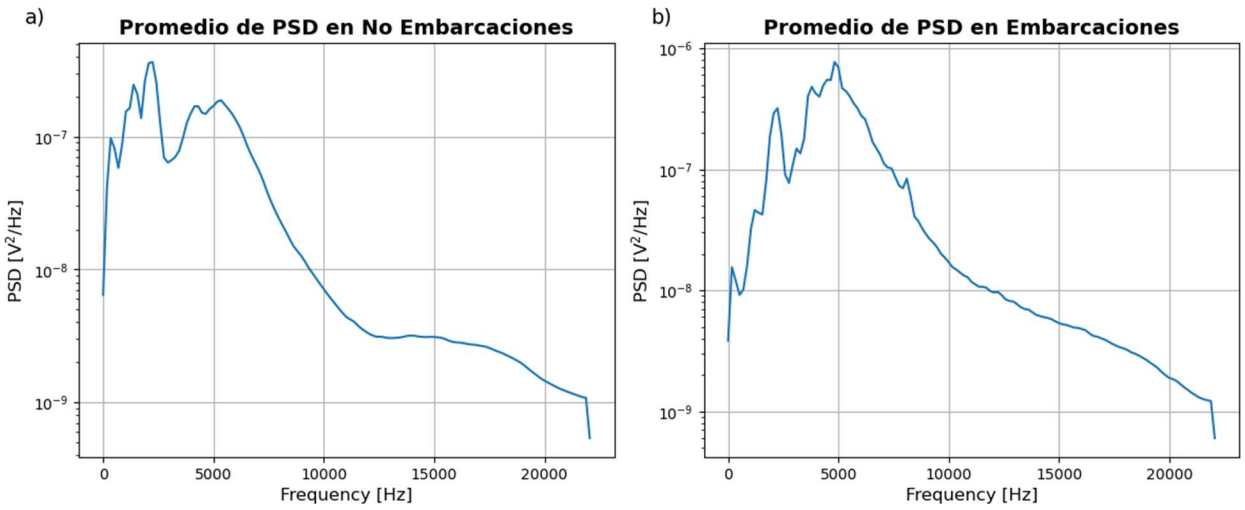

Fig. 3. Media del comportamiento de PSD en grabaciones (a) con ausencia de embarcaciones y (b) con presencia de embarcaciones.

Para el cálculo de los coeficientes cepstrales de Mel se aplica, inicialmente, una transformada de Fourier de tiempo corto (STFT) a la señal de audio ($x[n]$) utilizando una ventana periódica tipo *Hamming* ($w[n]$) de una longitud igual a 13 (N) puntos al igual que la longitud de la transformada y un *overlap* de 8 puntos. Para esto se aplica la ecuación 2.

$$X[k,m] = \sum_{n=0}^{N-1} x[n] \cdot w[n-m] \cdot e^{-j2\pi \quad /N} \qquad Ecuación\ 2$$

Una vez aplicada la STFT se expresan las frecuencias lineales *f* en la escala de Mel aplicando la ecuación 3.

$$\nu_{mel} = \frac{1000}{\log(2)} \cdot \log\left(1 + \frac{f}{1000}\right) \qquad Ecuación\ 3$$

Posterior a esto se crean los bancos de filtro Mel los cuales son, en este caso particular, 16 filtros triangulares definidos por su centro de frecuencia multiplicado por la magnitud de la STFT. Estos productos son sumados obteniendo la energía de cada filtro *(E_n)*. Por último, se obtienen los coeficientes cepstrales *(C[m])* aplicando una transformada discreta de Fourier (DFT) al logaritmo de la energía calculada para cada filtro como se muestra en la ecuación 4.

$$C[m] = \sum_{n=0}^{N-1} \log(E_n) \cdot e^{-j2\pi kn/N} \qquad Ecuación\ 4$$

Por otro lado, para el cálculo del PSD se aplica la técnica de Welch en donde, nuevamente, se calcula la STFT *(X[k, m])* usando la ecuación 2 con una longitud de la transformada de 8,192 puntos mientras que la ventana tiene una longitud de 4,096 tipo *Hamming* y un *overlap* de 50%. A estos valores resultantes se le calcula la amplitud y posteriormente se eleva al cuadrado y se realiza un promedio temporal de estos valores.

La distribución de las características descritas anteriormente para grabaciones con embarcaciones y sin embarcaciones se muestra en la figura 4.

Al algoritmo de regresión logística se le aplica una validación cruzada en donde los datos son divididos en 5 *Folds* o grupos de muestras de igual cantidad de datos con el fin de garantizar que los resultados son independientes de la partición entre datos de entrenamiento y prueba. Dicho proceso fue realizado para el entrenamiento y validación de dos modelos, el primero entrenado con una base de datos balanceada; mientras que el segundo modelo fue entrenado con unos datos desbalanceados que representan el 20% de la cantidad total de audios.

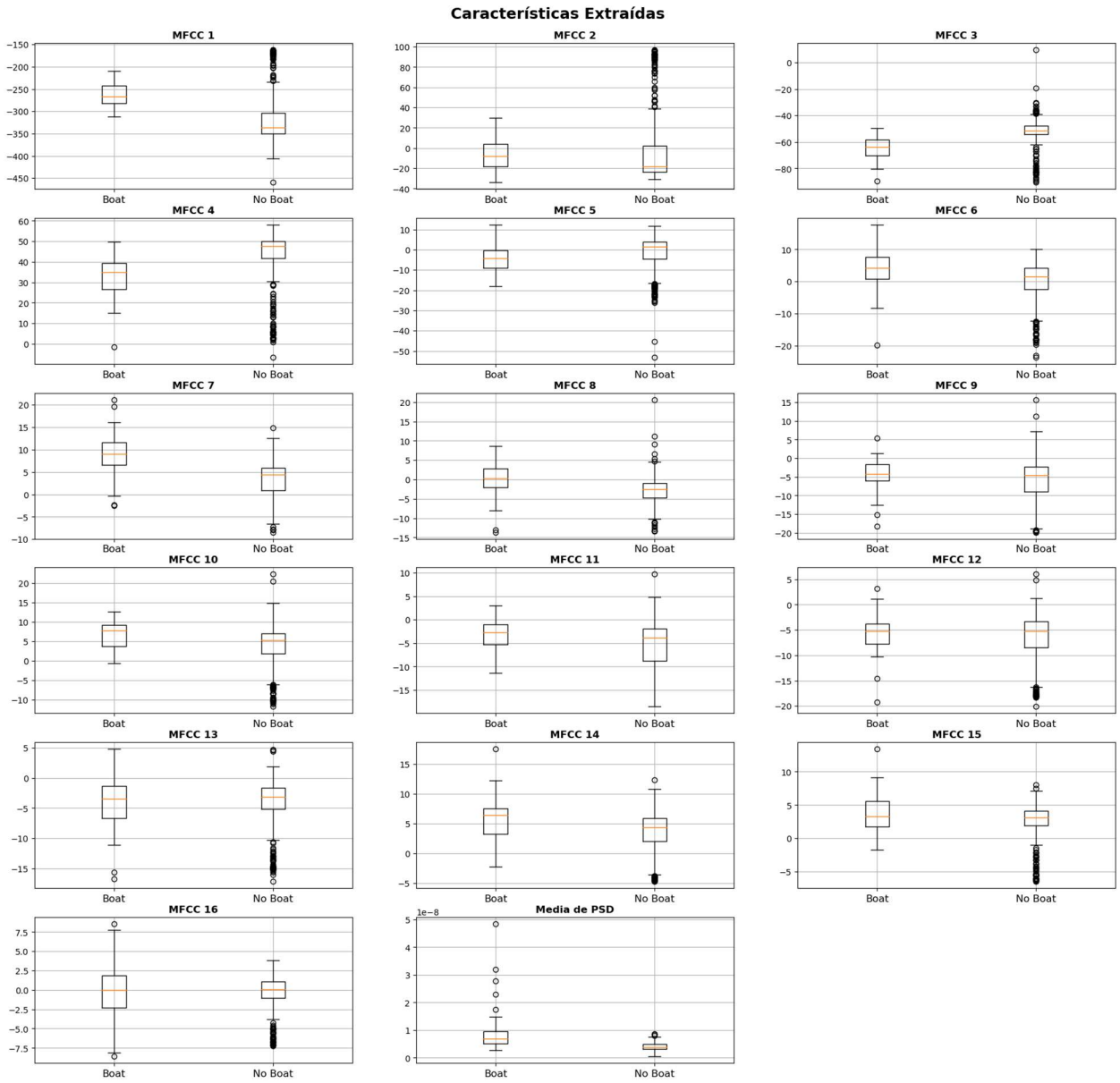

Fig. 4. Distribución de las características para las clases *Boat* y *No Boat*

4.2.1.3. Segmentación

Los sonotipos producidos por peces pueden variar en longitud y banda de frecuencia por lo que se hace necesario segmentar los audios captados con el fin de obtener la sección en tiempo y frecuencia del sonotipo permitiendo adquirir características representativas de este para alcanzar un buen rendimiento del algoritmo de agrupación. Es por esto por lo que el método de segmentación representa un paso importante en la identificación de fonaciones de peces.

El método de segmentación a usar es una modificación del algoritmo de Detección de Eventos Acústicos (AED) [3], [18], el cual extrae el espectrograma de cada audio y lo trata como una imagen para aplicar técnicas de procesamiento de imágenes. Dicha propuesta se compone de varias etapas: Reducción de ruido de fondo, aplicación de la umbralización Otsu para la binarización, operaciones morfológicas de Apertura y Cierre, y aplicación de *Bounding Boxes*.

En esta etapa se tiene por objetivo identificar los segmentos en donde se produce un sonotipo con el fin de obtener las características específicas de cada segmento permitiendo una mayor precisión en el momento de la aplicación del algoritmo de agrupación. Para esto se extrae el espectrograma de cada audio usando una ventana *Hamming* de tamaño igual a 1024, un *overlap* de 512 y 2048 puntos para la Transformada de Fourier de Tiempo Corto.

Como es bien sabido, las fonaciones de diferentes animales no tienen una duración y rango de frecuencia igual por lo que algunos sonidos ocupan una mayor área en el espectrograma que otros. Por esta razón, aunque las etapas son iguales, los parámetros varían para segmentos de área pequeña y área grande lo que permite una mayor adaptabilidad del algoritmo a los sonotipos presentes en las grabaciones.

Filtrado

Se aplica un filtro gaussiano el cual es un filtro de suavizado lineal que se basa en la función de distribución gaussiana puesto que elimina ruidos que siguen una distribución normal [19]. Para dicho filtro se aplica un *kernel* o núcleo cuadrado de 13x13 pixeles con desviación estándar igual a tres ($\sigma = 3$). También se implementa un algoritmo de sustracción espectral ya que el filtro gaussiano no elimina ciertas fuentes de ruido que se encuentran presentes en una banda de frecuencia amplia lo cual es considerado como ruido de fondo o *background noise*. Para esto se implementó la sustracción espectral propuesta por Xie [18]. Al aplicar dicho filtro y sustracción espectral se obtiene lo mostrado en la figura 5.

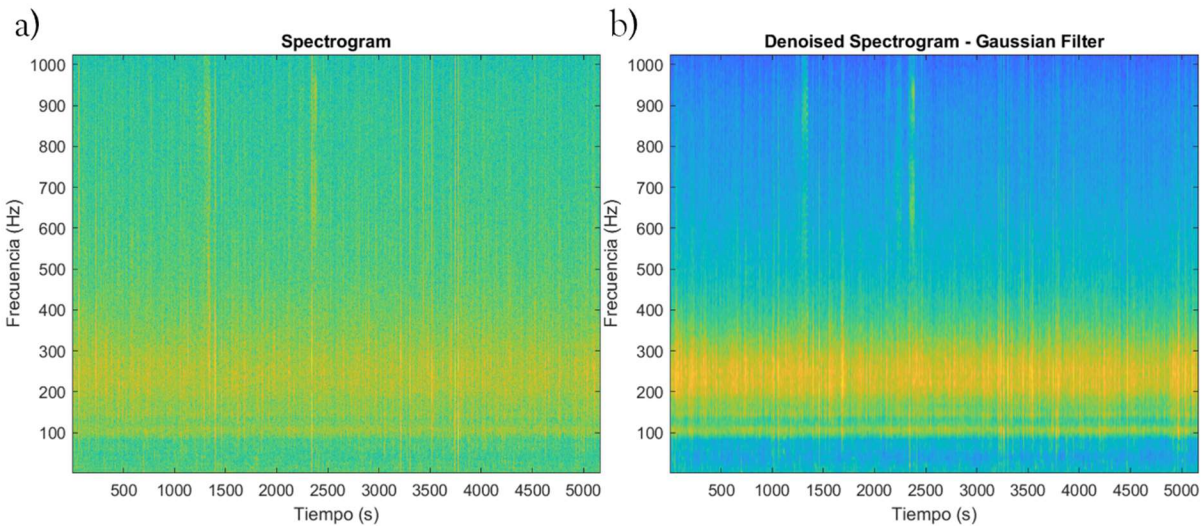

Fig. 5. Filtrado de un espectrograma aplicando filtro gaussiano y sustracción espectral

Binarización

Una vez es aplicado el filtro gaussiano y la sustracción espectral, se procede a una Binarización del espectrograma en donde, cuando la intensidad de un pixel supera un umbral, el valor de ese pixel será igual a uno. En caso contrario, el valor será igual a cero. Dicho umbral es calculado aplicando la Umbralización de Otsu el cual es un criterio discriminante lineal que supone que la imagen está compuesta por objetos (primer plano) y fondo, sin tener en cuenta la heterogeneidad y diversidad del fondo. En resumen, el método de Otsu segmenta la imagen en dos regiones, una clara y otra oscura permitiendo la binarización de la misma [20]. Este umbral se multiplica por un factor de 1.5 para segmentos pequeños y 1.1 para segmentos grandes con el fin de focalizar el procesamiento en los sonidos deseados pues, dependiendo del área del segmento, los parámetros son más o menos eficientes.

Operaciones Morfológicas

Son llamadas operaciones morfológicas porque realizan cambios en la forma y tamaño de los objetos de una imagen que, en este caso, está binarizada por lo que las operaciones serán sobre las regiones claras o igual a uno. El *Opening* o la apertura toma las pequeñas estructuras claras y las elimina dependiendo del área que ocupan la cual es medida por la cantidad de pixeles que componen dichas estructuras. Este proceso lo realiza tomando una estructura con un área dada, aquellos fragmentos claros que no cubren al elemento son llevados a cero. En la figura 6 se ejemplifica este proceso.

Por otro lado, el *Closing* o el cierre rellena (toman un valor igual a uno) las estructuras de fondo u oscuras que no pueden contener al elemento de estructuración determinado, similar al proceso de apertura. En la figura 6 se muestran estas operaciones morfológicas.

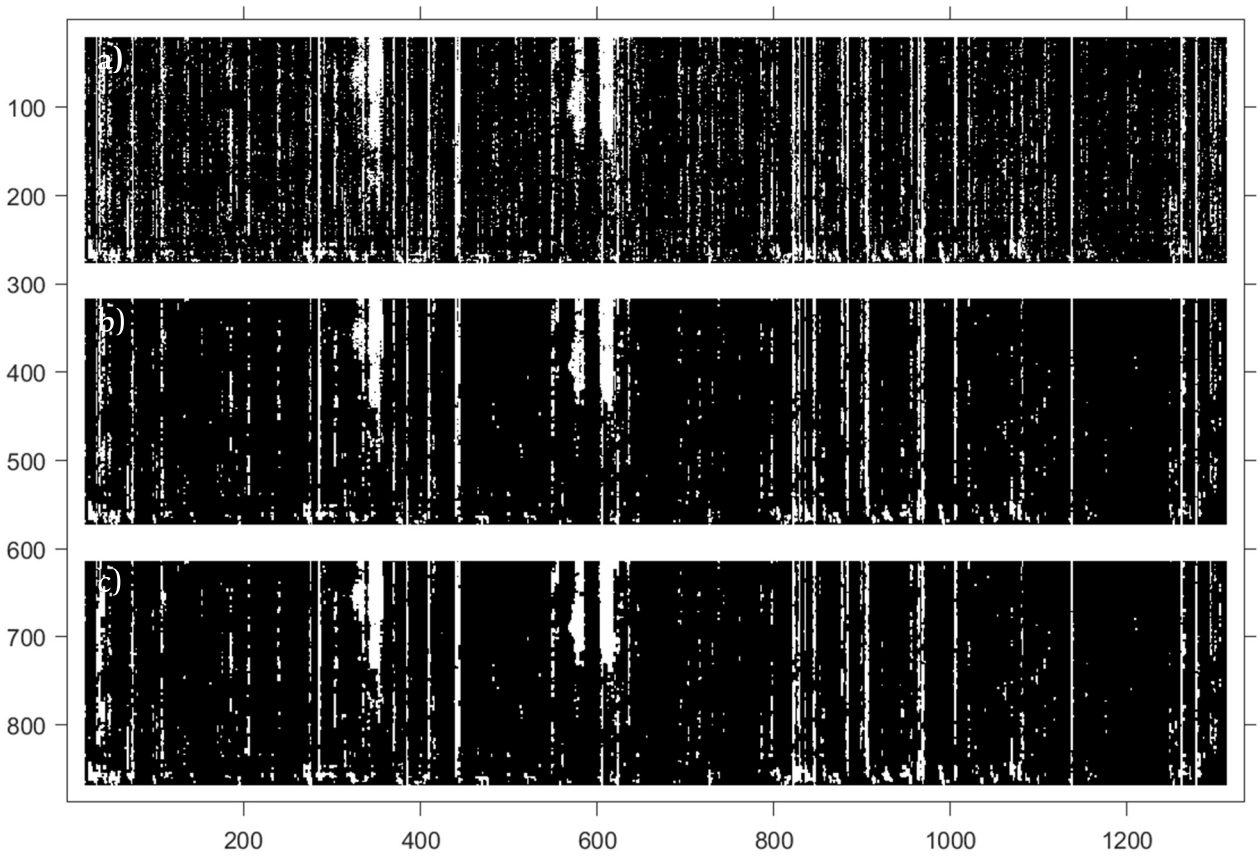

Fig. 6. Imagen original (a), proceso de apertura (b) y proceso de cierre (c).

Para la operación de *Opening* se aplica como *kernel* (núcleo) un rectángulo de 9x7 para el caso de segmentos grandes, mientras que para segmentos pequeños se aplica un rectángulo de 7x5. Para la operación de *Closing* se usa como *kernel* (núcleo) un cuadrado de 8x8 para segmentos grandes y de 3x3 para pequeños.

Bounding Boxes

Aplicando esta técnica de PDI se identifica una región de interés con base a la intensidad de la imagen que, en ese caso, se encuentra binarizada. Para segmentos grandes se identifican regiones con un área entre 10,000 y 40,000, mientras que, para segmentos pequeños, esta técnica busca regiones de alta intensidad que abarquen un área entre 500 y 10,000. Para ambos casos se determina una excentricidad mayor a 0.5 y un *Extent* mayor a 0.3. Con esto se obtienen las coordenadas de los *bounding boxes* que encierran las

fonaciones y dichas coordenadas están dadas en frecuencia y tiempo teniendo presente que todo el procesamiento se ha realizado con los datos del espectrograma de cada audio siendo que los valores en el eje *y* corresponden a los valores de frecuencia, mientras que los valores en el eje *x* a los valores de tiempo de acuerdo a la ventana y frecuencia de muestreo con la que se graficó el espectrograma y que están expuestos en la sección 4.2.1.3.

Los *bounding boxes* se definieron como rectángulos y delimitan un área en el que se encuentra un objeto de interés en la imagen. Son usados principalmente en visión por computadora y en detección de objetos con el fin de obtener la ubicación y el tamaño de un objeto representativo en una imagen. Por lo general se definen por las coordenadas de dos puntos: (1) coordenada de la esquina superior izquierda, (2) coordenada de la esquina inferior derecha [21].

4.2.1.4. Clustering – LAMDA

LAMDA es un algoritmo de lógica difusa de agrupamiento que tiene la capacidad de recibir características mixtas, es decir, datos lingüísticos y numéricos, siempre y cuando estos sean normalizados. Su funcionamiento se basa en funciones de distribución siendo la binomial la que presenta un mejor funcionamiento. Así, se define una función binomial diferente para cada atributo de cada cluster cuya media es actualizada cada que un dato es agrupado en el cluster, es decir es redefinida [22].

Inicialmente, LAMDA calcula grados de adecuación marginal (MAD) los cuales representan la contribución de cada característica a cada cluster existente. Un MAD para un cluster *c* y una característica *i*, con un vector normalizado $\hat{y}_i$ y una media ρ es calculada con la ecuación 5.

$$\mathrm{MAD}_{c,i} = \rho_{c,i}^{\hat{y}_i}(1-\rho_{c,i})^{1-\hat{y}_i} \qquad \textit{Ecuación } 5$$

En un primer momento se tiene un cluster predefinido, la Clase No Informativa (NIC), la cual presenta el mismo MAD para cualquier valor de una característica (0.5). El primer elemento que ingresa siempre será asignado a la NIC y cada que un dato sea asignado a esta clase se genera un nuevo cluster tomando como parámetros iniciales los definidos por la NIC ($\rho_{0,i}$) y modificados con los parámetros del nuevo dato entrante siguiendo la ecuación 6 en donde se actualiza la media característica del primer *cluster* ($\rho_{1,i}$) [3].

$$\rho_{1,i} = \frac{(\hat{y}_i + \rho_{0,i})}{2} \qquad \textit{Ecuación 6}$$

Cada vez que un nuevo segmento es analizado se obtienen los MAD ($M_{c,i}$) de cada característica para un *cluster* existente y son combinados aplicando un operador de agregación cuyo resultado es conocido como grado de adecuación global (GAD) de un elemento a un *cluster*. Esto se muestra en la ecuación 7 en donde N_f es el número de características que ingresan al algoritmo. El nuevo dato será agrupado al cluster con el que presente un mayor GAD (g_c) y se actualizan los parámetros de dicho cluster [3].

$$g_c = \frac{\prod_{i=1}^{N_i} M_{c,i}}{\prod_{i=1}^{N_i} M_{c,i} + \prod_{i=1}^{N_i}(1 - M_{c,i})} \qquad \textit{Ecuación 7}$$

Por otro lado, cuando un dato ingresa a un *cluster* creado anteriormente, los parámetros de dicho *cluster* son actualizados al aplicar la ecuación 8 en donde ${\rho_{c,i}}^{(k)}$ es la media actualizada, ${\rho_{c,i}}^{(k-1)}$ es la media anterior y ${n_c}^{(k)}$ es el número actual de elementos que conforman el *cluster.*

$${\rho_{c,i}}^{(k)} = {\rho_{c,i}}^{(k-1)} + \frac{\hat{y}_i + {\rho_{c,i}}^{(k-1)}}{{n_c}^{(k)}} \qquad \textit{Ecuación 8}$$

En esta etapa se lleva a cabo un aprendizaje no supervisado de agrupamiento aplicando un algoritmo difuso. Para esto, a cada segmento encontrado en la etapa de segmentación se le extrae un conjunto de características con las que el algoritmo realizará el agrupamiento. Dichas características son las propuestas por Guerrero [3] las cuales son: Frecuencia mínima, frecuencia máxima, frecuencia dominante y 24 LFCC. Para la extracción de los LFCC se sigue el mismo procedimiento para la extracción de MFCC expuesta en la sección 4.2.1.2., con la diferencia de que los coeficientes cepstrales son extraídos con la ecuación 9 en donde no se aplica logaritmo a la energía de los filtros.

$$C[m] = \sum_{n=0}^{N-1} E_n \cdot e^{-j2\pi k \ /N} \qquad \textit{Ecuación 9}$$

De acuerdo con la hipótesis de nicho acústico, cada especie emite sonidos en una banda de frecuencia en la cual otras especies no producen sonidos. Por esta razón las frecuencias mínima, máxima y dominante permiten caracterizar significativamente un sonido producido por una especie. Por otro lado, los LFCC permiten tener una representación compacta y eficiente de las características acústicas y fonéticas importantes en una señal de audio. Estos coeficientes, a su vez, se muestran linealmente en la escala de frecuencia posibilitando tener información relevante tanto para frecuencias bajas como altas.

4.2.1.5. Identificación de Cluster

LAMDA 3π propone un conjunto de grupos o *clusters* de los cuales el experto escogerá aquel o aquellos que mejor representen la fonación objetivo, en este caso, Peces 1 y Peces 2. Para este proceso se hizo uso del aplicativo propuesto por Guerrero [3] denominado Aureas en donde la identificación del *cluster* puede realizarse observando el elemento representativo del *cluster,* graficando los segmentos en el espectrograma de un audio de tal manera que encuentre un patrón en los segmentos agrupados y que corresponda a la fonación deseada, o realizando un barrido de los segmentos en Aureas.

Vale la pena aclarar que el uso del aplicativo Aureas sólo es usado en esta etapa de entrenamiento. El aplicativo que se implementará en la etapa de reconocimiento y que estará disponible para usuarios no requiere de este software.

Una vez identificado el *cluster* se entrena el algoritmo almacenando las especificaciones de dicho grupo para ser usadas al momento del reconocimiento, es decir, en la etapa de prueba. Este paso es importante para disminuir la cantidad de segmentos que entrarán a la red convolucional lo que permitirá ahorrar recursos computacionales.

4.2.1.6. Extracción de Segmentos como Imágenes

Los segmentos pertenecientes al *cluster* identificado en el paso anterior son extraídos como imágenes manteniendo un área constante con el fin de unificar los parámetros del procesamiento posterior para todos los segmentos. En el caso de la clase Peces 1 se establece una duración de 300 ms para cada segmento comprendiendo frecuencias menores a 1.3 kHz debido a que el nicho acústico de este sonotipo se encuentra

en estas frecuencias. Por el otro lado, para Peces 2 la duración se establece en 450 ms con frecuencias menores a 7 kHz.

4.2.1.7. PDI

Este procesamiento es similar al presentado en la etapa de segmentación, pero, en lugar de realizar PDI en el espectrograma completo, se realiza sobre cada segmento extraído como imagen. El primer paso es binarizar la imagen con base en un umbral calculado mediante la técnica Umbralización de Otsu, la cual propone un valor para este umbral que será multiplicado inicialmente por un factor de 1.9 con el fin de disminuir el ruido de las grabaciones provocado por el movimiento del agua. El algoritmo realiza una sumatoria del valor de los pixeles que conforman la imagen, si la sumatoria es menor a 17,000, resta a este factor 0.3 hasta que la sumatoria sea mayor o igual a 17,000 disminuyendo la posibilidad de eliminación de fonaciones de baja intensidad en la Binarización.

Cuando la imagen se encuentra binarizada se aplica una operación morfológica de *opening* o apertura con un *kernel* rectangular de 100x9 debido a la forma de las fonaciones objetivo. Por último, se realiza un esqueletizado de la imagen ya que, como se observa en la figura 7 (d) la forma de los peces tiende a ser un conjunto de líneas verticales horizontales de tal manera que se le facilita el patrón a la red neuronal. Al obtener líneas como entrada a la CNN se disminuyen los recursos computacionales necesarios para la clasificación abriendo la posibilidad de reducir la cantidad de capas de la ResNet de 50, como es aplicada en la propuesta de Waddell [16], a 18.

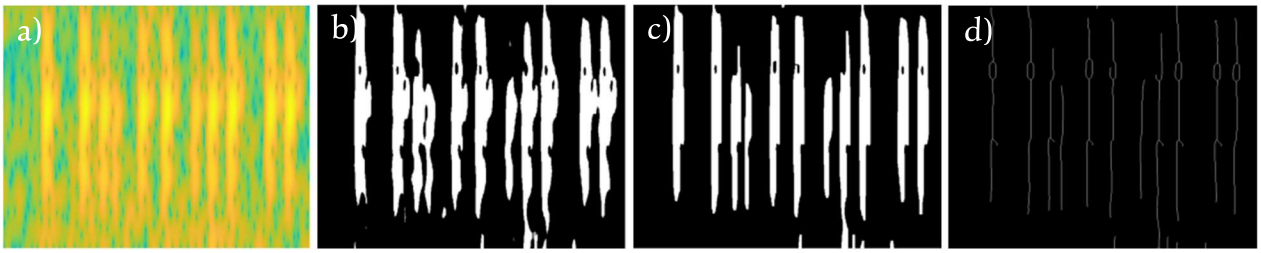

Fig. 7. Segmentos (a) original extraído del espectrograma, (b) binarizado, (c) con *opening* y (d) Esqueletizado

4.2.1.8. Etiquetado de Imágenes

Como en todo algoritmo supervisado, la CNN a utilizar requiere de un etiquetado de imágenes. Vale la pena aclarar que se entrenará una red convolucional por cada clase con el fin de detectar presencia o ausencia del sonotipo por lo que, los datos de entrenamiento

para Peces 1 tendrán como etiquetas “Peces” o “NoPeces”, mientras que para Peces 2 las etiquetas serán “Peces2” o “NoPeces2”.

4.2.1.9. Entrenamiento de ResNet18

Como se menciona en el apartado anterior, se entrenaron dos modelos diferentes para cada una de las clases objetivo: Peces 1 y Peces 2. En el primer caso se entrenó el modelo aplicando validación cruzada o *Cross-Validation* con 5 *Folds* o grupos y estableciendo una taza de aprendizaje o *learning rate* de 0.001, momentum de 0.9, tamaño de Batch igual a 20, como función de pérdida se define *Cross-entropy*, 20 épocas y optimización del descenso de gradiente estocástico con momento. Para Peces 2 se aplica la misma función de pérdida, optimizador y valores para los parámetros, exceptuando el número de épocas que se definió en 8. En este último caso no fue necesario realizar *Cross-Validation.*

Todo el flujo de trabajo de la etapa de entrenamiento se encuentra resumido en la figura 8 en donde se describen brevemente las fases de esta etapa.

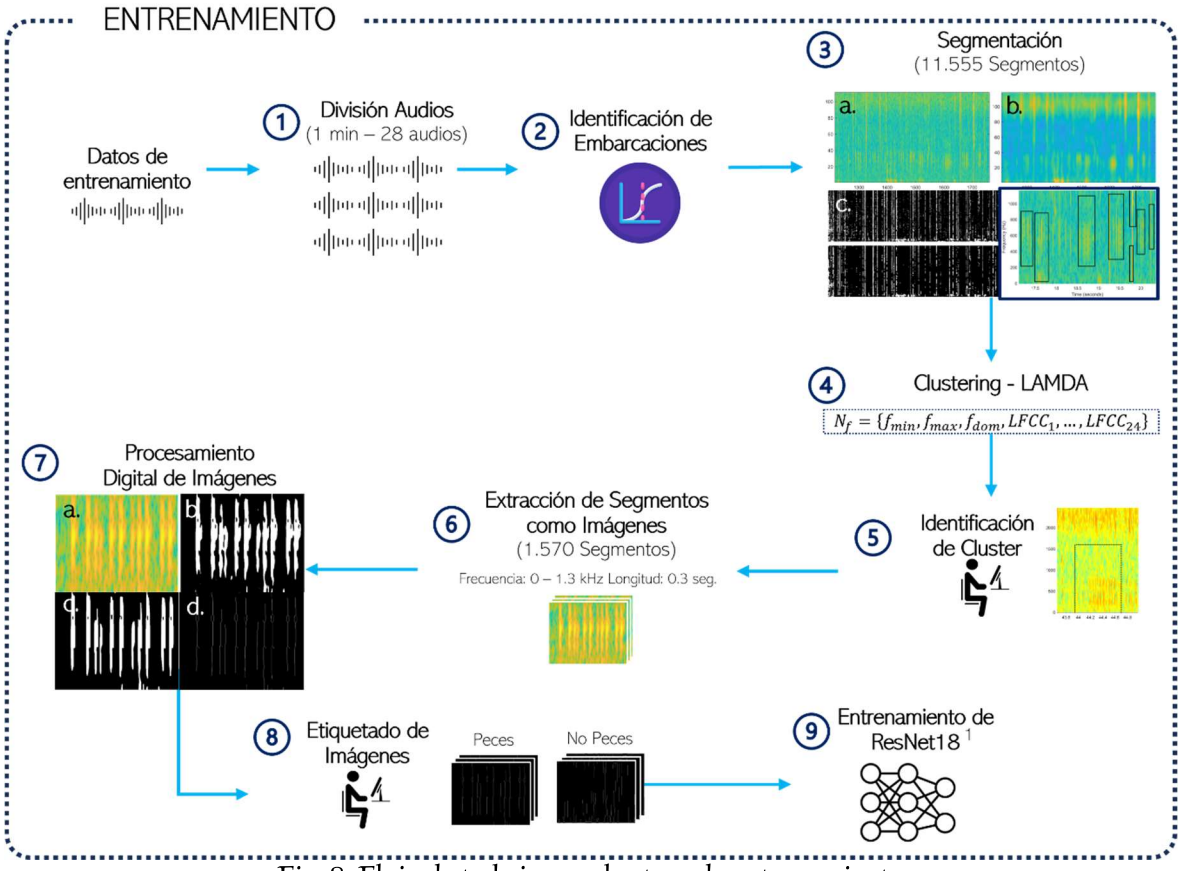

Fig. 8. Flujo de trabajo para la etapa de entrenamiento

4.2.2. Reconocimiento

Esta segunda etapa será la usada por el usuario para la detección de la clase que desee: Peces 1 o Peces 2. Para esto se solicitará la ruta de acceso a los audios en los que se desea realizar la búsqueda de fonaciones de peces, la clase de fonación a reconocer y un parámetro de exigencia. Para los resultados mostrados en la sección 5 los datos fueron sometidos a esta segunda etapa y son diferentes a los utilizados en la etapa de entrenamiento puesto que son los elementos reservados para prueba los cuales son expuestos en la sección 4.1.2.

Las fases del reconocimiento son similares a las expuestas en la sección de entrenamiento con la diferencia de que los modelos no son entrenados, sino que predicen.

4.2.2.1. Recorte de Audios

Igual a la sección 4.2.1.1. las grabaciones son divididas en señales de audio de un minuto de duración con el fin de hacer más eficiente el algoritmo.

4.2.2.2. Identificación de Embarcaciones

El modelo de regresión logística entrenado en la sección 4.2.1.2. predice las señales de audio de un minuto en donde hay presencia de ruido por embarcaciones y elimina dichos archivos con el fin de aumentar el rendimiento de los modelos de agrupamiento y *machine learning*.

4.2.2.3. Segmentación

Esta etapa se aplica a los espectrogramas de los audios restantes y el procedimiento, al igual que los valores de los parámetros, no varía con respecto a la sección 4.2.1.3. a partir de lo cual se obtienen los segmentos que serán clasificados por el algoritmo de *clustering*.

4.2.2.4. Reconocimiento – LAMDA

El algoritmo que realiza un reconocimiento del *cluster* almacenado en la etapa de entrenamiento, más específicamente, en la sección 4.2.1.5. con lo que se obtienen los segmentos clasificados como parte de dicho *cluster* a partir del cálculo de GAD mostrado en la ecuación 7 con las características extraídas como se muestra en la sección 4.2.1.4.

4.2.2.5. Extracción de Segmentos como Imágenes con parámetro de exigencia

Nuevamente, la extracción de segmentos sigue el mismo procedimiento expuesto en el apartado 4.2.1.6. con la diferencia de que en la etapa de reconocimiento se solicita la introducción de un parámetro de exigencia por parte del usuario. Dicho parámetro funciona

como umbral de tal manera que los segmentos que superen este valor continúen en el flujo de trabajo y sean extraídos como imágenes, de lo contrario no son tenidos en cuenta.

El parámetro de exigencia se basa en el grado de pertenencia *(Membership)* entregado por parte de LAMDA 3π al momento del reconocimiento; a mayor valor, mayor pertenencia al *cluster*. Con esto se eliminan segmentos que No contenían el sonotipo de interés, aunque también se corre el riesgo de disminuir la cantidad de segmentos que Si contienen la fonación objetivo. En otras palabras, con un menor parámetro de exigencia se obtiene mayor cantidad de falsos positivos.

4.2.2.6. PDI

Esta etapa sigue el mismo procedimiento y valor de parámetros descritos en la sección 4.2.1.7. aplicado a los segmentos que superaron el umbral impuesto por el parámetro de exigencia del apartado anterior.

4.2.2.7. Reconocimiento de ResNet18

Por último, se detecta la presencia o ausencia del sonotipo de interés en los segmentos obteniendo la lista de los segmentos que contienen la fonación deseada, así como datos de nombre de archivo, tiempo inicial y final, y frecuencia mínima y máxima.

La etapa de reconocimiento y las subetapas que la conforman, se encuentran resumidas en la figura 9.

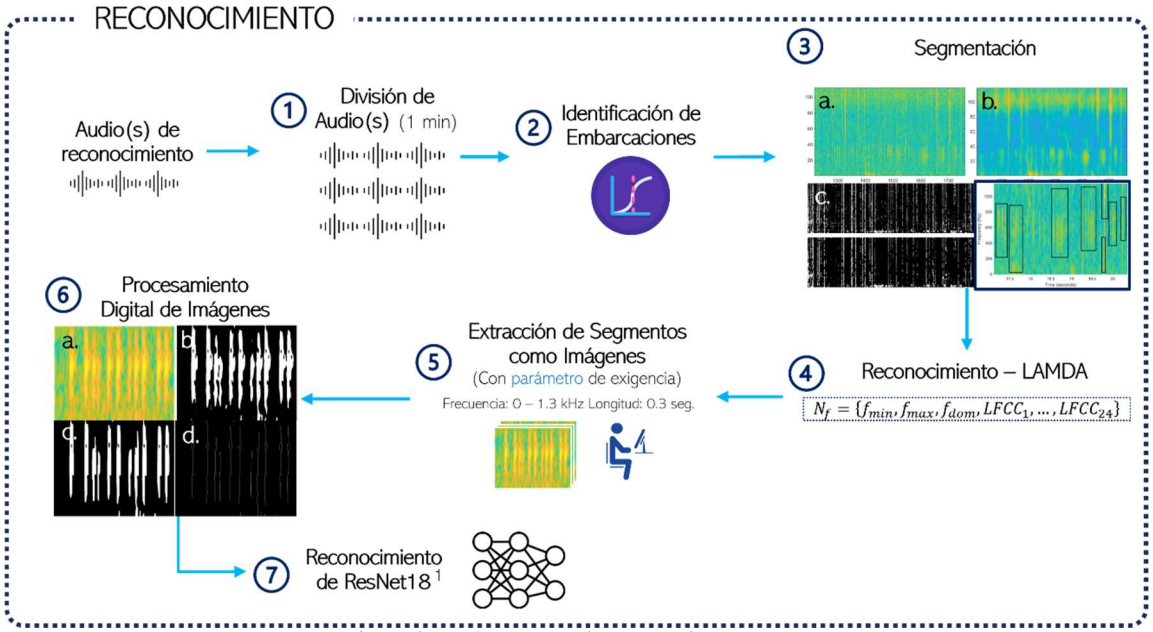

Fig. 9. Flujo de trabajo para la etapa de reconocimiento

4.3. Métricas de evaluación

La métrica con la que se evaluó el desempeño de los modelos es ***F1-Score*** ya que ofrece una relación entre el ***Recall*** y la Precisión permitiendo medir el desempeño de los modelos al momento de detectar la presencia o ausencia del sonotipo puesto que el primero da cuenta de la proporción de datos positivos ("Peces" o "Peces2") clasificados correctamente, mientras que el segundo mide la capacidad del modelo para no etiquetar incorrectamente ejemplos negativos como positivos. En la ecuación 10 se muestra el cálculo del ***Recall***, en la ecuación 11 de la Precisión y en la ecuación 12 del ***F1-Score***.

$$\text{Recall} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Negativos (FN)}} \qquad \textit{Ecuación } 10$$

$$\text{Precisión} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Positivos (FP)}} \qquad \textit{Ecuación } 11$$

$$\text{F1} - \text{Score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}} \qquad \textit{Ecuación } 12$$

Por otro lado, el *Accuracy* también fue calculado con el fin de realizar una comparación con otros estudios encontrados en la literatura. En la ecuación 13 se muestra el cálculo de esta métrica.

$$\text{Accuracy} = \frac{\text{Número de prediccione correctas}}{\text{Total predicciones}} \qquad \textit{Ecuación } 13$$

5. RESULTADOS

Para la exposición de los resultados se tendrán cuatro casos de estudio: Identificación de embarcaciones, identificación de Peces 1, identificación de Peces 2 y análisis de sensibilidad.

5.1. Identificación de Embarcaciones

Para la identificación de embarcaciones se realizó la prueba con los 1,856 audios que conforman la base de datos obteniendo un *F1-score* igual a 0.9942 y un *Accuracy* de 0.9989 para el modelo entrenado con datos no balanceados mostrando un buen desempeño al momento de identificar las embarcaciones, mientras que para el modelo entrenado con datos balanceados se obtuvo un *F1-Score* de 0.5942 por la falta de detección de no embarcaciones, y un *Accuracy* de 1 debido a la total detección de embarcaciones. En la figura 10 (b) se muestra que el modelo no pudo identificar correctamente los audios que no contenían embarcaciones por lo que se requiere un mejor entrenamiento de esta clase. Dicha estrategia se llevó a cabo al entrenar el modelo con la base de datos desbalanceada.

Es importante tener en cuenta que entre los 1,856 audios sólo existían 39 grabaciones que contenían embarcaciones por lo que, en aras de evaluar el desempeño y la robustez del modelo, es importante someter la predicción a grabaciones en diferentes lugares y con una base de datos que contenga un mayor número de embarcaciones.

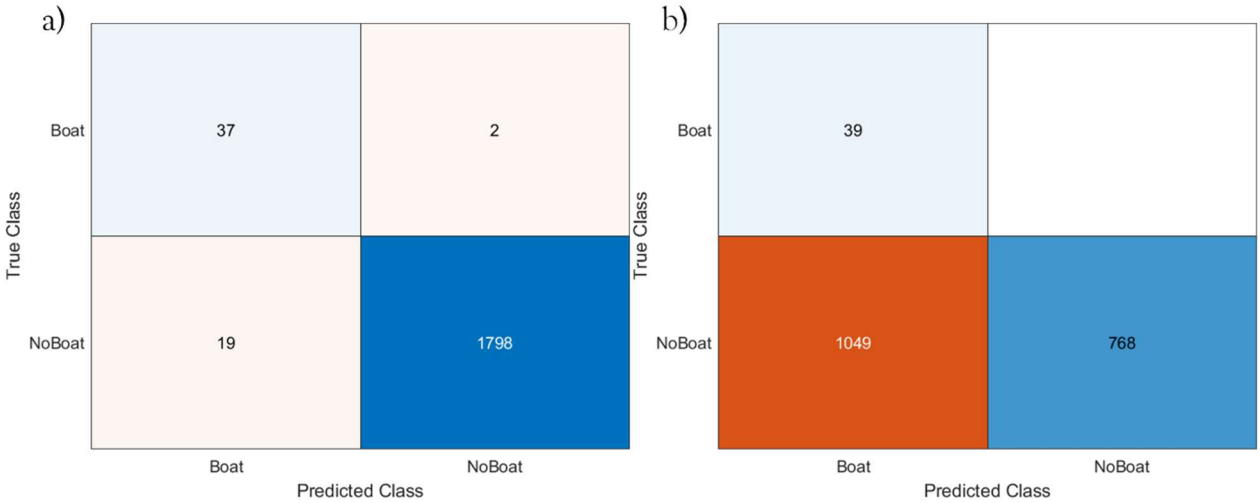

Fig. 10. Matriz de confusión de la predicción de embarcaciones por (a) el modelo entrenado con datos desbalanceados y (b) modelo entrenado con datos balanceados

Por otro lado, se observa en la figura 10 (a) que 19 audios que no contenían embarcaciones fueron mal clasificadas como embarcaciones por parte del clasificador. Dichos falsos positivos se deben a que en estas grabaciones existe una potencia espectral constante similar al presentado en los audios con embarcaciones, principalmente en frecuencias medias como se muestra en la figura 11 (a). En este caso el rendimiento del modelo puede ser mejorado si se entrena con una base de datos que contenga más embarcaciones.

En cuanto a las dos grabaciones que fueron clasificadas como no embarcaciones incorrectamente, se observa que los audios tenían una baja potencia espectral, específicamente en las frecuencias bajas ya que las embarcaciones presentes en estas señales de audios se encontraban distantes del hidrófono puesto que se estaban acercando o alejando como se muestra en la figura 11 (b).

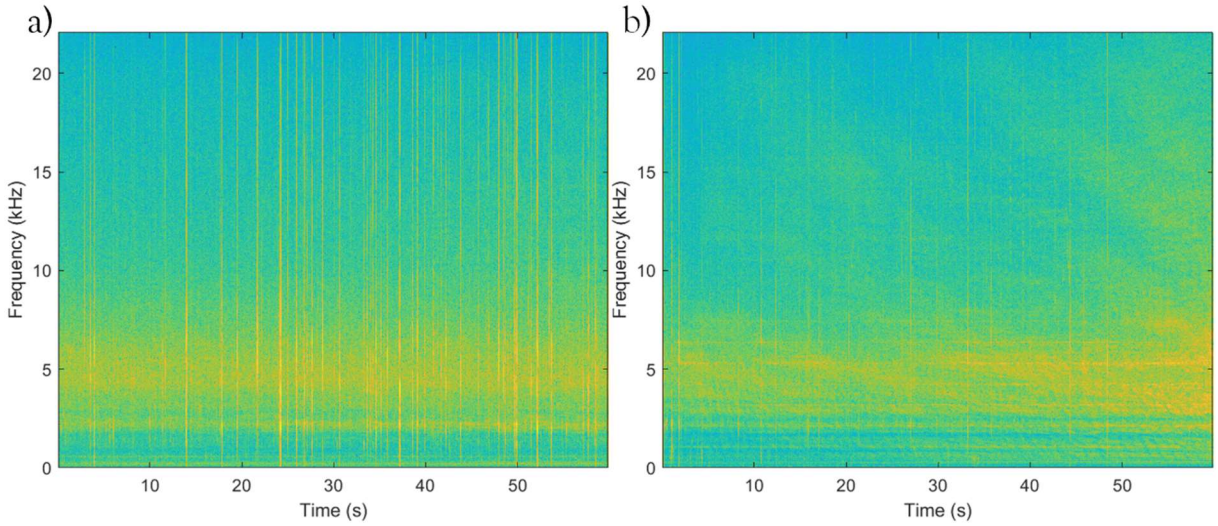

Fig. 11. Espectrograma de (a) audio clasificado como embarcación y (b) audio clasificado como no embarcación incorrectamente.

5.2. Peces 1

El entrenamiento de esta clase fue realizado mediante la técnica de *Cross-Validation* obteniendo una mediana de *F1-score* igual a 0.5956 ± 0.0418 mientras que la mediana del *Accuracy* es de 0.7531 ± 0.0165. Por otro lado, en la etapa de reconocimiento del modelo se estableció un parámetro de exigencia de 9.9 y se obtuvo un *F1-score* igual a 0.7805 y un Accuracy igual a 0.8269 con los 301 segmentos de prueba descritos en la sección 4.1.2.2. Esta diferencia entre las métricas se debe a la diferencia en la cantidad de segmentos que

fueron utilizados en cada etapa puesto que, en el primer caso, fueron usados 2,510 segmentos, mientras que para el segundo se utilizaron 301 segmentos. Como se expuso en la sección 5.1., es importante someter este algoritmo a una mayor cantidad de grabaciones para comprobar su robustez y rendimiento.

En la figura 12 se muestra que existen cuatro falsos negativos, es decir, segmentos que fueron clasificados como No Peces cuando realmente contienen fonaciones de Peces. Dichos casos se dan principalmente cuando existe una baja intensidad de las fonaciones o cuando existe una cantidad importante de ruido de fondo que enmascara los sonotipos. Esto se observa en la figura 13 (b) puesto que las fonaciones presentes tienen una baja intensidad que, en la etapa de PDI, son eliminadas al momento de realizar la binarización. Para estos casos es importante realizar una nivelación de ganancias, tanto para segmentos con baja intensidad como para aquellos de alta intensidad con ruido de fondo importante de tal manera que sea posible contrastar mejor las fonaciones aumentando el rendimiento del modelo.

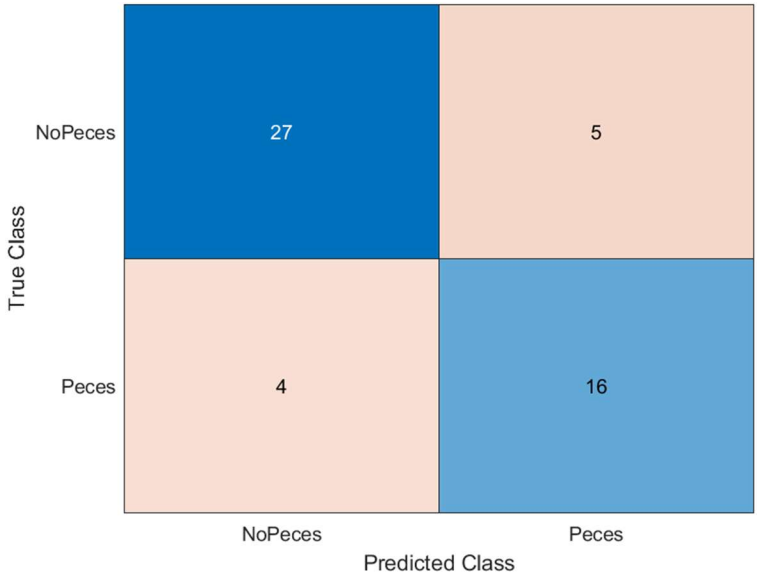

Fig. 12. Matriz de confusión de la predicción de la clase Peces 1 por el modelo propuesto

Por otro lado, en cuanto a los segmentos que fueron clasificados como peces pero que no contenían fonaciones, es decir, los falsos positivos, se observa en la figura 13 (a) que hay picos de potencia causados por el movimiento del agua posiblemente por un desplazamiento de algún objeto o ser vivo ocurrido cerca al hidrófono y que fue identificado mediante sonido de manera manual. Dicho ruido tiene una forma similar al patrón exhibido

por las fonaciones de los peces de la figura 1. Esto puede ser mejorado con una mayor cantidad de ejemplos obtenidos con la toma de datos en diferentes zonas y en diferentes épocas enriqueciendo el entrenamiento y la generalización del modelo de inteligencia computacional.

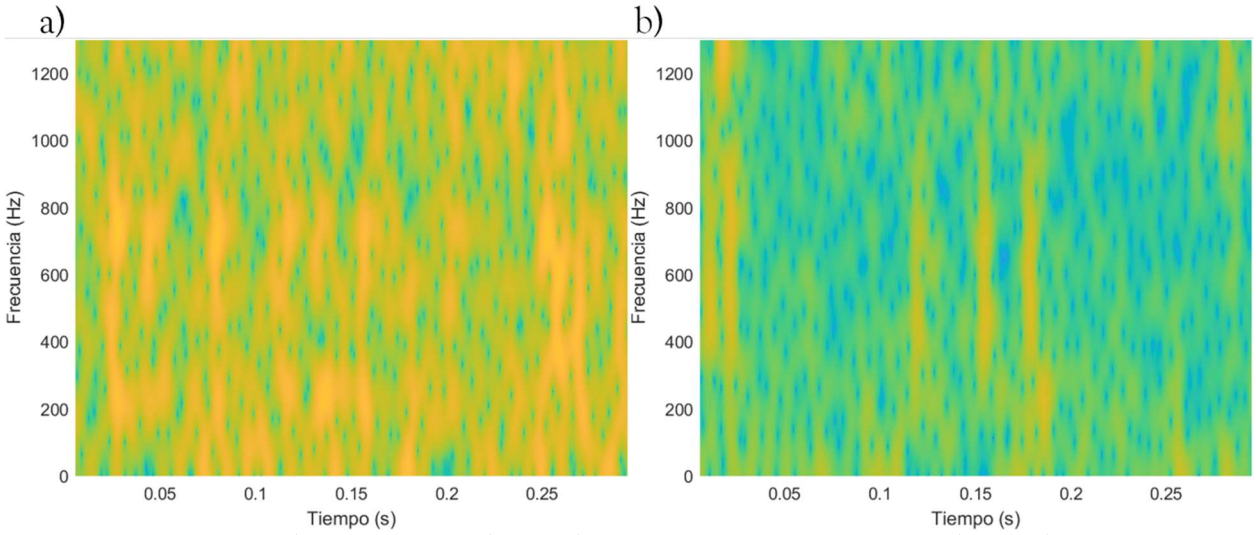

Fig. 13. Espectrograma de (a) segmento clasificado como Peces 1 y (b) segmento clasificado como No Peces 1 incorrectamente.

5.3. Peces 2

Para este tercer modelo se realizó la etapa de reconocimiento con 1,678 segmentos y un parámetro de exigencia igual a 9.9 dando como resultado un *F1-Score* igual a 0.8272 y un *Accuracy* de 0.7224 mostrando una buena detección de la clase Peces 2. Sin embargo, en la figura 14 se observa que 174 segmentos fueron mal clasificados como "NoPeces2". En dichos casos se aprecia presencia de ruido que enmascara las fonaciones de los peces lo que entorpece tanto la etapa de PDI como la etapa de reconocimiento como se observa en la figura 19 (b). Para estos casos es oportuno un procesamiento de la señal de audio de tal manera que se disminuya el ruido presente en las grabaciones.

Por otro lado, se observa un buen desempeño por parte del modelo al momento de reconocer audios que no contienen fonaciones de los peces objetivo. Se observa que los dos falsos positivos son principalmente por presencia de un sonotipo diferente a los buscados en este estudio el cual es mostrado en la figura 19 (a). Para este caso es conveniente tener

una mayor cantidad de segmentos que no contengan fonaciones de Peces 2 en la etapa de entrenamiento con el fin de aumentar la capacidad de generalización que tiene el modelo.

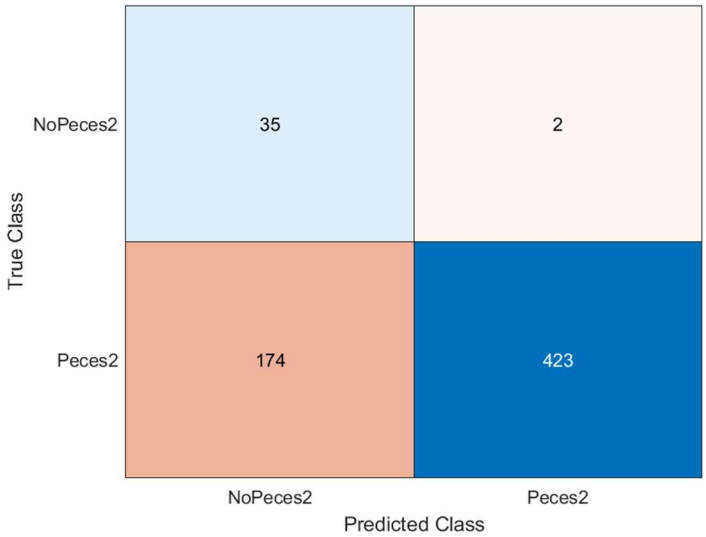

Fig. 14. Matriz de confusión de la predicción de la clase Peces 2 por el modelo propuesto

5.4. Análisis de sensibilidad

5.4.1. Peces 1

El efecto del parámetro de exigencia sobre el rendimiento del modelo es mostrado en la figura 15 en donde se observa que para valores bajos de este parámetro no se aprecia una diferencia considerable del *F1-Score* mientras que, para valores mayores a 7 se observa un aumento en estos valores. Esto se debe principalmente a la distribución del grado de pertenencia entregado por el algoritmo de lógica difusa implementado, LAMDA 3π. Dicha distribución se muestra en la figura 16 en donde se exhibe que la mayor cantidad de segmentos presentan un factor de pertenencia mayor a 0.9 por lo que la distribución de los datos no es lineal, razón por la cual, al realizar una interpolación lineal, a bajos valores del parámetro de exigencia el umbral de Membership toma valores bajos y la cantidad de segmentos que entran a la CNN no varía significativamente dando resultados muy similares.

A pesar de esto se muestra que, a partir de un parámetro de exigencia mayor a 7, hay un aumento en el *F1-Score* alcanzando un valor 0.7805 a medida que se aumenta dicho factor. Sin embargo, es evidente que, a medida que se aumenta la exigencia, la cantidad de segmentos que entran al clasificador es menor y se corre el riesgo de disminuir la detección

de fonaciones en segmentos puesto que los segmentos que no superen el umbral impuesto por el parámetro de exigencia no serán tenidos en cuenta en la clasificación, aun cuando puedan contener fonaciones de peces por lo que, este factor deberá ser establecido de acuerdo con los requerimientos del usuario y a la aplicación que se desea llevar a cabo.

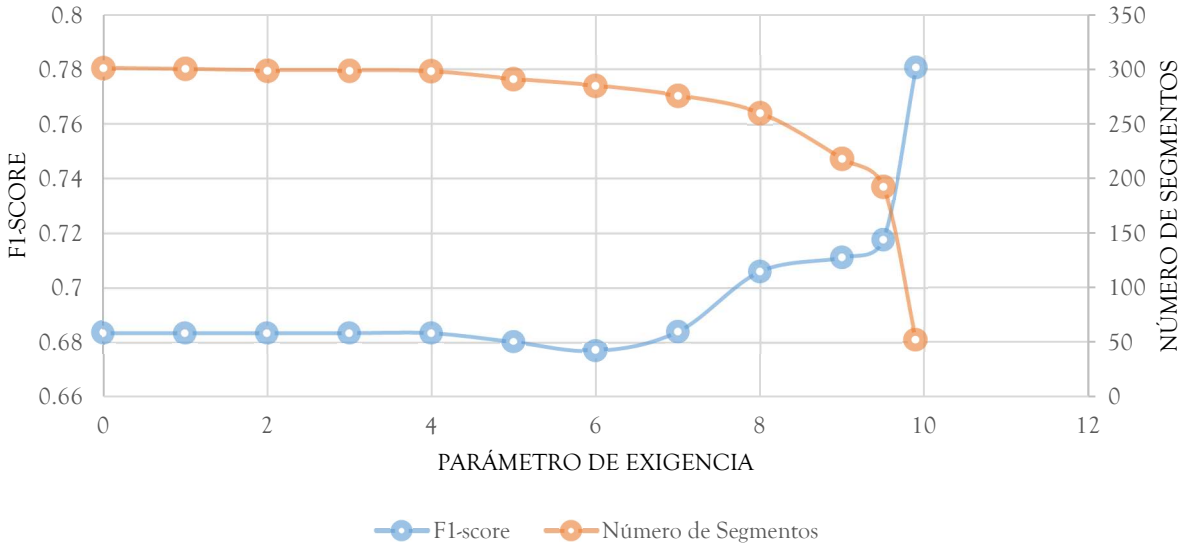

Fig. 15. Barrido del parámetro de exigencia y F1-Score resultante para la clase Peces 1

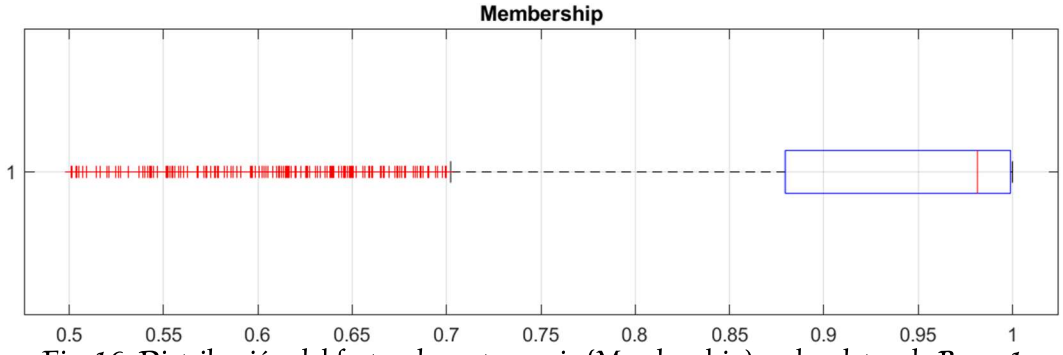

Fig. 16. Distribución del factor de pertenencia (Membership) en los datos de Peces 1

5.4.2. Peces 2

En este caso se presenta una situación similar a la mostrada en la sección anterior, aunque, para esta clase se muestra una mayor diferencia entre valores de *F1-Score* a medida que se varía el parámetro de exigencia como se puede ver en la figura 17, en donde se muestra una curva ascendente a partir de un parámetro de exigencia igual a 5. Sin embargo, al igual que en la clase de Peces 1, para Peces 2 hay un gran aumento en el *F1-score* cuando se tiene un parámetro de exigencia de 9.9 alcanzando un pico de 0.8272 debido a la distribución de los grados de pertenencia de los segmentos que son mostrados en la figura

18. Nuevamente el Membership presenta principalmente valores cercanos a la unidad, aunque, a diferencia del caso anterior, este grado de pertenencia es más variable lo que permite una mayor diferencia en el *F1-Score* al momento de variar el parámetro de exigencia.

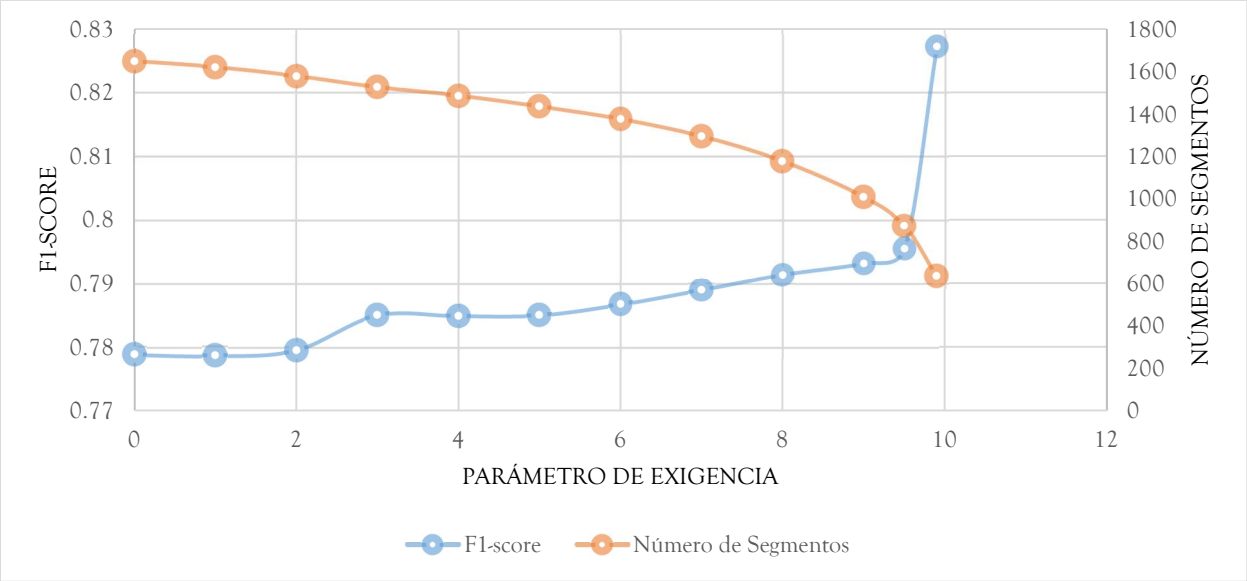

Fig. 17. Barrido del parámetro de exigencia y F1-Score resultante para la clase Peces 2

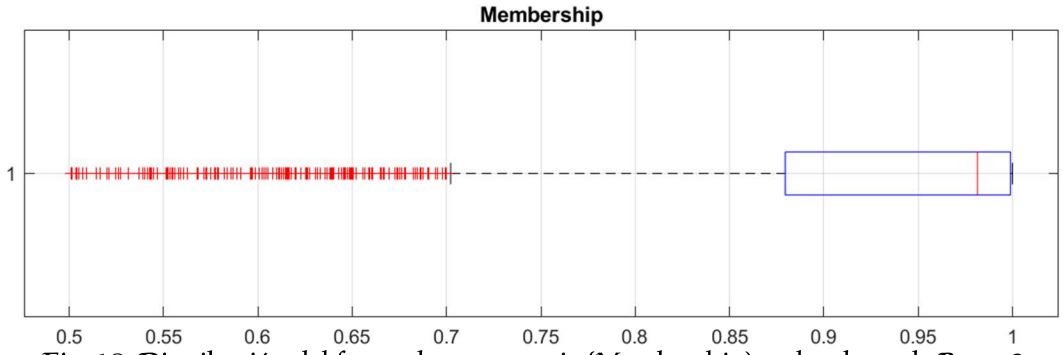

Fig. 18. Distribución del factor de pertenencia (Membership) en los datos de Peces 2

Al tener un parámetro de exigencia con valores bajos, se presenta un *F1-Score* menor debido a segmentos con presencia de ruido en donde, nuevamente, se ven enmascaradas las fonaciones con ruido presente en las grabaciones. Ante este ruido el Membership disminuye y es por este motivo por el que se da una mayor cantidad de falsos negativos al tener la exigencia con valores bajos resultando en métricas menores. Un ejemplo de este caso se muestra en la figura 19 (a), mientras que en la figura 19 (b) se muestra un ejemplo en donde un segmento que no contenía fonaciones de Peces 2 fue mal clasificado debido a la presencia de un sonotipo diferente a las clases objetivo de este estudio. Para mejorar los

resultados es importante encontrar mayores ejemplos de este nuevo sonotipo de tal manera que pueda ser diferenciado.

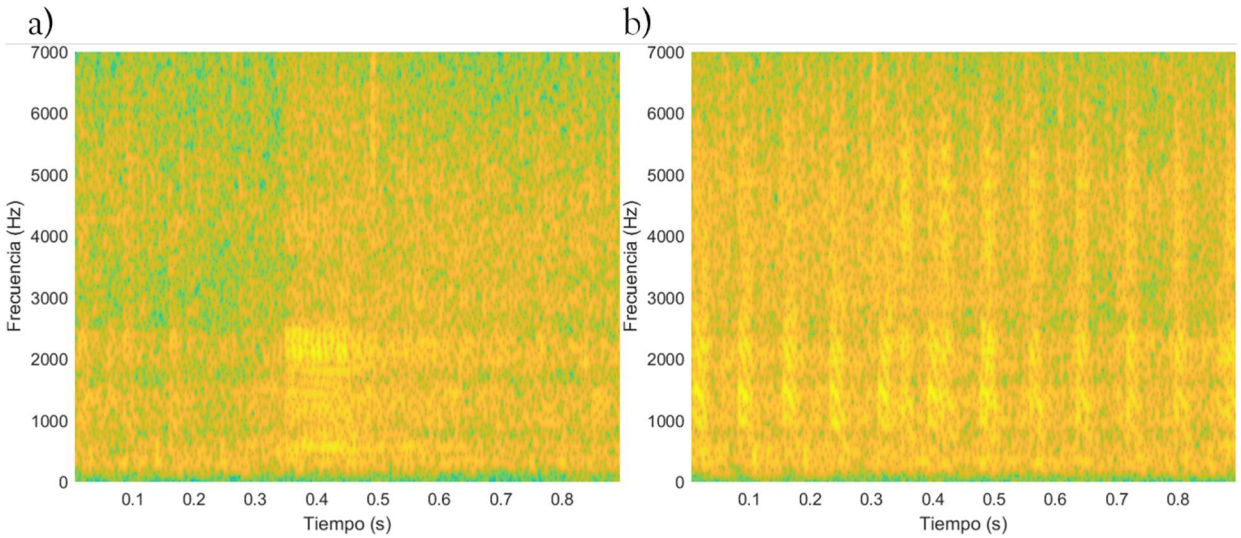

Fig. 19. Espectrograma de (a) segmento clasificado como Peces 2 y (b) segmento clasificado como No Peces 2 incorrectamente.

7. CONCLUSIONES

En conclusión, el algoritmo alcanza un *F1-Score* máximo de 0.8272 y un *Accuracy* de 0.7224 al momento de clasificar fonaciones de Peces 2, mientras que obtiene un *F1-Score* igual a 0.7805 y un *Accuracy* igual a 0.8269 para la clase Peces 1. Dichas métricas son similares a las presentados por Ibrahim [10], Harakawa [14] y Waddell [16] quienes tenían unas bases de datos más enriquecidas y con una mayor cantidad de ejemplos para el entrenamiento de los modelos. Por otro lado, esta metodología presenta un mejor *Accuracy* al presentado por Ozanich [13] debido a que el enfoque no supervisado aportado por el algoritmo LAMDA 3-pi fue complementado por un enfoque supervisado dado por la red ResNet18.

Con esto se muestra que el algoritmo propuesto no se encuentra lejos de otros trabajos realizados en donde se contaron con bases de datos más grandes permitiendo mayor entrenamiento de los modelos. Sin embargo, entrenar el modelo de este estudio con una mayor cantidad de grabaciones permitirá una mayor generalización y un mejor rendimiento. Lo mismo se concluye de la identificación de embarcaciones, en donde, a pesar de tener unas buenas métricas (*F1-Score* de 0.9942), se tenían pocas grabaciones en donde se pudiera evaluar correctamente el modelo de regresión logística.

Otra problemática que influye en el rendimiento es el ruido producido por el movimiento del agua, corrientes o actividad biológica inherente al ambiente en el que las grabaciones son tomadas, el cual puede llegar a enmascarar las fonaciones objetivo, principalmente por el hecho de que los sonidos son atenuados más rápidamente en el agua por lo que en muchos segmentos se presentan fonaciones de baja intensidad que no pueden ser correctamente detectadas gracias a la presencia del ruido ambiental. Es por este motivo que es importante realizar un preprocesamiento de la señal de audio de tal manera que este ruido sea atenuado permitiendo un mayor contraste con los sonotipos que se desean identificar.

Por último, el parámetro de exigencia permitió obtener mejores resultados en el rendimiento del algoritmo permitiéndole al usuario elegir la sensibilidad deseada de acuerdo con las necesidades de la aplicación. Sin embargo, se debe buscar que la

interpolación de este valor al momento de obtener el Membership umbral tenga una mejor representación de la distribución de los datos.

8. TRABAJO FUTURO

Una de las principales problemáticas presentes en el desarrollo de este estudio fue la falta de una mayor cantidad de ejemplos con los cuales la red neuronal pudiese ser entrenada por lo que, para futuros trabajos, sería recomendable evaluar la robustez del modelo con nuevos datos tomados en diferentes locaciones en donde se presenten las clases de peces objetivo de este trabajo y, en caso de ser necesario, reentrenar el modelo de tal manera que se puedan obtener mejor desempeño.

Por otro lado, el ruido producido por el ambiente en el que las grabaciones fueron tomadas, es decir, por corrientes marinas, movimiento de objetos cercano al hidrófono, o la interferencia inherente del agua disminuye el rendimiento del modelo tanto supervisado como no supervisado. Por esta razón es importante realizar mayores estudios en el procesamiento de las señales de audio tomadas con el fin de disminuir este ruido de fondo permitiendo un mayor contraste de las fonaciones y, por lo tanto, una mejor detección de estos sonotipos.

Debido a que en el agua la atenuación de los sonidos se produce más rápidamente, una nivelación de ganancias podría proponerse, de tal manera que las fonaciones de baja intensidad no sean eliminadas en la etapa de PDI para lograr una correcta clasificación y, por lo tanto, un mejor rendimiento de la metodología propuesta.

Por último, sería bastante interesante evaluar el comportamiento de la metodología propuesta al momento de detectar otras clases de fonaciones como pueden ser *Burst*, Clicks, otras especies de peces, silbidos, etc., de tal manera que esta metodología pueda escalar a una aplicación más general.

REFERENCIAS

[1] B. C. Pijanowski, A. Farina, S. H. Gage, S. L. Dumyahn, y B. L. Krause, "What is soundscape ecology? An introduction and overview of an emerging new science", *Landsc. Ecol.*, vol. 26, núm. 9, pp. 1213–1232, nov. 2011, doi: 10.1007/s10980-011-9600-8.

[2] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap", *PeerJ*, vol. 10, p. e13152, mar. 2022, doi: 10.7717/peerj.13152.

[3] M. J. Guerrero, C. L. Bedoya, J. D. López, J. M. Daza, y C. Isaza, "Acoustic animal identification using unsupervised learning", *Methods Ecol. Evol.*, vol. n/a, núm. n/a, 2023, doi: 10.1111/2041-210X.14103.

[4] J. Sueur y A. Farina, "Ecoacoustics: the Ecological Investigation and Interpretation of Environmental Sound", *Biosemiotics*, vol. 8, núm. 3, pp. 493–502, dic. 2015, doi: 10.1007/s12304-015-9248-x.

[5] M. Minello, L. Calado, y F. C. Xavier, "Ecoacoustic indices in marine ecosystems: a review on recent developments, challenges, and future directions", *ICES J. Mar. Sci.*, vol. 78, núm. 9, pp. 3066–3074, nov. 2021, doi: 10.1093/icesjms/fsab193.

[6] F. Ladich, "Fish bioacoustics", *Curr. Opin. Neurobiol.*, vol. 28, pp. 121–127, oct. 2014, doi: 10.1016/j.conb.2014.06.013.

[7] J. Wimmer, M. Towsey, P. Roe, y I. Williamson, "Sampling environmental acoustic recordings to determine bird species richness", *Ecol. Appl.*, vol. 23, núm. 6, pp. 1419–1428, sep. 2013, doi: 10.1890/12-2088.1.

[8] C. Bedoya, C. Isaza, J. M. Daza, y J. D. López, "Automatic recognition of anuran species based on syllable identification", *Ecol. Inform.*, vol. 24, pp. 200–209, nov. 2014, doi: 10.1016/j.ecoinf.2014.08.009.

[9] C. Bergler *et al.*, "ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning", *Sci. Rep.*, vol. 9, núm. 1, Art. núm. 1, jul. 2019, doi: 10.1038/s41598-019-47335-w.

[10] A. K. Ibrahim *et al.*, "An approach for automatic classification of grouper vocalizations with passive acoustic monitoring", *J. Acoust. Soc. Am.*, vol. 143, núm. 2, pp. 666–676, feb. 2018, doi: 10.1121/1.5022281.

[11] P. Scholar, "A wavelet based time-frequency descriptor for automatic classification of acoustic signals of fishes", 2019.

[12] J. Noda, C. Travieso, y D. Sánchez-Rodríguez, "Automatic Taxonomic Classification of Fish Based on Their Acoustic Signals", *Appl. Sci.*, vol. 6, núm. 12, p. 443, dic. 2016, doi: 10.3390/app6120443.

[13] E. Ozanich, A. Thode, P. Gerstoft, L. A. Freeman, y S. Freeman, "Unsupervised clustering of coral reef bioacoustics", dic. 2020.

[14] R. Harakawa, T. Ogawa, M. Haseyama, y T. Akamatsu, "Automatic detection of fish sounds based on multi-stage classification including logistic regression via adaptive feature weighting", *J. Acoust. Soc. Am.*, vol. 144, núm. 5, pp. 2709–2718, nov. 2018, doi: 10.1121/1.5067373.

[15] J.-F. Laplante, M. A. Akhloufi, y C. Gervaise, "Deep Learning for Marine Bioacoustics and Fish Classification Using Underwater Sounds", en *2022 IEEE*

Canadian Conference on Electrical and Computer Engineering (CCECE), Halifax, NS, Canada: IEEE, sep. 2022, pp. 288–293. doi: 10.1109/CCECE49351.2022.9918242.

[16] E. E. Waddell, J. H. Rasmussen, y A. Širović, "Applying Artificial Intelligence Methods to Detect and Classify Fish Calls from the Northern Gulf of Mexico", *J. Mar. Sci. Eng.*, vol. 9, núm. 10, p. 1128, oct. 2021, doi: 10.3390/jmse9101128.

[17] W. Shi y X. Fan, "Research on armored vehicle classification based on MFCC and SVM", en *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, dic. 2017, pp. 1938–1941. doi: 10.1109/CompComm.2017.8322876.

[18] J. Xie, M. Towsey, M. Zhu, J. Zhang, y P. Roe, "An intelligent system for estimating frog community calling activity and species richness", *Ecol. Indic.*, vol. 82, pp. 13–22, nov. 2017, doi: 10.1016/j.ecolind.2017.06.015.

[19] M. Wang, S. Zheng, X. Li, y X. Qin, "A new image denoising method based on Gaussian filter", en *2014 International Conference on Information Science, Electronics and Electrical Engineering*, Sapporo, Japan: IEEE, abr. 2014, pp. 163–167. doi: 10.1109/InfoSEEE.2014.6948089.

[20] Jamileh Yousefi, "Image Binarization using Otsu Thresholding Algorithm", 2015, doi: 10.13140/RG.2.1.4758.9284.

[21] J. Redmon, S. Divvala, R. Girshick, y A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection". arXiv, el 9 de mayo de 2016. doi: 10.48550/arXiv.1506.02640.

[22] B. Lamrini, M.-V. Le Lann, A. Benhammou, y E. K. Lakhal, "Detection of functional states by the 'LAMDA' classification technique: application to a coagulation process in drinking water treatment", *Comptes Rendus Phys.*, vol. 6, núm. 10, pp. 1161–1168, dic. 2005, doi: 10.1016/j.crhy.2005.11.017.