# Détection de clics d'Odontocètes basé sur une approche de réseaux de neurones profonds

Lucie Gabagnou, Armand L'Huillier, Yanis Rehoune
*Université Paris 1 Panthéon-Sorbonne*

**Abstract**

La recherche en bioacoustique a considérablement progressé ces dernières années, notamment grâce à l'amélioration des technologies d'enregistrement des sons et à l'avénement des réseaux de neurones. Au-delà l'amélioration de notre compréhension de la communication inter-espèces, les techniques bioacoustiques offrent également la capacité de surveiller la santé de la biodiversité, contribuant ainsi à la préservation des écosystèmes. Dans notre analyse, nous nous concentrons sur les clics d'écholocalisations des Odontocètes (cétacés à dents) tels que les dauphins et les baleines. Nous démontrons comment les algorithmes basés sur les réseaux de neurones peuvent efficacement détecter la présence de ces sons en utilisant les spectogrammes comme des images et en les passant dans des modèles adaptés comme les réseaux de neurones convolutifs (CNNs). Ce projet est réalisé dans le cadre d'un Challenge Data proposé par l'Ecole Normale Supérieure.

# 1 Introduction

## 1.1 Problématique

L'objectif de cette étude est de détecter les clics d'écholocalisation produits par les Odontocètes dans des enregistrements audio. L'écholocalisation est une technique employée par ces animaux pour éviter les obstacles et localiser leurs proies. Elle repose sur l'émission de clics à haute fréquence qui se réfléchissent sur les surfaces environnantes, puis sur l'analyse de l'écho produit par ces clics. Le principal défi de cette étude réside dans la reconnaissance de ces sons dans les enregistrements d'hydrophone réalisés via la méthode du "Passive Acoustic Monitoring (PAM)". Ces enregistrements peuvent être perturbés par des bruits transitoires (tels que les bruits produits par des crevettes ou des récifs), ce qui peut altérer la qualité de l'enregistrement.

## 1.2 Description du jeu de données

La base de données utilisée pour cette étude provient du projet européen CARI'MAM (Caribbean Marine Mammals Preservation Network), dont l'objectif est d'améliorer la conservation des espèces marines dans les Caraïbes, y compris les Odontocètes, qui sont très présents dans cette région du monde. Les enregistrements audio ont été effectués entre 2017 et 2021 sur huit sites des Antilles. La base d'entraînement se compose d'environ 23 000 fichiers, dont 41% contiennent un clic d'Odontocète. La base de test, utilisée pour soumettre les résultats sur la plateforme, comprend quant à elle moins de 1 000 observations. Chaque enregistrement audio a une durée de 200 millisecondes (ms).

# 2 Traitement des données audios

Les données audio requièrent l'extraction de caractéristiques avant d'être utilisées dans un modèle d'apprentissage profond. Cette section aborde la préparation des données audio, en se concentrant sur l'utilisation des spectrogrammes, tout en écartant d'autres méthodes basées sur les fréquences telles que les spectres centroids et les approches de transformation (comme la Transformation de Fourier ou la transformation en ondelettes). Après avoir filtré avec une bande passante, le signal est transformé en spectrogramme. Après cela, différentes méthodes de traitement d'images sont appliqués avant la modélisation.

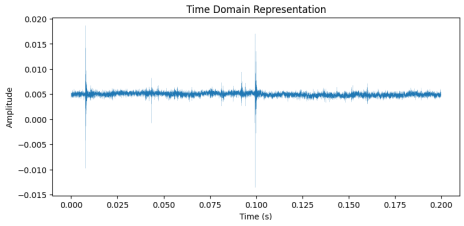


Figure 1: Aperçu du signal sonore brut

## 2.1 Filtrage bande-passante

Dans un premier temps, un filtre bande-passante est appliqué entre 5 kHz et 100 kHz pour sélectionner les fréquences pertinentes. Effectivement, différentes études, tel que l'étude *Analyse des risques pour les mammifères marins liés à l'emploi des méthodes acoustiques en océanographie* [1] de 2007, montre que les sons émis par les Odontocètes sont compris dans cet intervalle.

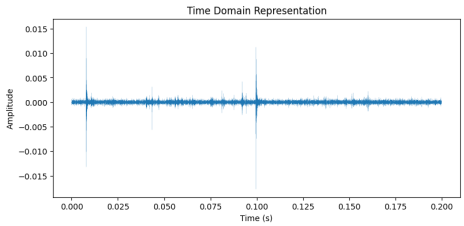


Figure 2: Aperçu du signal filtré entre 5kHZ et 100kHZ

## 2.2 Spectrogramme

Le spectrogramme est une représentation en temps-fréquence d'un signal basée sur la théorie du traitement du signal. Pour construire le spectrogramme, la transformée de Fourier à court terme (STFT) est utilisée:

$$STFT(x[n])(m,k) = \sum_{n=0}^{N-1} x[n+mR]w[n]e^{-j\frac{2\pi}{N}nk} \tag{1}$$

où $x[n]$ est le signal d'entrée, $w[n]$ est la fenêtre temporelle (dans ce cas, une fenêtre de Hamming de taille $N = 2048$), $R$ est le pas de décalage (taille de chevauchement de 128), et $m$ et $k$ sont les indices temporels et fréquentiels discrets, respectivement. La fenêtre Hamming est définie comme suit :

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

Le type de fenêtre, la taille de la fenêtre ainsi que le pas de chevauchement inter-fenêtre a été décidé en visionnant l'ensemble des spectrogrammes à retrouver en annexe. Grille de paramètres du STFT et spectrogrammes associés pour un signal sonore

Finalement, une réduction de bruit par filtrage adaptatif avec la décomposition de Wiener a été appliquée pour tenter d'enlever du bruit sur l'audio et faire apparaître davantage les clics. Ce signal est conservé en logarithme, car l'échelle permet de capter davantage d'informations. Les mel-spectrogrammes ont également été testés, mais n'ont pas permis d'améliorer les performances.

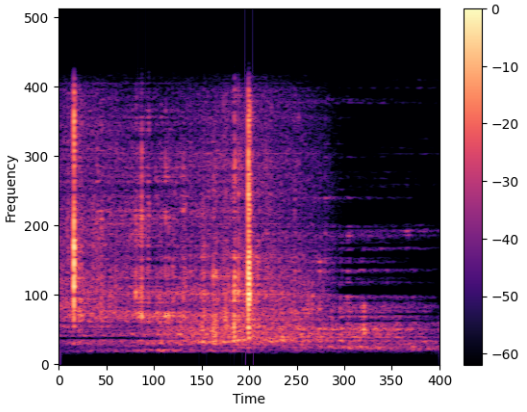


Figure 3: spectrogramme d'un audio positif (comportant au moins un clic)

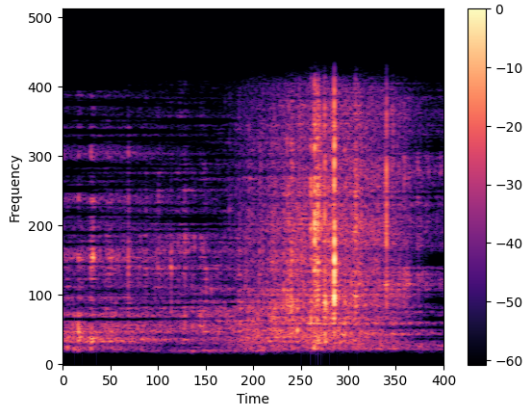


Figure 4: spectrogramme d'un audio négatif

Les spectrogrammes des clics semblent être caractérisés par des faisceaux plus distincts (en jaune clair) et sont espacés dans le temps cintrairemermnt aux bruits.

## 2.3 Traitement d'image

Les spectrogrammes, considérés comme des images, nécessitent un prétraitement afin d'obtenir des valeurs numériques exploitables pour les modèles de réseaux de neurones. Ainsi, les images sont d'abord normalisées dans l'intervalle [0, 1] et ensuite redimensionnées sur une échelle de 255 pour permettre un traitement efficace des images en couleur. Pour correspondre aux valeurs entrantes des modèles de réseaux de neurones, les images sont également converties en une représentation à 3 canaux (RGB).

# 3 Modèles

## 3.1 Objectif

Suite au traitement des enregistrements audio permettant d'obtenir des spectrogrammes sous forme d'images, plusieurs modèles de classification binaire ont été développés en utilisant diverses architectures de réseaux de neurones adaptées au traitement d'image.

## 3.2 Approche

En se basant sur *Computational bioacoustics with deep learning: a review and roadmap*[2], l'état de l'art en matière de modèles de Deep Learning adaptés à la classification bioacoustique a été exploré. Plusieurs architectures et techniques ont été utilisées pour développer des modèles en commençant par un modèle de base et en augmentant progressivement la complexité. Finalement l'ensemble des méthodes reposent sur les CNN:

- Un modèle de base utilisant une architecture simple de CNN a été développé pour établir une référence (baseline) en termes de performance (OdontoceteCNN)
- Une implémentation des architectures de CNN connues VGG et Res a été adaptée aux données de la base pour évaluer les performances sur notre problème spécifique.
- Un modèle de Transfer Learning* a été appliqué à partir des architectures de CNN MobileNetV2 et ResNet50, permettant d'exploiter les connaissances préalables apprises par ces modèles pour améliorer les performances sur notre tâche de classification.

Chacune de ces architectures est complétée par une couche de sortie avec une fonction d'activation sigmoïde. Grâce à sa nature, cette fonction produit des probabilités (comprises entre 0 et 1) qui représentent l'appartenance à chaque classe. La sigmoïde est donc particulièrement adaptée pour les tâches de classification.

### 3.2.1 OdontoceteCNN: baseline

OdontoceteCNN est un modèle de baseline reposant sur une architecture de CNN relativement simple. Il se compose de quatre blocs de

*Le Transfer Learning consiste à utiliser un modèle pré-entraîné sur un ensemble de données différent (généralement plus large) et à le réutiliser pour résoudre un problème similaire.

convolution, chacun comprenant des couches de convolution 2D pour l'extraction des caractéristiques, de normalisation par batch pour améliorer la stabilité, d'activation ReLU pour introduire la non-linéarité et de MaxPooling2D pour réduire la dimensionnalité. Finalement, une couche de Global Average Pooling et une couche dense (entièrement connectée) sont employées.

#### 3.2.2 VGG

VGG est une architecture de CNN développée par l'équipe Visual Geometry Group de l'Université d'Oxford. La principale caractéristique de VGG est l'utilisation de couches convolutives avec de petits filtres (3x3) et un stride (pas de déplacement du filtre) de 1, empilées les unes sur les autres pour capturer des caractéristiques. L'architecture comprend également des couches de MaxPooling pour réduire la dimensionnalité et des couches entièrement connectées pour la classification finale. La version VGG16 a été implémentée dans l'étude.

#### 3.2.3 MobileNet

MobileNet (V2) est une architecture de CNN développée par Google en 2018 dans le but d'être utilisé des appareils mobiles avec des ressources limitées en calcul. Elle utilise une technique appelée "bottleneck residual block" pour réduire la dimensionnalité du réseau, tout en conservant des performances élevées.

#### 3.2.4 Techniques d'amélioration

Data Augmentation et Transfer Learning ont été utilisés pour améliorer les performances.

**Transfer Learning** Bien que les modèles pré-entraînés soient principalement entraînés sur des images naturelles, des tentatives ont été faites pour fine-tuner des modèles tels que VGG et MobileNet, qui offrent un bon compromis entre performances et complexité (nombre de paramètres).

**Data Augmentation** Il s'agit d'une technique générant des données supplémentaires. En utilisant les images de la base d'origine, la Data Augmentation consiste en réaliser différentes transformations aléatoires (recadrage, orientation, bruit supplémentaire) dans le but d'enrichir la variabilité des données (et d'éviter le sur-apprentissage). Cette technique sera utilisée en supplément du Transfer Learning sur MobileNetV2: une augmentation du nombre de données induit une augmentation du temps de calcul mais le modèle MobileNetV2 est optimisé en termes de nombre de paramètres.

### 3.3 Entraînement

#### 3.3.1 Métriques

Les métriques de précision (accuracy) et d'AUC (Area Under the Curve) ont été utilisées pour guider les ajustements du modèle. La précision est calculée comme le ratio entre le nombre de prédictions correctes et le nombre total de prédictions, tandis que l'AUC est une mesure de la capacité du modèle à distinguer entre les classes.

La précision est définie comme suit:

$$Accuracy = \frac{Nombredeprédictionscorrectes}{Nombretotaldeprédictions} \tag{3}$$

L'AUC est calculée en traçant la courbe ROC (Receiver Operating Characteristic) du modèle, qui représente le taux de vrais positifs par rapport au taux de faux positifs à différents seuils de classification. L'aire sous cette courbe est ensuite calculée pour obtenir l'AUC.

$$AUC = \int_0^1 ROC(x)dx \tag{4}$$

L'AUC évalue la capacité du modèle à distinguer les exemples positifs et négatifs en fonction des seuils de probabilité séparant les classes "positives" et "négatives". Il s'agit de la métrique utilisée pour évaluer les performances du modèle pour le challenge.

### 3.3.2 Fonction de perte

La fonction de perte utilisée dans le modèle est la fonction de perte entropique croisée binaire (binary cross-entropy loss), qui mesure la différence entre les prédictions du modèle et les valeurs réelles de la variable cible. La fonction de perte est définie comme suit:

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}[y_i log(\hat{y_i}) + (1 - y_i) log(1 - \hat{y_i})] \quad (5)$$

Où $y_i$ est la valeur réelle de la variable cible pour l'observation $i$, $\hat{y_i}$ est la prédiction du modèle pour l'observation $i$, et $n$ est le nombre total d'observations.

### 3.3.3 Choix des Hyper-paramètres

L'échantillon d'apprentissage a été divisé en ensembles d'entraînement et de validation pour régler les hyperparamètres du modèle et veiller à ne pas sur-apprendre sur les données de l'échantillon. Les hyperparamètres ajustés comprenaient le nombre d'époques, le taux d'apprentissage et l'optimiseur.

Chaque modèle a été soumis à une procédure d'exploration, initialement fixée à 20 époques et en utilisant l'optimiseur Adam. L'exploration des hyperparamètres a été effectuée en utilisant une recherche aléatoire (random search) sur une grille prédéfinie de valeurs possibles pour chaque hyperparamètre. Les résultats de chaque modèle ont été évalués en utilisant les métriques de précision et d'AUC sur l'ensemble de validation.

## 3.4 Performances

Les performances des différents modèles ont été évaluées avec l'AUC sur la plateforme du Challenge. Les performances des modèles principaux se trouvent dans le tableau ci-dessous:

| Modèle | AUC |
|---|---|
| OdontoceteCNN | 0.89 |
| VGG16 | 0.84 |
| Transfer Learning VGG16 (sur ImageNet) | 0.87 |
| Transfer Learning MobileNetV2 (sur ImageNet) + DataAugmentation | 0.56 |

Table 1: AUC values for different models.

# 4 Conclusion

A partir des enregistrements récoltés par PAM (Passive Acoustic Monitoring), différents traitements ont pu être appliqués permettant de récuperer de l'information provenantn de sons plutôt bruités. L'utilisation de spectrogramme et les traitements associées (filtrage, choix des caractéristiques pour STFT) et de diverses architectures de CNN ont permis d'obtenir des performances satisfaisantes en termes d'AUC pour la tâche de classification proposée. Parmi les architectures étudiées, l'approche de base avec le modèle OdontoceteCNN a montré les meilleures performances avec un AUC de 0.89.

Il est important de noter que cette étude a principalement exploré les approches basées sur les CNN, mais d'autres techniques pourraient également être envisagées pour améliorer davantage les performances. Par exemple, l'exploration d'architectures de réseaux de neurones récurrents (RNN) ou d'approches hybrides (CRNN) pourrait permettre de mieux prendre en compte les aspects temporels des enregistrements audio.

# References

[1] Antoine Loic Lurton Xavier. Analyse des risques pour les mammifères marins liés Ã l'emploi des méthodes acoustiques en océanographie. Report, 2007.

[2] Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap, 2021.

# 5 Annexes

## 5.1 Annexe 1

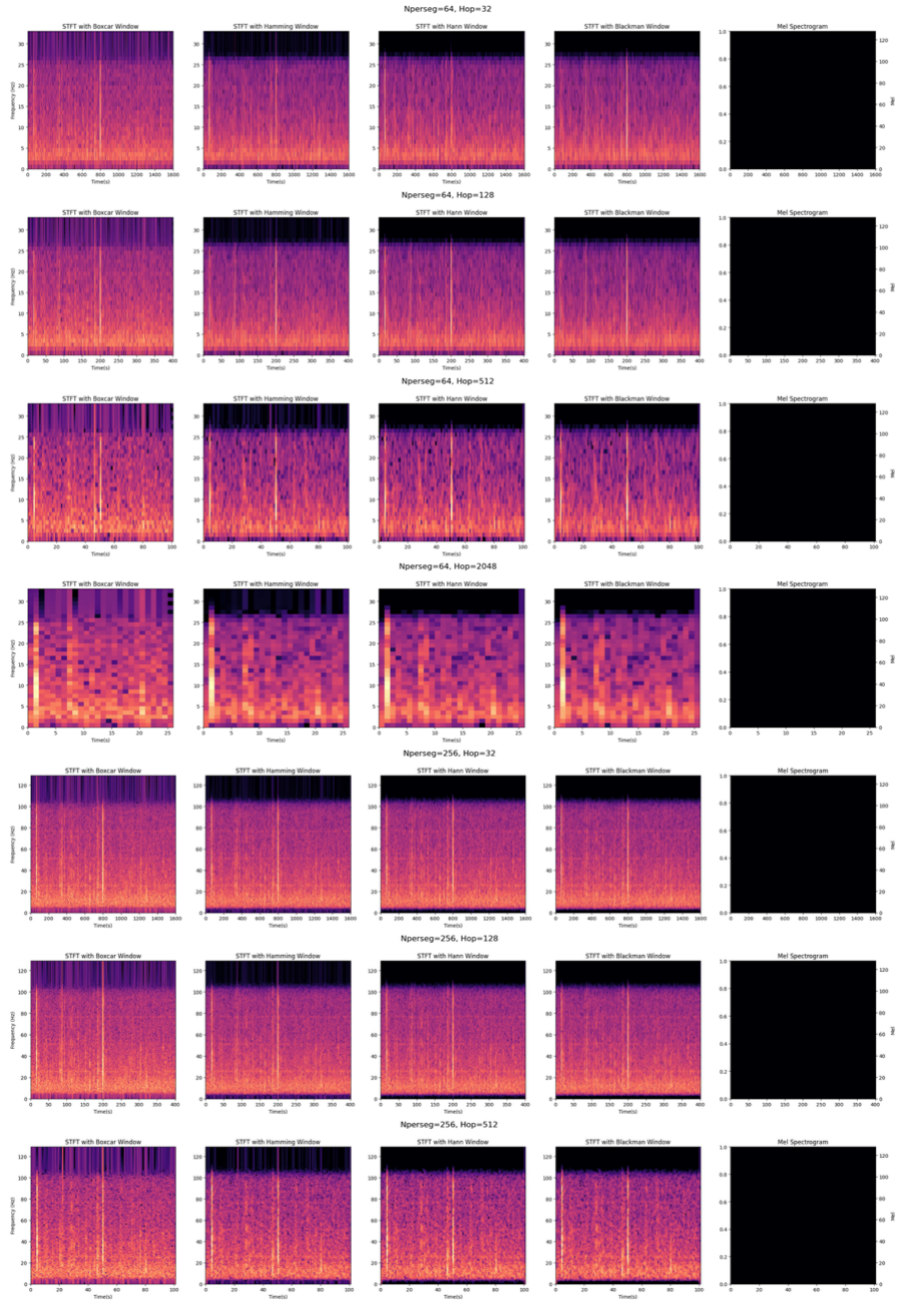


Figure 5: Grille de paramètres du STFT et spectrogrammes associés pour un signal sonore