Apprentissage par transfert pour la détection et la classification automatiques de grandes baleines dans l'océan austral

Lucie JEAN-LABADYE[1, 2, 3] Gabriel DUBUS[1, 2] Dorian CAZAU[2] Nicolas FARRUGIA[3] Axel MARMORET[3] Olivier ADAM[1]

[1]Institut ∂'Alembert, Sorbonne Université, UMR CNRS 7190, F-75005, Paris

[2]ENSTA Bretagne, Lab STICC, UMR CNRS 6285, F-29200, Brest

[3]IMT Atlantique, Lab STICC, UMR CNRS 6285, F-29238, Brest

Résumé – Les tâches de détection et de classification des vocalisations de cétacés capturées par acoustique passive sont complexes Les méthodes d'apprentissage profond à l'état de l'art souffrent d'une généralisation limitée. L'apprentissage par transfert, exploitant des modèles pré-entraînés sur de plus larges jeux de données, offre une solution. Dans ce cadre, nous avons évalué l'encodeur Perch, entraîné sur Xeno-Canto, à l'aide de métriques assurant une comparaison équitable.

Abstract – Automatic detection and classification of cetacean vocalizations in passive acoustic recordings are complex tasks While convolutional neural networks are widely used, their generalization is often constrained by scarce annotated data and high recording variability (geographic, temporal, equipment-related, etc.). Transfer learning, leveraging larger pretrained networks, offers a potential solution. In this context, we investigated the Perch encoder and evaluated it using metrics ensuring a fair comparison.

1 Introduction

La bioacoustique consiste en l'étude des sons émis par le vivant. Elle est particulièrement pertinente pour l'étude des cétacés [15], difficiles à observer visuellement, et qui utilisent le son comme canal privilégié pour toutes leurs activités vitales (chasse, socialisation, apprentissage, etc.) [3]. L'eau propage par ailleurs le son jusqu'à 4.5 fois plus vite que l'air. Les grandes baleines émettant généralement des vocalisations entre 20 et 125 Hz, la faible atténuation des basses fréquences offerte par le milieu sous-marin, ainsi que les pertes quasi nulles lors de la propagation de l'onde acoustique permettent aux hydrophones d'enregistrer des sons dont la source se trouve parfois à plusieurs centaines de kilomètres.

Les hydrophones, qui sont des microphones étanches, sont les premiers maillons de l'acoustique passive, qui consiste à enregistrer des sons sans en émettre. Cette méthode d'observation à distance est particulièrement intéressante car elle ne requiert pas d'intervention humaine constante : elle est donc non invasive, indépendante des conditions météo, et peut s'inscrire sur le long-terme. Dans ce contexte, l'acoustique passive permet d'avoir accès à de précieuses informations pour le suivi des cétacés et peut être un outil privilégié pour accompagner les décideur·euse·s quant à des politiques de conservation.

Toutefois, le traitement de ces données nécessite l'intervention d'expert·e·s capable·s d'identifier la grande variété de sources sonores, processus extrêmement chronophage et exigeant. La question de l'automatisation des processus de détection et de classification s'est donc rapidement posée, à laquelle l'apprentissage profond constitue aujourd'hui la réponse la plus répandue. S'inspirant des techniques de vision par ordinateur, la plupart des travaux actuels se basent sur des réseaux de neurones convolutifs (CNN) prenant en entrée des spectrogrammes [13]. Si ces méthodes permettent d'obtenir des performances satisfaisantes sur des données proches de celles ayant servi à l'entraînement des modèles (même période de temps, même région géographique ou mêmes espèces d'intérêt par exemple), la question de la généralisation reste encore très largement ouverte, notamment dans un contexte où les annotations —plus que les données— sont rares, coûteuses et de qualité variable.

Pour y répondre, une solution peut être l'apprentissage par transfert, qui consiste en l'utilisation d'un autre réseau pré-entraîné comme pré-encodeur pour le réseau réalisant la tâche d'intérêt . En effet, le modèle réalisant la tâche finale de détection et/ou de classification tire ainsi profit des représentations latentes d'un réseau généralement plus important et entraîné sur un volume de données annotées plus conséquent. Dans cette optique, notre contribution s'est intéressée au modèle Perch, publié en 2023 par Ghani et al. [2] dans un article mettant en lumière ses capacités à extraire des *embeddings* (i.e. des représentations latentes) pertinents pour des tâches de classification en bioacoustique. Notre but est d'étudier le comportement du modèle sur des données très différentes de ses données d'entraînement, pour évaluer sa capacité de généralisation et sa pertinence pour de l'apprentissage par transfert.

Après avoir mis l'étude dans le contexte de son état de l'art dans la Section 2, nous décrirons les modèles employés ainsi que leurs données d'entraînement et les métriques d'évaluation utilisées dans la Section 3, avant d'exposer les résultats obtenus dans la Section 4.

2 Etat de l'art

Une des problématiques actuelles de la bioacoustique est le traitement des gros volumes de données enregistrées. Avec l'accessibilité grandissante des hydrophones et l'augmentation des capacités de stockage et de l'autonomie, l'acquisition des sons pose désormais nettement moins problème que leur annotation. Les campagnes d'acquisition étant menées sur des

périodes de temps pouvant aller jusqu'à plusieurs années, le volume à traiter est gigantesque : un hydrophone enregistrant en continu pendant 24 heures en mono sur 24 bits à 44 kHz génère par exemple 11 GB de données par jour. Le processus d'annotation est donc extrêmement long et coûteux, en plus de nécessiter généralement l'intervention d'expert·e·s capables d'identifier les sources sonores. Son automatisation est donc devenue une question centrale.

Cette automatisation s'est d'abord faite en exploitant les théories du traitement du signal avancé, avant l'émergence du *machine learning* à la fin des années 2000. Une dizaine d'années plus tard, le *deep learning* est venu créer un nouvel état de l'art [13], avec des réseaux de neurones profonds (DNN) dépassant très largement les résultats des détecteurs précédents [12]. Agissant en approximateurs universels de fonction, les réseaux de neurones font en effet preuve d'une grande versatilité, permettant d'être utilisés pour différentes tâches comme la détection et la classification automatiques de sources sonores, le débruitage, etc. A l'échelle plus fine de la détection et de la classification automatiques de vocalisations de cétacés, ces réseaux ont été utilisés aussi bien pour détecter des sifflements et des clics de dauphins, qui se situent sur des bandes fréquentielles allant du kHz à l'ultrason que pour classifier des vocalisations de grandes baleines [1], situées entre la centaine de Hz et l'infrason, et ce pour des enregistrements provenant de régions géographiques, périodes temporelles et/ou configurations d'enregistrement différentes. Grâce à l'utilisation de spectrogrammes, les DNN ont pu bénéficier des avancées de la vision par ordinateur, plus ancienne et plus étudiée : l'usage de CNN prenant comme entrées des spectrogrammes s'est ainsi rapidement imposé comme la nouvelle norme [13], et constitue la *baseline* des rares benchmarks publiés [11].

Toutefois, et dû aux nombreuses problématiques des données acoustiques sous-marines et de leur annotation (rapports signal sur bruit (RSB) extrêmement bas, superposition de signaux d'intérêt, masquage acoustique, déséquilibre des jeux de données, variabilité inter-annotateur·ice·s etc.[1]), les contributions les plus récentes s'intéressent à de nouvelles méthodes moins gourmandes en données annotées[10],[7] ou à des techniques nécessitant peu de données comme l'apprentissage par transfert.

2.1 Apprentissage par transfert avec Perch

La première étape de l'apprentissage par transfert consiste à choisir un modèle pré-entraîné qui soit pertinent pour les données à disposition. En 2023, Ghani et al. rendent public Perch [2], un modèle profond basé sur une architecture d'EfficientNet B1 [14]. Présenté comme un encodeur pour les tâches de bioacoustique, c'est-à-dire un modèle capable de générer des représentations latentes des sons issus de la biophonie qui soient pertinentes et riches en informations, il est comparé à d'autres modèles convolutifs ou attentionnels sur différentes tâches, et se révèle être le plus robuste et le plus versatile. Cette comparaison est faite au moyen d'une régression logistique appliquée aux *embeddings* générés par l'encodeur (*i.e.* aux vecteurs contenant ses représentations latentes). Volontairement simples, ces expériences démontrent que les bonnes performances du modèle sont plus dûes au fait que l'encodeur a effectivement appris à capturer les informations pertinentes pour la classification des signaux d'entrée qu'au traitement appliqué aux *embeddings* par la suite.

Perch a été entraîné sur des sons émis par des oiseaux, ce qui a motivé notre intérêt pour deux raisons. Premièrement, les données d'oiseaux sont plus largement étudiées par les bioacousticien·ne·s car plus nombreuses et souvent de bonne qualité. Ceci s'explique par la relative simplicité à capturer des sons d'oiseaux, ainsi que par la possibilité de le faire en conditions contrôlées. Ainsi, Xeno-Canto (XC), qui a servi de base d'entraînement pour Perch, regroupe plus de 16 000 heures d'enregistrement, distribuées entre plus de 10 000 taxons[8]. Par ailleurs, les sons d'oiseaux présentent des similarités avec les données acoustiques de mammifères marins : ce sont tous les deux des sons issus de la biophonie, avec une structure temporelle de chants et des répétitions de motifs dans le temps [5]. Ces similarités peuvent peut-être expliquer les bonnes performances de Perch sur la Watkins Marine Mammal Sound Database (WMMSD) [9], base de données regroupant plus de 15 000 enregistrements provenant de plus de 60 espèces de mammifères marins.

Toutefois, les vocalisations d'oiseaux et de grandes baleines présentent aussi de grandes différences, notamment en termes de plage de fréquence, où le décalage est important. La plupart des vocalisations d'oiseaux se situent en effet entre 1 et 10 kHz, quand les vocalisations basse fréquence des grandes baleines Antarctique se situent entre 15 et 125 Hz. [6]. De plus, les données de XC sont des clips audios, qui sont tout le contraire de données d'acoustique passive sous-marine, qui peuvent impliquer de longues distances entre l'hydrophone et la source, ainsi que de longues périodes sans signaux d'intérêt. La pertinence des *embeddings* générés par Perch pour nos données n'est donc pas garantie.

3 Méthodologie

3.1 Modèle

Afin d'évaluer la pertinence de Perch sur des données d'acoustique passive centrées sur les grandes baleines de l'océan austral, nous avons utilisé sa partie encodeur pour générer des représentations *a priori* pertinentes du signal. Pour rester proches des expériences de Ghani et al. [2], ces *embeddings* ont été passés dans un Perceptron à deux couches (MLP) entraîné avec une fonction de coût d'entropie croisée binaire, le cadre étant celui de la détection présence/absence d'une vocalisation. Ce dernier a été entraîné pendant quinze *epochs* en validation croisée, avec une partie du jeu de données dédiée à l'évaluation, l'autre étant séparée en données d'entraînement (80 %) et de validation (20 %). Le code utilisé est en accès libre[1].

3.2 Données

Les données utilisées ont été regroupées et annotées par Miller et al. [6]. Disponibles en accès libre, elles ont été enregistrées à sept localisations différentes autour de l'Antarctique et au cours de différentes années, pour un total de onze paires site-année. Regroupant 1880 heures d'enregistrements (soit huit fois moins que XC), ce jeu de données est particulièrement intéressant pour évaluer les capacités de généralisation des

[1] https://github.com/LucieJL/shelf

modèles, dans la mesure où il rassemble des données issues de lieux, d'années et d'hydrophones différents.

L'effort d'annotation, mené par cinq annotateur·ices différent·es, s'est concentré sur les vocalisations de baleines bleues (*Balaenoptera musculus*, Bm) et de rorquals communs (*Balaenoptera physalus*, Bp). Les annotations sont réparties en trois classes : les vocalisations ABZ (BmA, BmB, BmZ), spécifiques des baleines bleues, les sons pulsés (Bp20, Bp20plus) spécifiques des rorquals communs et les *downsweeps* (BmD, BpD), émis par les deux espèces. Elles ont été faites sous la forme de boîtes en temps/fréquence sur les spectrogrammes des sons enregistrés. A l'inverse de la base de données WMMSD, constituée d'extraits audio pré-segmentés autour des vocalisations et généralement de bonne qualité, les données compilées par Miller et al [6] sont issues de l'acoustique passive, et présentent par nature un grand déséquilibre entre bruit de fond et signaux d'intérêt (9.6 % du jeu de données est considéré positif pour les trois classes décrites ci-dessus), ainsi que des signaux pouvant se superposer, voire se masquer.

3.3 Pré-traitement

A l'instar de la plupart des modèles pré-entraînés, Perch prend en entrée des vecteurs de taille fixe. Entraîné sur des segments de 5 secondes échantillonnés à 32 kHz, le vecteur d'entrée doit donc contenir $5 \times 32.10^3 = 160.10^3$ points.

Toutefois, dans la mesure où les vocalisations d'intérêt de cette étude se situent principalement entre 15 et 125 Hz, et où certaines vocalisations ont une durée moyenne de douze secondes, nous avons choisi de rééchantillonner. Trois configurations ont été testées : une première avec une fréquence d'échantillonnage de 250 Hz, fréquence minimale pour nos signaux et fréquence minimale de certains des enregistrements réalisés par Miller et al. [6], impliquant une durée de $160.10^3/250 = 640$ secondes ; une deuxième en créant des segments de 299 secondes, durée minimale d'un enregistrement non-découpé, et en rééchantillonnant donc à $160.10^3/299 = 535$ Hz ; et une dernière en créant des segments de 15 secondes rééchantillonnés à $1,07.10^4$ Hz. Cette dernière configuration a été motivée par la publication du *benchmark* de Schall et. al [11] utilisant des extraits de cette durée, pour une comparaison équitable.

Les segments ainsi obtenus ont été ré-annotés automatiquement par des labels binaires de présence/absence pour chaque type de vocalisation, un segment étant considéré positif s'il contient au moins 50% de la vocalisation. La validation croisée a été mise en place en prenant une paire "site-année" pour l'évaluation à chaque bloc. Pour réduire le déséquilibre inhérent des données sans s'éloigner de scénarios réalistes, les ensembles d'entraînement et de validation contiennent autant de segments avec des signaux d'intérêt que de segments sans, tandis que la paire "site-année" servant à l'évaluation est conservée telle quelle.

3.4 Métriques

Les jeux de données issus de l'acoustique passive étant par nature hautement déséquilibrés, les modèles de détection et de classification appliqués à ces données ne peuvent être évalués par *accuracy* : dans notre cas, un détecteur indiquant systématiquement une absence de signaux d'intérêt obtiendrait une *accuracy* de 90.4%. En suivant les recommandations de [4] ainsi que les usages de la communauté, nous avons évalué les différentes configurations par des courbes de *precision-recall*, ainsi que par la moyenne harmonique de ces deux métriques : le F_1-score. Ce score nous a par ailleurs servi à fixer des seuils afin de déterminer les meilleures versions de nos modèles : différents seuils de binarisation ont été testés sur les sorties du modèle en validation, et le seuil obtenant le meilleur F_1 a été choisi. Après binarisation des sorties, nous avons évalué notre modèle avec les 4 métriques proposées par le *benchmark* auquel nous nous comparons [11] : le taux de classification correcte (TCR, pour *true classification rate*, le taux de mavaise classification du bruit (NMR, pour *noise missclassification rate*), le taux de mauvaise classification intra-classes (CMR, pour *class missclassification rate*) et une métrique d'adéquation (*fitness metric*), qui est une moyenne des trois autres. Elles sont définies par :

$$\text{TCR} = \frac{1}{3}\left(\frac{\text{TP}_{c_1}}{(\text{TP}_{c_1} + \text{FP}_{c_1})} + \frac{\text{TP}_{c_2}}{(\text{TP}_{c_2} + \text{FP}_{c_2})}, \frac{\text{TP}_{c_3}}{(\text{TP}_{c_3} + \text{FP}_{c_3})}\right)$$

$$\text{NMR} = \frac{\text{FP}_{\text{bruit (vérité } c_1)} + \text{FP}_{\text{bruit (vérité } c_2)} + \text{FP}_{\text{bruit (vérité } c_3)}}{(\text{TP}_{\text{bruit}} + \text{FP}_{\text{bruit}})}$$

$$\text{CMR} = \frac{1}{6}\sum_{i=1}^{(i \bmod 3)+1}\left(\frac{\text{TP}_{c_i(\text{vérité } c_{i+1})}}{\text{total}_{c_i}} + \frac{\text{TP}_{c_i(\text{vérité } c_{i+2})}}{\text{total}_{c_i}}\right)$$

$$\text{Fitness} = \frac{1}{4}\left[\text{TCR} + (1-\text{NMR}) + (1-\text{NMR}) + (1-\text{CMR})\right]$$

Un poids supplémentaire est accordé à la minimisation du NMR dans le calcul de la *fitness* car cette métrique reflète une attente clé de l'utilisateur·ice : détecter les signaux d'intérêt tout en évitant les fausses alertes.

4 Résultats et discussion

Comme le montre la Figure 1, c'est le modèle entraîné avec des segments de 15 secondes rééchantillonnés à $1,07.10^4$ Hz qui obtient les meilleures performances en termes de *fitness*. Ce résultat peut être expliqué par le découpage du jeu de données : des segments de 15 secondes ont plus de chance d'être étiquetés "bruit" (et donc de diminuer le NMR en augmentant le bruit total) que des segments de 299 ou 640 secondes, qui ont eux plus de chance de contenir au moins une vocalisation, et d'être donc considérés positifs dans notre cadre de classification binaire présence / absence.

Toutefois, notre modèle atteint le deuxième meilleur NMR par rapport à ceux comparés par Schall et al. [11], comme le montre le Tableau 1. Cette performance confirme les capacités de notre modèle à distinguer entre le bruit et les signaux d'intérêt, dans la mesure où tous les modèles présentés dans le tableau ont aussi été entraînés sur des segments de 15 secondes. Cependant, malgré le fait qu'un poids plus important soit accordé à la minimisation du NMR dans le calcul de la *fitness*, notre classifieur se place 3ème pour cette métrique. Cela peut s'expliquer par un TCR plutôt bas, qui peut trouver une justification dans la durée de 15 secondes, qui a sans doute impliqué des découpages dans les vocalisations les plus longues, rendant difficile la tâche de reconnaissance de motif.

Les performances du modèle présenté ici constituent un premier essai plutôt encourageant pour la suite, notamment

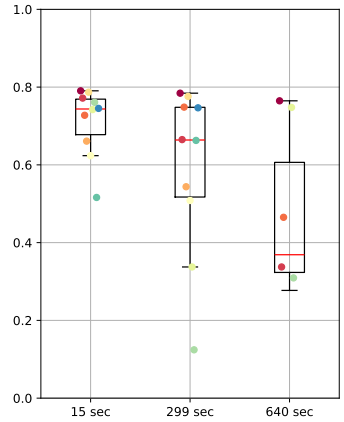

FIGURE 1 : Score de fitness pour les 3 durées. Un point correspond à une paire site-année.

Métrique		ANIMAL-SPOT	Koogu	Custom CNN	Ours
TCR	Moyenne	0.67	0.32	**0.73**	0.27
	Ecrat-type	0.24	0.18	0.10	0.20
NMR	Moyenne	**0.16**	0.32	0.26	0.17
	Ecart-type	0.11	0.24	0.15	0.20
CMR	Moyenne	**0.04**	0.14	0.06	0.09
	Ecart-type	0.03	0.10	0.03	0.11
Fitness	Moyenne	**0.83**	0.62	0.79	0.71
	Ecart-type	0.07	0.15	0.09	0.08

TABLE 1 : Scores pour les 3 modèles comparés par Schall et al. [11] et pour notre réseau encodeur/MLP.

dans la perspective de répondre à des cas d'usage où la capacité d'un classifieur à distinguer un signal d'intérêt du bruit ambiant est très importante. Par ailleurs, il est à noter que cette performance est intéressante dans la mesure où les données d'entraînement de Perch ne sont pas du tout issues de l'acoustique passive, et présentent donc nettement moins de bruit de fond que les données compilées par Miller et al. [6]. Enfin, l'entraînement de notre modèle est peu gourmand en ressources dans la mesure où l'encodage des signaux d'entrée est déjà opéré par Perch et qu'il n'est pas nécessaire de réentraîner cette partie pour obtenir des performances intéressantes.

5 Conclusion

Cet article présente une évaluation de la pertinence de Perch pour des données issues de l'acoustique passive sous-marine. Les performances sont encourageantes dans la mesure où les expériences réalisées sont restées dans un cadre relativement simple, et où de nombreuses pistes d'amélioration, comme le fait de mieux entraîner la tête de classification ou l'adaptation de domaine existent. Les performances obtenues restent comparables à l'état de l'art avec une métrique mesurant la pertinence d'un détecteur pour de futurs cas d'usage, ce qui est d'autant plus intéressant que notre modèle n'est pas très gourmand, et pourrait donc convenir à une configuration embarquée.

Références

[1] Gabriel DUBUS *et al.* : From citizen science to ai models : Advancing cetacean vocalization automatic detection through multi-annotator campaigns. *Ecological Informatics*, 81:102642, 2024.

[2] Burooj GHANI *et al.* : Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1), 2023.

[3] Louis M. HERMAN : The multiple functions of male song within the humpback whale (egaptera novaeangliae) mating system : review, evaluation, and synthesis. *Biological Reviews*, 92(3):1795–1818, 2017.

[4] John A. HILDEBRAND, Kaitlin E. FRASIER, Tyler A. HELBLE et Marie A. ROCH : Performance metrics for marine mammal signal detection and classification. *JASA*, 151(1):414–427, 2022.

[5] III MERCADO, Eduardo et Stephen HANDEL : Understanding the structure of humpback whale songs (l). *JASA*, 132(5):2947–2950, 11 2012.

[6] Brian S. MILLER *et al.* : An open access dataset for developing automated detectors of antarctic baleen whale sounds and performance evaluation of two commonly used detectors. *Scientific Reports*, 11, 2021.

[7] Ilyass MOUMMAD, Nicolas FARRUGIA et Romain SERIZEL : Regularized contrastive pre-training for few-shot bioacoustic sound detection. *In ICASSP 2024-2024*, pages 1436–1440. IEEE, 2024.

[8] Robert PLANQUÉ et Willem-Pier VELLINGA : Xeno-canto : a 21st century way to appreciate neotropical bird song. *Neotropical Birding*, pages 17–23, 2008.

[9] Laela SAYIGH *et al.* : The watkins marine mammal sound database : an online, freely accessible resource. *In Proceedings of Meetings on Acoustics*, volume 27. AIP Publishing, 2016.

[10] Julian C SCHÄFER-ZIMMERMANN *et al.* : animal2vec and meerkat : A self-supervised transformer for rare-event raw audio input and a large-scale reference dataset for bioacoustics. *arXiv preprint arXiv :2406.01253*, 2024.

[11] Elena SCHALL *et al.* : Deep learning in marine bioacoustics : a benchmark for baleen whale detection. *Remote Sensing in Ecology and Conservation*, 10(5):642–654, 2024.

[12] Yu SHIU *et al.* : Deep neural networks for automated detection of marine mammal species. *Scientific Reports*, 10:607, 2020.

[13] Dan STOWELL : Computational bioacoustics with deep learning : a review and roadmap. *PeerJ*, 10, 2022.

[14] Mingxing TAN et Quoc LE : Efficientnet : Rethinking model scaling for convolutional neural networks. *In ICML*, pages 6105–6114. PMLR, 2019.

[15] Maëlle TORTEROTOT : *Traitement et analyse de données bioacoustiques dans l'océan Indien austral : application aux baleines bleues*. Thèse de doctorat, Université de Bretagne occidentale - Brest, 2020.