FACULTÉ DES SCIENCES
UNIVERSITÉ D'ANGERS

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

ENSTA INSTITUT POLYTECHNIQUE DE PARIS

Master Mathématiques et Applications

M2 DATA SCIENCE

Année académique 2024-2025

Suivi de la bioacoustique marine à l'aide de méthodes utilisant la factorisation en matrices non négatives (NMF)

Rapport de stage

Auteur :
Mahamat Mahamat Nour Bachar

Tuteurs :
Axel Marmoret
Dorian Cazau
Mikael Escobar-Bach

Engagement de non plagiat

Nous, soussignés, déclarons être pleinement conscients que le plagiat de documents ou d'une partie d'un document publiés sur toutes formes de support, y compris internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée. En conséquence, nous nous engageons à citer toutes les sources que nous avons utilisées pour écrire ce rapport.

Résumé

La surveillance des océans est essentielle pour mieux comprendre et protéger les populations de baleines. Les enregistrements recueillis sont riches : ils contiennent à la fois des vocalisations d'intérêt et de nombreux bruits (trafic maritime, vent, etc.). Ce rapport présente un pipeline complet pour la détection d'événements sonores en bioacoustique, appliqué aux données de Miller et al. [11]. La méthode repose d'abord sur des transformations des audios en spectrogrammes par STFT. Ceux-ci sont ensuite analysés par la factorisation en matrices non négatives, qui divise le signal en spectrogrammes élémentaires interprétables. Ces spectrogrammes élémentaires sont ensuite utilisés pour détecter des événements sonores. Une version séquentielle de la NMF est également proposée pour traiter de suites d'audios. Au-delà de ce cadre standard, une approche few-shot learning est introduite : à partir de peu d'exemples connus, on oriente la détection. Enfin, une méthode est étendue à la séparation de sources, en combinant la NMF avec un filtre de Wiener, ce qui permet d'isoler les sons biologiques.

Mots-clés : STFT, NMF, détection d'événements, bioacoustique, PELT, few-shot learning, séparation de sources, filtre de Wiener.

Abstract

Monitoring the oceans is essential to better understand and protect whale populations. The collected recordings are rich : they contain both vocalizations of interest and numerous noises (shipping traffic, wind, etc.). This report presents a complete pipeline for the detection of acoustic events in bioacoustics, applied from the data of Miller et al. [11]. The method first relies on transforming audio signals into spectrograms using the STFT. These are then analyzed through non-negative matrix factorization, which decomposes the signal into interpretable elementary spectrograms. These elementary spectrograms are subsequently used to detect acoustic events. A sequential version of NMF is also proposed to handle sequences of audio recordings. Beyond this standard framework, a few-shot learning approach is introduced : starting from a small number of known examples, detection is guided accordingly. Finally, the method is extended to source separation, by combining NMF with a Wiener filter, which makes it possible to isolate biological sounds.

Keywords : STFT, NMF, event detection, bioacoustics, PELT, few-shot learning, source separation, Wiener filter.

Table des matières

Table des figures

Liste des tableaux

List of Algorithms

Présentation de la structure d'accueil

Lieu et contexte

Mon stage se déroule à Brest, au sein d'IMT Atlantique, avec des échanges avec ENSTA Bretagne. Le sujet porte sur le suivi de la biodiversité marine à partir d'enregistrements audios sous-marins (acoustique passive) et sur le développement d'outils d'IA non supervisée pour la détection d'événements sonores et la séparation. L'objectif est de proposer des méthodes efficaces, sobres en calcul et interprétables.

Encadrants

Axel Marmoret (IMT Atlantique) Maître de conférences à IMT Atlantique (Lab-STICC) dans l'équipe BRAIn, où j'ai été accueilli. Axel Marmoret travaille sur des méthodes d'IA pour le traitement du signal audio. Ses recherches portent notamment sur les factorisations non négatives (NMF/NTD) et leurs applications. Il intervient sur les choix méthodologiques.

Dorian Cazau (ENSTA) Maître de conférences à ENSTA (Lab-STICC) dans l'équipe OSE, Dorian Cazau travaille sur l'acoustique passive sous-marine pour l'océanographie observationnelle, avec un intérêt pour la bioacoustique (mammifères marins), l'écologie par des méthodes issues de l'intelligence artificielle et du traitement du signal. Il apporte l'expertise domaine (bioacoustique marins) et les cas d'usage.

Équipe d'accueil : BRAIn (IMT Atlantique)

Mon stage a lieu dans l'équipe **BRAIn**(BRoader Artificial Intelligence), rattachée au département MEE (Mathématiques, Électronique et Informatique) d'IMT Atlantique et au laboratoire Lab-STICC (UMR CNRS 6285).

Ses axes de recherche portent sur les représentations dans l'intelligence artificielle, particulièrement :

— Thrifty AI : par exemples apprendre avec peu de données ou peu d'annotations (few-shot learning),

— l'IA compressée : réduire les coûts en calcul et en énergie, favoriser des méthodes sobres.

Parmi ses domaines applicatifs figurent la vision par ordinateur, la neuroimagerie et les signaux audio.

Chapitre 1

Introduction

L'océan est un milieu riche en biodiversité, mais il est difficile à observer directement. Pour mieux comprendre la présence et le comportement des animaux marins, les chercheurs utilisent l'*acoustique passive*. Elle consiste à écouter les sons produits sous l'eau grâce à des capteurs, appelés *hydrophones*. Ces capteurs enregistrent en continu de longs fichiers audio, qui contiennent à la fois :

— des sons d'animaux (biophonie),
— des sons liés aux activités humaines, comme ceux des bateaux (anthropophonie),
— et des sons naturels comme la pluie ou le vent (géophonie).

L'analyse de ces enregistrements est un enjeu scientifique et pratique important. Elle aide à mieux suivre les populations marines et donc à les protéger. Mais ces données sont énormes : il devient vite impossible de tout analyser manuellement. La question est donc la suivante : comment détecter automatiquement les sons intéressants, par exemple les chants de baleines, au milieu de tout ce bruit ?

> La problématique de ce travail est donc : *Comment développer des méthodes, simples à comprendre et sobres, pour détecter des événements sonores dans de longs enregistrements marins et séparer les différents sons mélangés, tout en évitant le recours aux modèles d'apprentissage statistique ?*

La motivation est double :

— Proposer des outils aux biologistes marins, qui disposent déjà de ces enregistrements mais manquent de moyens adaptés pour les traiter.
— Étudier des nouvelles approches afin d'améliorer l'état de l'art en détection bioacoustique.

Ce rapport débute par un chapitre consacré à l'état de l'art, qui présente les approches existantes en bioacoustique pour les tâches de la détection et séparation. Le chapitre suivant introduit les représentations audio et la factorisation en matrices non négatives (NMF), puis décrit un premier pipeline de détection basé sur la segmentation par ruptures, ainsi qu'une extension séquentielle de la NMF. Le troisième chapitre propose une approche de type *few-shot*, permettant d'exploiter quelques exemples annotés pour améliorer la détection. Le quatrième chapitre explore l'utilisation de la NMF pour la séparation de sources sonores. Enfin, le rapport se conclut par des perspectives.

Chapitre 2

Travaux existants

2.1 Méthodes de détection

2.1.1 Détection par l'énergie

Une méthode simple pour détecter des événements sonores est la détection par l'énergie. Il s'agit de mesurer l'énergie dans une bande de fréquences et de déclencher une détection lorsque cette énergie dépasse un seuil prédéfini. Ce détecteur « seuil d'énergie » ne requiert ni modèle de signal ni phase d'apprentissage, ce qui en fait une solution facile à implémenter. Il fonctionne bien pour des événements occupant une bande de fréquence distincte [10]. Par exemple, si une baleine produit des appels entre 15 et 20 kHz, un détecteur énergétique centré sur cette bande peut identifier ces appels tant que le bruit de fond dans cette bande reste faible. Cependant, cette approche a des limites : elle est sensible au bruit ambiant et génère potentiellement de fausses détections si des bruits non biologiques font des élévations d'énergie dans la bande considérée. En présence de bruit ou d'autres appels dans la même plage de fréquences, un simple seuil d'énergie manque de sélectivité et pourrait confondre du bruit avec un son cible.

2.1.2 Filtrage adapté stochastique

Cette méthode est une variante de filtre adapté(MF), où on modélise l'événement sonore d'intérêt comme un processus stochastique doté de certaines propriétés statistiques connues. Techniquement, le filtre adapté stochastique construit une banque de filtres ou un filtre paramétrique qui prend en compte la variabilité du signal et du bruit. Bouffaut et al.[3] en ont démontré l'efficacité pour la détection des appels de baleine bleue antarctique(appels Z). Cette méthode est difficile à utiliser (elle requiert d'estimer et/ou de supposer certaines statistiques du signal et du bruit) [3].

2.1.3 Méthodes basées sur l'apprentissage profond

Les méthodes d'apprentissage profond, notamment les réseaux convolutifs (CNN) et récurrents, se sont imposées depuis les années 2010 pour la détection de vocalisations marines. Entraînés sur de grands ensembles de données annotées, ces modèles surpassent les approches traditionnelles. Ils excellent dans l'extraction de motifs complexes et permettent non seulement de détecter, mais aussi de classifier les vocalisations par espèce ou type d'appel, ce qui facilite la mise en place de systèmes de surveillance autonomes. Toutefois, leur efficacité repose sur la disponibilité de larges bases de données annotées, coûteuses en ressource, et ils souffrent de problèmes d'interprétabilité (boîte noire) et de généralisation (performances volatiles selon le temps et l'espace de l'enregistrement).

2.2 Méthodes de séparation de sources basées sur NMF

Depuis quelques années, des méthodes issues du traitement des musiques ont été reprises en bioacoustique. Parmi elles, la factorisation en matrices non négatives semble intéressantes pour extraire des vocalisations d'enregistrements, mais on y reviendra plus détails sur cette méthode.

Une extension de cette approche est la NMF à codage de périodicité (PC-NMF) proposée par Lin et al. dans [9]. Cette méthode introduit une contrainte supplémentaire pour analyser des audios : elle regroupe des parties de l'audio qui ont la même périodicité. Grâce à ce procédé, Lin et al. ont réussi à séparer les vocalisations de cétacés et les chants de poissons du bruit non biologique (le nombre de sources à séparer est fourni, mais sans aucun exemple étiqueté).

L'inconvénient de la PC-NMF réside dans l'utilisation de la périodicité des vocalisations cibles, ainsi que dans la nécessité de fixer à l'avance le nombre de composantes à extraire.

En résumé, les méthodes classiques de détection restent simples et performants dans des cas bien définis, mais elles montrent rapidement leurs limites face au bruit, à la variabilité des signaux. Les approches par apprentissage profond offrent des performances élevées mais dépendent fortement de jeux de données annotées et posent des problèmes de généralisation et d'interprétabilité. Enfin, les méthodes basées sur la NMF constituent une alternative intéressante pour séparer des sources et détecter des vocalisations et ouvrent ainsi de perspectives.

Chapitre 3

Méthodologie : de l'audio à la détection des événements sonores

3.1 Représentation temps–fréquence des signaux audio

3.1.1 Transformée de Fourier à court terme (STFT)

Les audios sont des séries temporelles $x(t)$, dont le traitement direct est difficile car trop bruité. Une solution classique est d'utiliser la *transformée de Fourier à court terme (STFT)*, qui décompose le signal en une suite de spectres fréquentiels calculés sur des fenêtres temporelles "glissantes".

En pratique, on utilise la version discrète et échantillonnée :

$$X(f,t) = \sum_{n=0}^{N-1} x[n+tH]\, w[n]\, e^{\frac{-i\,2\pi\, fn}{N}}, \tag{3.1}$$

où :

— $w[n]$ est une fenêtre(par ex. Hann)(de taille win_length),
— N est la longueur de la fenêtre(n_fft),
— H est le pas de décalage (hop_length) entre deux fenêtres successives,

Le *spectrogramme* est défini comme le module au carré de la STFT :

$$S(f,t) = |X(f,t)|^2. \tag{3.2}$$

Il est aussi l'énergie en fonction du temps t et de la fréquence f.

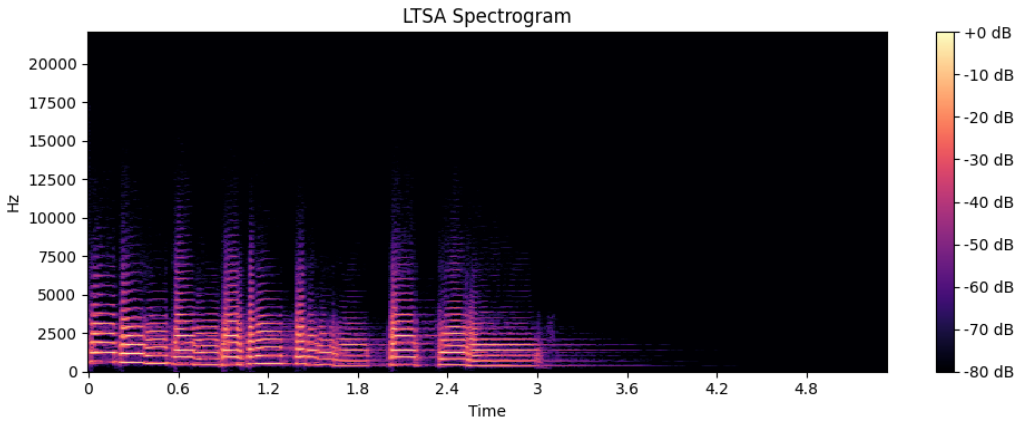


FIGURE 3.1 – Spectrogramme d'un extrait audio obtenu avec la STFT

La taille des fenêtres est un paramètre crucial, car elle détermine la résolution temps–fréquence :

— une fenêtre large améliore la résolution fréquentielle mais dilue l'information temporelle,
— une fenêtre courte améliore la localisation temporelle mais réduit la précision fréquentielle.

Ainsi, la STFT constitue donc un compromis entre l'analyse purement fréquentielle (transformée de Fourier classique) et l'analyse temporelle.

Toutefois, lorsqu'on s'intéresse à des enregistrements de longue durée, le spectrogramme complet devient rapidement lourd à manipuler, tant en mémoire qu'en calcul. Une solution consiste alors à **compresser** l'information spectrale sans perdre la structure acoustique.

3.1.2 Long-Term Spectral Average (LTSA)

Une approche utilisée en bioacoustique est le Long-Term Spectral Average (LTSA). Traditionnellement, la LTSA consiste à calculer la moyenne spectrale sur de longues fenêtres temporelles (d'où le terme Average). Dans notre cas, on remplace la moyenne par la *médiane*, afin de rendre la représentation plus robuste aux outliers.

Formellement, pour un spectrogramme $S(f,t)$, la LTSA est donnée par :

$$\mathrm{LTSA}_{\mathrm{med}}(f,t) = \mathrm{med}_{m\in\mathcal{W}_k}\, S(f,m), \tag{3.3}$$

où $\mathcal{W}_k$ désigne une fenêtre temporelle(de k secondes).

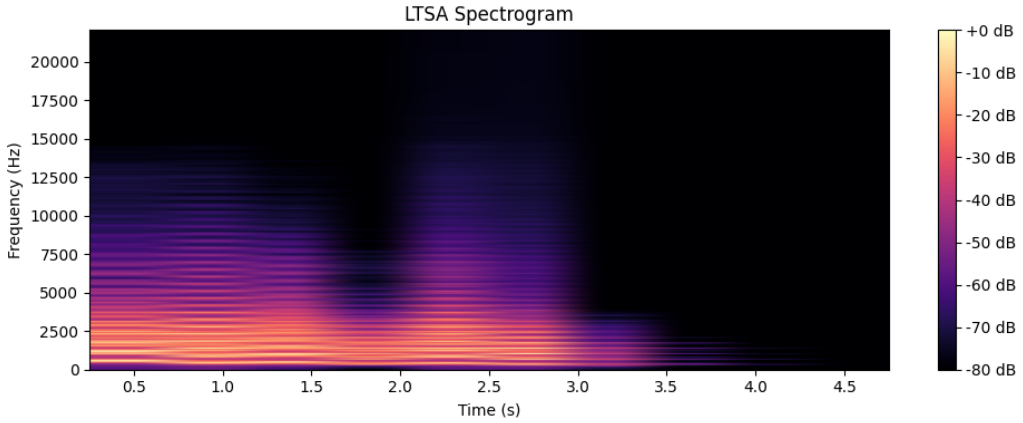


FIGURE 3.2 – Exemples de représentation LTSA

Le spectrogramme contient à la fois des événements d'intérêt (par ex. appels de baleines) et des sons parasites (par ex. bruit, fond marin...). La factorisation en matrices non négatives (NMF) constitue une méthode adaptée pour décomposer ce spectrogramme en dictionnaires ou bases spectraux *interprétables* et en activations temporelles associées.

3.2 Factorisation matricielles non négatives(Non-negative Matrix Factorization, NMF)

3.2.1 Principe général

La NMF, initialement proposée par Lee et al.[8] pour décomposer des images en parties additives (cf.§3.3), approxime une matrice non négative $V \in \mathbb{R}_+^{F\times T}$ (par exemple un spectrogramme fréquence–temps issu de la STFT ou du LTSA) par une factorisation en deux matrices non négatives $W \in \mathbb{R}^{F\times K}+$ et $H \in \mathbb{R}_+^{K\times T}$ Formellement, on cherche :

$$V \approx WH, \qquad W \in \mathbb{R}_+^{F\times K}, \quad H \in \mathbb{R}_+^{K\times T}. \tag{3.4}$$

Le paramètre K désigne le rang de factorisation, autrement dit le nombre de composantes. Les colonnes de W constituent des *bases fréquentielles* ou *dictionnaires spectraux*, tandis que les lignes de H correspondent à leurs *activations temporelles*. Cette décomposition est interprétable : chaque base spectrale (colonne de W) représente une **partie additive du spectrogramme**, et H indique les instants où cette base est activée. Dans la pratique, W et H sont estimés en minimisant une mesure de divergence entre V et WH (par ex. la divergence euclidienne, de Kullback–Leibler ou d'Itakura–Saito), ce qui conditionne la nature de l'approximation obtenue.

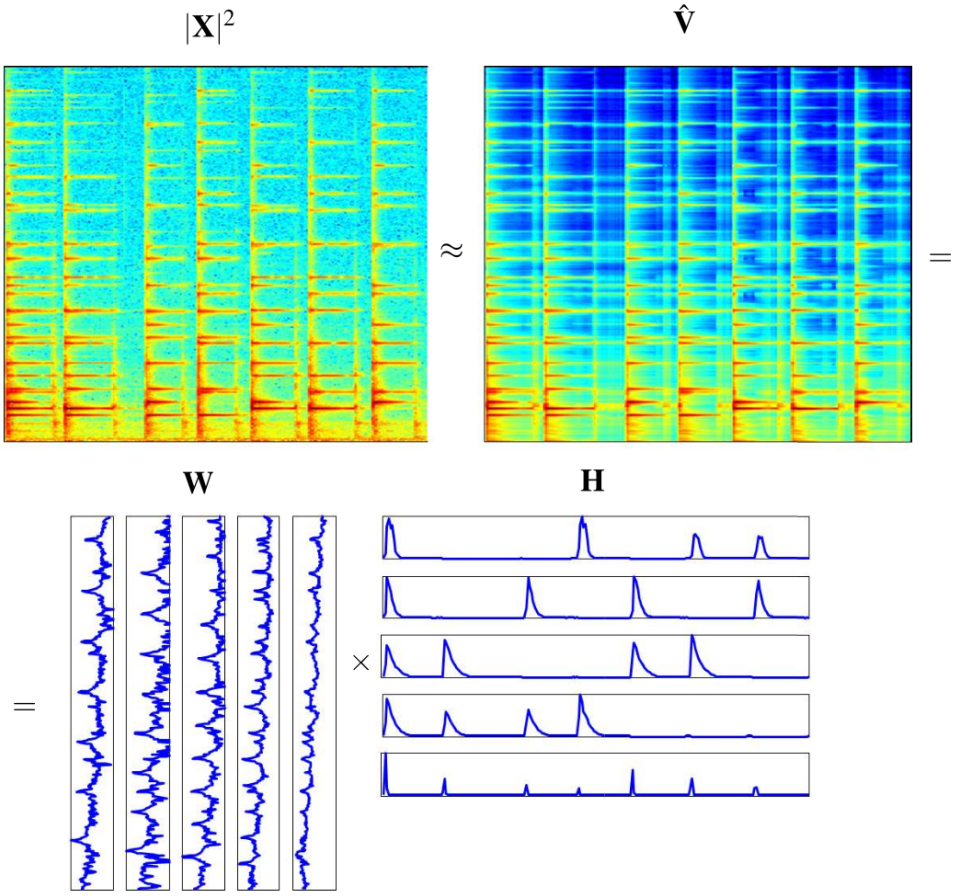


FIGURE 3.3 – Décomposition NMF d'un spectrogramme en bases fréquentielles W et activations temporelles H(source : Févotte et al. [6])

3.2.2 Optimisation par divergence

La factorisation s'obtient en minimisant une *divergence* entre V et WH. Formellement, elle s'écrit comme suit :

$$\min_{W\geq 0,\, H\geq 0} \; D_\beta(V \,\|\, WH) \tag{3.5}$$

où D_β est la **divergence** β :

$$\beta = 2 \;\Rightarrow\; D_\beta = \|V - WH\|_F^2 \quad \text{(norme de Frobénius)},$$
$$\beta = 0 \;\Rightarrow\; D_\beta = D_{\mathrm{IS}}(V \,\|\, WH) \quad \text{(Itakura–Saito)}.$$

En pratique, on utilise des mises à jour itératives (*multiplicative updates*, HALS, etc.) adaptées à la divergence choisie plutôt qu'une descente de gradient afin d'assurer la positivité des coefficients dans WH.

Dans ce travail, la divergence Itakura–Saito (IS) servira de fonction de perte, couramment utilisée pour l'analyse des audios et adaptée à la bioacoustique grâce à sa propriété d'invariance d'échelle [5]. La formule de cette divergence est comme suit :

$$D_{\mathrm{IS}}(V\|WH) = \sum_{f,t}\left(\frac{V_{ft}}{(WH)_{ft}} - \log\left(\frac{V_{ft}}{(WH)_{ft}}\right) - 1\right). \tag{3.6}$$

Les règles multiplicatives de mise à jour(*Multiplicative Updates*, MU) introduites par [8], dans le cas de la divergence IS, elles s'écrivent comme suit :

$$H \leftarrow H \odot \frac{W^\top (V \odot (WH)^{-2})}{W^\top (WH)^{-1}}, \qquad W \leftarrow W \odot \frac{(V \odot (WH)^{-2})\, H^\top}{(WH)^{-1} H^\top}, \tag{3.7}$$

où $\odot$ désigne le produit terme par terme (Hadamard).

3.2.3 Choix du rang K

Le rang K fixe le nombre de motifs élémentaires. Il véhicule une hypothèse forte : *combien* de composants sont nécessaires pour expliquer le mélange fond+événements. Un K trop faible mélange des phénomènes distincts (perte d'interprétation) ; un K trop grand fragmente l'information fréquentielle (sorte de sur-apprentissage).

Plusieurs critères empiriques permettent d'optimiser ce choix dont :

- Courbe d'épaule ("elbow") sur la divergence $D_\beta(V\|WH)$ en fonction de K : choisir le plus petit K au-delà duquel le gain dans la fonction de perte devient faible.
- Validation orientée tâche : si l'objectif est la *détection* d'événements, sélectionner K qui maximise une métrique (F1, PR et R) sur un jeu de validation annoté.
- Tâches de séparation : choisir K en fonction de la qualité de séparation obtenue (par ex. SNR, SI-SDR) sur un jeu de validation de référence.

Le pseudo-algorithme suivant illustre le principe de la NMF avec divergence d'Itakura–Saito et mises à jour multiplicatives :

Algorithm 1: NMF avec divergence IS et MU

Entrée: Matrice non négative $V \in \mathbb{R}_+^{F\times T}$, rang K, itérations N.
Sortie: $W \in \mathbb{R}_+^{F\times K}$, $H \in \mathbb{R}_+^{K\times T}$ tels que $V \approx WH$

Initialiser $W^{(0)}, H^{(0)}$ (aléatoire, SVD non négatives, ...)
for $t \leftarrow 1$ **to** N **do**
  `// Mise à jour de` W
  $W \leftarrow W \odot \dfrac{(V \odot (WH)^{-2})\, H^\top}{((WH)^{-1})\, H^\top}$
  `// Mise à jour de` H
  $H \leftarrow H \odot \dfrac{W^\top (V \odot (WH)^{-2})}{W^\top ((WH)^{-1})}$
  if $\mathcal{L}_{\mathrm{IS}}(V\|WH) < \varepsilon$ **then**
    arrêt

où $\odot$ désigne le produit de Hadamard(produit terme à terme)

Remarque. Dans la suite, je choisis d'utiliser la LTSA avec la **médiane** comme outil de condensation. Ce choix permet de réduire le coût des calculs tout en conservant une information suffisante pour les tâches finales.

Après ces étapes, on peut désormais aborder la question de la détection d'événements sonores, qui constitue l'un des objectifs de ce travail.

3.3 Détection de ruptures

3.3.1 Principe général

La détection de ruptures (ou *changepoint detection*) consiste à identifier les instants $t = \tau_1, \ldots, \tau_m$ où les propriétés statistiques d'une série temporelle changent [13]. Une rupture peut correspondre à une modification significative d'une ou plusieurs statistiques. Dans notre contexte, une rupture dans une activation $H_{i,:}$ est supposée l'apparition ou la disparition d'un événement sonore.

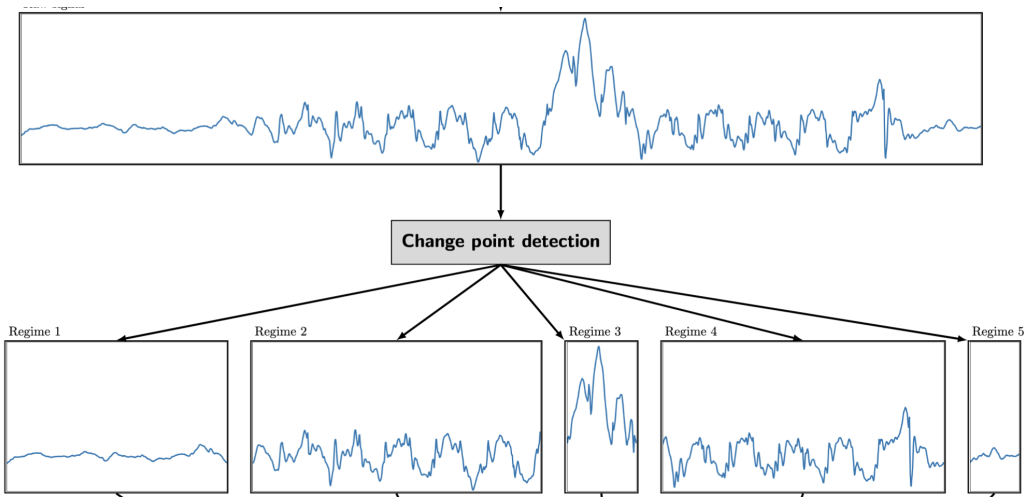


FIGURE 3.4 – Exemple d'une segmentation par détection de ruptures(Source : Truong et al.[13])

3.3.2 Méthode PELT

L'algorithme Pruned Exact Linear Time (PELT)[7] est une méthode optimisée pour la détection de points de changement dans une série $y = (y_1, \ldots, y_T)$. Il repose sur l'optimisation du critère pénalisé suivant :

$$\min_{m,\ \tau_{1:m}} \sum_{i=1}^{m+1} C\big(y_{(\tau_{i-1}+1):\tau_i}\big) \ + \ \lambda\, f(m), \qquad \text{avec } \tau_0 = 0,\ \tau_{m+1} = T. \tag{3.8}$$

où $C(\cdot)$ est un coût intra-segment (par exemple, variance quadratique). Le paramètre de pénalité λ contrôle le nombre de ruptures détectées : une valeur trop faible entraîne une surdétection, tandis qu'une valeur trop forte risque de négliger certains changements. PELT garantit une solution exacte et s'exécute en temps linéaire dans la longueur de la série d'après [7].

3.3.3 La librairie `ruptures`

Dans la pratique, la librairie Python `ruptures`, introduite dans [13], implémente PELT ainsi que d'autres méthodes de segmentations. Dans la pratique, nous utilisons la librairie Python ruptures [13], qui implémente PELT ainsi que d'autres méthodes de segmentation. Ainsi, elle encode :

- le coût intra-segment $C(\cdot)$ via la méthode `model()` (p. ex. `"l2"` pour une perte quadratique, `"l1"`, etc.) ;
- la pénalité $\lambda f(m)$ via le paramètre `pen` dans `predict()`. Par défaut, `ruptures` suppose une pénalité linéaire $f(m) = m$, si bien que `pen` joue directement le rôle du λ de (3.8).

Listing 3.1 – Exemple PELT avec coût quadratique.

```
n = 200
y = npr.normal(0, 1, n) #signal jouet

algo = rpt.Pelt(model="l2").fit(y) #choix du cout C(.) via model="l2"

lambda_pen = 10.0  #pen=lambda*f(m) avec f(m)=m implicite

bkps = algo.predict(pen=lambda_pen) #indices de rupture d tect s(y
   compris y[n])
```

Ainsi, l'utilisation de l'algorithme PELT permet de trouver les instants de rupture dans les activations $H_{i,:}$, interprétés comme l'apparition ou la disparition d'événements sonores. Les points de rupture détectés sont notés $\{\tau_1^i, \ldots, \tau_{M_i}^i\}$, ce qui définit une partition de $H_{i,:}$ en segments.

$$S_{i,j} = [\tau_{j-1}^i, \tau_j^i[, \quad j = 1, \ldots, M_i, \tag{3.9}$$

avec $\tau_0^i = 0$ et $\tau_{M_i}^i = T$.

L'aire d'un segment $S_{i,j}$ est donnée par :

$$A_{i,j} = \sum_{t=\tau_{j-1}^i}^{\tau_j^i - 1} H_{i,t}. \tag{3.10}$$

Pour chaque composante i, on définit un seuil θ_i comme le quantile q des aires segmentaires $\{A_{i,j}\}_{j=1}^{M_i}$:

$$\theta_i = \text{Quantile}_q\Big(\{A_{i,j} \mid j = 1, \ldots, M_i\}\Big), \quad q \in (0, 1).$$

Par exemple, $q = 0.9$ conserve uniquement les 10% des segments les plus "forts".

Un segment $S_{i,j}$ est conservé si

$$A_{i,j} \geq \theta_i,$$

ce qui permet de filtrer les ruptures dues à des segments faibles.

Pour la suite, nous utiliserons le coût dit normal dans `ruptures`, basé sur une approximation gaussienne. On suppose que chaque segment $S_{i,j}$ suit une loi normale $\mathcal{N}(\mu, \Sigma)$.

Le coût associé est alors défini par :

$$C_{\text{Normal}}(S_{i,j}) \;=\; (\tau_j^i - \tau_{j-1}^i)\, \log\det\big(\hat{\Sigma} + \varepsilon I\big), \tag{3.11}$$

où $\hat{\Sigma}$ est la covariance empirique du segment $S_{i,j}$ et εI un terme de régularisation. Ce critère s'est montré le meilleur pour la tâche de détection.

3.4 Données

Une partie des enregistrements audio publiés par Miller et al. [11], aussi utilisés comme données d'entraînement dans la *tâche 2* du BioDCASE Challenge 2025, sert de base à la tâche de détection. Ces données comprennent plusieurs sessions d'enregistrement réalisées sur différents sites en Antarctique, contenant des appels de baleines bleues ainsi que de rorquals bleus.

On remarque que les données présentent une grande variabilité en termes de taille et de richesse événementielle. Par exemple, le jeu de données d'Elephant Island 2014 est très riche en événements, avec plus de 2500 enregistrements et un ratio d'événements de 13%, ce qui en fait

Site-year	Hydrophone	Nb recordings	Total duration (h)	Event duration	Event/total ratio (%)
Balleny Islands 2015	PMEL-AUH	205	204.0	02 :46.32	1.4
Casey 2014	AAD-MAR	194	194.0	14 :12.00	7.3
Elephant Island 2013	AURAL	2247	187.0	16 :06.58	8.6
Elephant Island 2014	AURAL	2595	216.0	28 :08.25	13.0
Greenwich 2015	Sono.Vault	190	31.7	02 :03.11	6.5
Kerguelen 2005	HARP	200	200.0	03 :31.37	1.8
Maud Rise 2014	AURAL	200	83.3	05 :42.45	6.9
Ross Sea 2014	PMEL-AUH	176	176.0	08 :51.00	5.0
TOTAL	—	6007	1292.0	—	5.2

TABLE 3.1 – Statistiques des données

une source majeure d'exemples positifs. À l'inverse, certains sites comme Kerguelen présentent un ratio d'événements beaucoup plus faible ($< 2\%$), ce qui est un défi pour la détection.

L'évaluation est faite au **niveau segment**, en faisant correspondre la détection et les vérités terrain par recouvrement temporel.

Un segment détecté $\hat{S}$ est un *vrai positif* (TP) s'il recouvre un segment de référence S avec un IoU (Intersection over Union) dépassant un seuil τ_{IoU} :

$$\text{IoU}(\hat{S}, S) \;=\; \frac{|\hat{S} \cap S|}{|\hat{S} \cup S|} \;\geq\; \tau_{\text{IoU}}.$$

Sinon, $\hat{S}$ est un FP. Les segments de référence non appariés sont des FN. Sauf mention contraire, τ_{IoU} utilisé pour le seuillage vaut 0.3.

Paramètres

Les paramètres utilisés pour l'échantillonnage des audios sont ceux du *baseline* du BioDCASE Challenge 2025.

TABLE 3.2 – Paramètres d'échantillonage des audios

Paramètre	Valeur
Fréquence d'échantillonnage (sr)	250
Longueur fenêtre FFT (n_fft)	512
Taille fenêtre(Hann) (win_length)	512
Décalage (hop_length)	20
Fenêtre temporelle (pour LTSA, $\mathcal{W}_k$)	0,5 s

Cette fréquence d'échantillonnage basse est adaptée aux chants des baleines, qui se situent dans les basses fréquences. De plus, l'utilisation d'une fenêtre et d'un décalage réduits permet d'obtenir une meilleure résolution temporelle, ce qui favorise la détection précise des événements courts.

TABLE 3.3 – Hyperparamètres de la méthode

Paramètre	**Valeur défaut**	**Intervalles testées**
NMF		
Rang	6	4–12
Initialisation	nndsvd	–
Itérations max.	500	–
Tolérance	10^{-10}	–
Détection		
Pénalité	–	0–35
Seuil percentile détection	–	70-99

Concernant les hyperparamètres, un rang de décomposition $K = 6$ s'est avéré le meilleur pour la détection. L'initialisation par nndsvd a favorisé une convergence rapide. Les paramètres de détection traduisent un compromis entre rappel et précision : des seuils plus bas auraient augmenté la sensibilité mais au prix de nombreux faux positifs, tandis que des seuils plus stricts auraient réduit la détection.

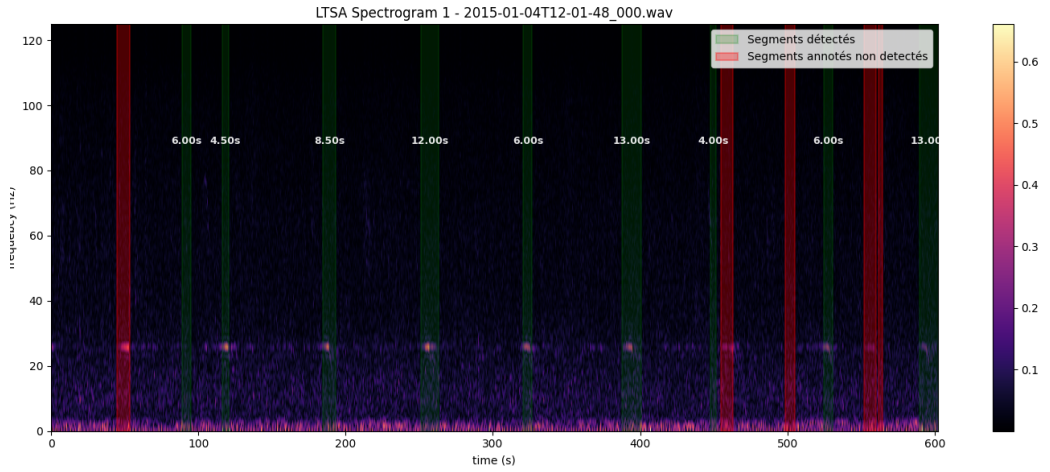
3.5 Résultats

Une étude quantitative (F1, Precision, Recall) est faite sur certains jeux de données, en se basant sur des critères tels que le ratio événement/durée et la durée totale des enregistrements. Sauf mention contraire, les hyperparamètres par défaut sont ceux des tableaux 3.3 et 3.2.

TABLE 3.4 – Résultats de détection pour quelques jeux de données ($\tau_{\text{IoU}} = 0,3$).

Jeux de données	**Précision**	**Rappel**	**F1-score**
Elephant Island 2014	0,17	0,57	0,24
Elephant Island 2013	0,15	0,48	0,198
Greenwich 2015	0,06	0,49	0,10
Casey 2014	0,06	0,46	0,09

Un exemple de détection sur un échantillon d'enregistrement de Greenwich :



Le tableau 3.5 présente les scores obtenus avec un modèle YOLOv11(basé sur des réseaux de neurones), utilisé comme baseline dans le BioDCASE Challenge 2025 [2]. Il est important

de rappeler qu'une comparaison directe n'est pas pertinente : les réseaux de neurones profonds exploitent des capacités d'apprentissage bien supérieures à celles d'une factorisation NMF non supervisée, notamment grâce à un finetuning spécifique pour la détection.

Néanmoins, cette mise en relation aide à situer nos résultats. Par exemple, sur Elephant Island 2014, notre approche NMF atteint un F1-score de $0,24$(Table 3.4), proche de celui du YOLOv11 $(0,34)$. Sur Casey 2014, l'écart est plus marqué $(0,09$ contre $0,24)$, tandis que sur Greenwich 2015, nos résultats (F1=$0,10$) restent bien en deçà du baseline $(0,39)$.

Ainsi, la factorisation NMF, beaucoup plus simple et légère en calcul, peut-être une alternative intéressante dans des situations où l'entraînement d'un réseau est difficile (ressources limitées, rareté d'annotations, besoin d'interprétabilité).

Par ailleurs, ces résultats suggèrent que des améliorations sont possibles : une étape de prétraitement plus ciblée (filtre bande passante selon l'espèce visée) ou un ajustement adaptatif des seuils de détection selon l'heure ou le contexte de l'enregistrement, ainsi qu'une analyse fine des erreurs (faux positifs, faux négatifs liés à des chants masqués par un bruit). Enfin, une extension vers une détection différenciée selon le type de chant de baleine constituerait une piste prometteuse pour affiner les performances.

Les scores du modèle baseline de DCASE[2] sont les suivants :

TABLE 3.5 – Scores par jeux (toutes classes confondues) avec le YOLOV11.

Jeux de données	Precision	Recall	F1
ballenyislands2015	0,193	0,213	0,202
casey2014	0,370	0,180	0,243
elephantisland2013	0,474	0,360	0,409
elephantisland2014	0,346	0,334	0,340
greenwich2015	0,710	0,271	0,393
kerguelen2005	0,366	0,232	0,284
maudrise2014	0,439	0,050	0,090
rosssea2014	0,571	0,038	0,072

La factorisation NMF permet d'obtenir des dictionnaires fréquentiels intéressants, mais elle considère chaque audio de manière indépendante. Or, dans de nombreux cas, les audios successifs présentent une certaine similarité. Pour utiliser cette continuité et améliorer l'efficacité de la méthode, une variante a été développée : la NMF séquentielle (SeqNMF).

3.6 NMF séquentielle

3.6.1 Principe

Le *NMF séquentielle* (SeqNMF) est une variante qui vise à exploiter la similarité entre plusieurs audios consécutifs, et paraît plus rapide que la NMF classique. Chaque audio $i = 1, 2, \ldots$ est représenté par un spectrogramme $V^{(i)} \in \mathbb{R}_+^{F \times T_i}$. On cherche une factorisation $V^{(i)} \approx W^{(i)} H^{(i)}$ avec $W^{(i)} \in \mathbb{R}_+^{F \times K}$ et $H^{(i)} \in \mathbb{R}_+^{K \times T_i}$ en minimisant la divergence β (§3.2) :

$$\min_{W^{(i)} \geq 0,\, H^{(i)} \geq 0} D_\beta\big(V^{(i)} \,\|\, W^{(i)} H^{(i)}\big). \tag{3.12}$$

La spécificité de l'approche est l'initialisation sous forme d'une *suite* des bases fréquentielles, inspiré des réseaux de neurones :

$$W_{(0)}^{(i)} \leftarrow \begin{cases} W^{\text{custom}} & \text{si } i = 1 \quad \text{(bases initiales)}, \\ W^{(i-1)} & \text{si } i \geq 2 \quad \text{(transfert des bases)}, \end{cases} \qquad H_{(0)}^{(i)} \leftarrow \text{NNDSVD}_H\big(V^{(i)}, K\big).$$

3.6.2 Algorithme

Algorithm 2: NMF séquentielle (SeqNMF)

Entrée: $\{V^{(i)}\}_{i=1}^{T}$, rang K, itérations max $N_{\max}$, tolérance ε.
Sortie: $\{W^{(i)}, H^{(i)}\}_{i=1}^{T}$ tels que $V^{(i)} \approx W^{(i)}H^{(i)}$.

Construire/charger W^{custom}
for $i \leftarrow 1$ **to** T **do**
  `// Initialisation pour le bloc` i
  if $i = 1$ **then**
    $W \leftarrow W^{\text{custom}}$
    $H \leftarrow \text{NNDSVD}_H\big(V^{(1)}, K\big)$
  else
    $W \leftarrow W^{(i-1)}$
    $H \leftarrow \text{NNDSVD}_H\big(V^{(i)}, K\big)$
  for $t \leftarrow 1$ **to** $N_{\max}$ **do**
    `// Mises à jour multiplicatives (IS,` β`=0) cf.` (3.7)
    $$W \leftarrow W \odot \frac{\left(V^{(i)} \odot (WH)^{-2}\right) H^{\top}}{\left((WH)^{-1}\right) H^{\top}}$$
    $$H \leftarrow H \odot \frac{W^{\top}\left(V^{(i)} \odot (WH)^{-2}\right)}{W^{\top}\left((WH)^{-1}\right)}$$
    if $\mathcal{L}_{\text{IS}}\big(V^{(i)} \| WH\big) < \varepsilon$ **then**
      break
  $W^{(i)} \leftarrow W$
  $H^{(i)} \leftarrow H$

On applique ensuite la détection de ruptures (§3.3) séparément sur l'activation de chaque composante $H^{(t)}_{i,:}$. Les segments gardés par filtrage-percentile définissent les *prédictions* $\hat{S}$ évaluées par IoU (§3.4).

3.6.3 Avantages et limites

Le SeqNMF présente quelques avantages :
— accélération de la convergence par réutilisation des bases du $W^{(i-1)}$,
— stabilisation des composantes fréquentielles lorsque les audios successifs sont *similaires* géométriquement,
— possibilité d'intégrer des bases personnalisées W^{custom} au premier audio(par exemple des bases fréquentielles issues d'annotations manuelles du 1er audio).

Sur des données consécutives (par ex. *Elephant Island 2013–2014*), on remarque des performances équivalentes à une factorisation indépendante, mais avec environ 10% de temps en moins.

Ses limites apparaissent lorsque les audios successifs diffèrent : cette approche pourrait ralentir la convergence, voire piéger l'algorithme dans un minimum local, surtout si $N_{\max}$ est faible. Ainsi, sur des données hétérogènes ou désordonnées (par ex. *Greenwich 2015*), une dégradation des performances de la détection est observée.

En résumé, le SeqNMF peut être très efficace lorsque les audios successifs sont similaires, mais il atteint vite ses limites dès que les données deviennent trop variées. Pour mieux gérer ces situations, une autre approche a été explorée : la NMF Few-shot, une sorte d'apprentissage avec quelques annotations, qui cherche à utiliser la capacité d'adaptation aux nouveaux contextes.

Chapitre 4

NMF à apprentissage par peu d'exemples

Dans ce chapitre, nous présentons une approche avec peu d'exemples (*few-shot*) fondée sur la NMF (§3.2). À partir de quelques segments annotés de l'événement cible, nous apprenons un dictionnaire spectral *cible* W_{tar}, que nous complétons par un dictionnaire de *fond* W_{bg}. Le dictionnaire composé $W_{\text{train}} = [W_{\text{tar}}, W_{\text{bg}}]$ sert ensuite à factoriser la partie non annotée de l'audio. Les activations associées au dictionnaire cible serviront alors pour détecter les occurrences de l'événement.

4.1 Principe

Notons *tar* la partie *cible* et *bg* la partie *fond ou non visée*. Soit $V \in \mathbb{R}_+^{F\times T}$ un spectrogramme. La NMF cherche $W \in \mathbb{R}_+^{F\times K}$ et $H \in \mathbb{R}_+^{K\times T}$ de la suite du spectogramme tels que $V \approx WH$ (cf. (3.5)), avec une décomposition de rang

$$K \;=\; K_{\text{tar}} + K_{\text{bg}}, \qquad W \;=\; \begin{bmatrix} W_{\text{tar}} & W_{\text{bg}} \end{bmatrix}.$$

À partir de k segments annotés, on apprend W_{tar} sur les trames correspondantes, puis W_{bg} sur les trames restantes. En inférence, on garde W fixé et on estime H ; les activations H_{tar} (lignes de H associées à W_{tar}) sont alors utilisées pour la détection (voir §3.3 pour la segmentation par ruptures et le filtrage par quantile).

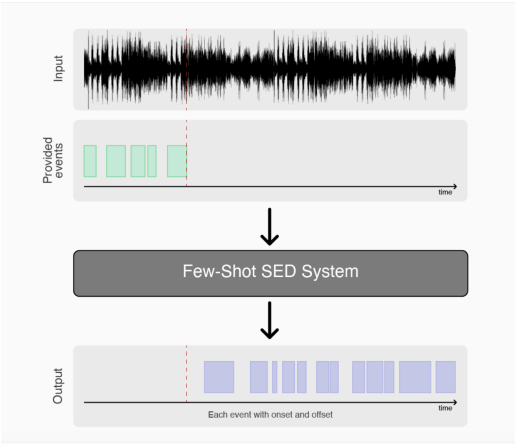


FIGURE 4.1 – Illustration de FewShot Learning(figure provenant de [1])

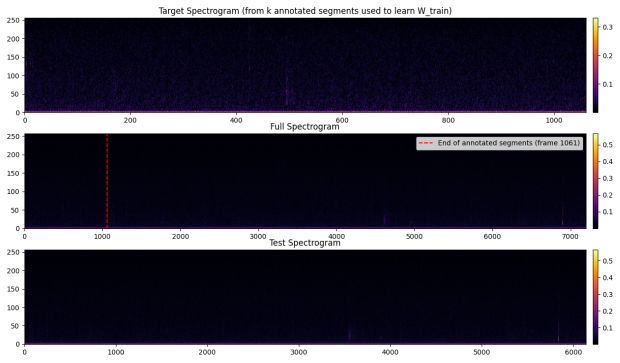


FIGURE 4.2 – Illustration de NMFewShot

4.2 Construction de la partie d'apprentissage

On suppose disposer de k segments annotés $\{[a_\ell, b_\ell]\}_{\ell=1}^{k}$ (en secondes) dans un audio d'index i. On convertit ces temps en indices de trames $(\alpha_\ell, \beta_\ell)$ et on extrait ces segments :

$$V_{\text{tar}} \;=\; \big[V_{:,\alpha_1:\beta_1}, \ldots, V_{:,\alpha_k:\beta_k}\big] \;\in\; \mathbb{R}_+^{F\times T_{\text{tar}}}. \tag{4.1}$$

On définit un masque temporel $m \in \{0,1\}^T$ qui met à 0 les trames utilisées pour V_{tar} et toutes les trames à partir de la fin du k-ième segment (pour réserver une partie test). Le fond est alors

$$V_{\text{bg}} \;=\; V_{:,\{t:\ m_t=1\}} \;\in\; \mathbb{R}_+^{F\times T_{\text{bg}}}. \tag{4.2}$$

On obtient W_{tar} et W_{bg} par factorisation respective de V_{tar} et V_{bg} :

$$\min_{W_{\text{tar}}\ge 0,\, H_{\text{tar}}\ge 0} D_\beta\big(V_{\text{tar}} \,\|\, W_{\text{tar}}H_{\text{tar}}\big), \qquad W_{\text{tar}} \in \mathbb{R}_+^{F\times K_{\text{tar}}}. \tag{4.3}$$

$$\min_{W_{\text{bg}}\ge 0,\, H_{\text{bg}}\ge 0} D_\beta\big(V_{\text{bg}} \,\|\, W_{\text{bg}}H_{\text{bg}}\big), \qquad W_{\text{bg}} \in \mathbb{R}_+^{F\times K_{\text{bg}}}. \tag{4.4}$$

On forme ensuite la base fréquentielle de la partie d'entraînement :

$$W_{\text{train}} \;=\; \big[\,W_{\text{tar}} \;\; W_{\text{bg}}\,\big].$$

Algorithm 3: Apprentissage des dictionnaires W_{tar} et W_{bg}

Entrée: V, segments $\{[a_\ell, b_\ell]\}_{\ell=1}^{k}$, K_{tar}, K_{bg}, divergence β
Sortie: $W_{\text{train}} = \big[W_{\text{tar}}, W_{\text{bg}}\big]$, indice γ

1 Convertir $\{[a_\ell, b_\ell]\} \to \{[\alpha_\ell, \beta_\ell]\}$ (temps $\to$ trame)
2 Construire $V_{\text{tar}} = \text{Concat}_\ell\, V_{:,\alpha_\ell:\beta_\ell}$ et un masque m mettant ces trames à 0
3 Poser $\gamma = \beta_k$ et annuler dans m toutes les trames $t \ge \gamma$(partie test)
4 Construire $V_{\text{bg}} = V_{:,\{t:\ m_t=1\}}$
5 Résoudre (4.3) pour obtenir W_{tar}
6 Résoudre (4.4) pour obtenir W_{bg}
7 **return** $W_{\text{train}} = \big[W_{\text{tar}}, W_{\text{bg}}\big]$, γ

4.3 Inférence à dictionnaire fixé

On scinde V en une partie test $V_{\text{test}} = V_{:,\gamma:T}$, où $\gamma = \beta_k$ est l'indice de fin du k-ième segment utilisé pour l'apprentissage. On estime uniquement H en gardant W fixé(on n'optimise pas le dictionnaire fréquentiel) :

$$\min_{H\ge 0} D_\beta\big(V_{\text{test}} \,\|\, W_{\text{train}}\, H\big). \tag{4.5}$$

Pour l'étape finale, on reconstruit une matrice complète des activations $H_{\text{full}} \in \mathbb{R}_+^{K\times T}$ en complétant par des zéros la partie des activations ayant servi durant l'entraînement :

$$H_{\text{full}}[:, 1:\gamma-1] = 0, \qquad H_{\text{full}}[:, \gamma:T] = H.$$

4.4 Détection à partir des activations cibles

On applique ensuite une détection de ruptures (PELT, §3.3) sur H_{full} (composante par composante), puis un filtrage par quantile des aires segmentaires (§3.3) afin de ne conserver que les segments "forts" comme dans le cas du NMF standard. L'évaluation s'effectue au niveau segment via l'IoU (§3.4).

Algorithm 4: Inférence à W fixé et détection

Entrée: V, W_{train}, γ, divergence β
Sortie: H_{full}, segments détectés $\widehat{\mathcal{S}}$

$V_{\text{test}} \leftarrow V_{:,\gamma:T}$
Estimer H en résolvant (4.5) avec $W \leftarrow W_{\text{train}}$ fixé (mises à jour de H uniquement)
Construire H_{full} en préfixant des zéros sur $[1, \gamma-1]$ et en posant $H_{\text{full}}(:, \gamma : T) \leftarrow H$
Détecter les ruptures (PELT) sur $H_{\text{full}}(i,)$ et filtrer par quantile des aires segmentaires (cf. §3.3)
return H_{full}, $\widehat{\mathcal{S}}$

4.5 Résultats

Le tableau 4.1 montre un rappel élevé (jusqu'à 0,74 à Greenwich 2015) mais une précision faible (entre 0,07 et 0,16), conduisant à des F1 modestes (0,13–0,25).

Ces résultats restent proches de ceux obtenus avec la NMF standard (cf. tableau 3.4). Quelques facteurs peuvent expliquer ces résultats plutôt limités, dont :

1. $K_{\text{tar}} = 20$ peut sur-factoriser l'événement cible, tandis que $K_{\text{bg}} = 60$ peut être insuffisant pour absorber la variété du fond sur certains jeux.

2. Lors du transfert du dictionnaire partiel, $k = 5$ segments peuvent être *trop homogènes* : la base W_{tar} manque de diversité en label de chants et s'active sur des événements ressemblants.

L'effet de k_{shots} n'est pas linéaire à l'échelle d'un même audio : la *diversité* des segments annotés compte davantage que leur nombre. Sur le fichier `2014-01-13T13-00-00_000.wav` du jeu *Elephant Island 2014*, on observe :

— $k = 2 \Rightarrow$ F1 $= 0{,}31$,
— $k = 3 \Rightarrow$ F1 $= 0{,}29$,
— $k = 5 \Rightarrow$ F1 $= 0{,}33$.

L'ajout d'un 3e *shot* (2 `bpd` et 1 `bma`) a dégradé la détection, tandis qu'augmenter k_{shots} à 5 avec des segments plus *divers* (2 `bma`, 3 `bpd`) améliore le F1. Notons qu'il y a au total 2 `bma` et 10 `bpd` dans l'audio, ce qui souligne l'importance de couvrir la variabilité inter-classes.

TABLE 4.1 – Résultats des détections par NMFewShot

Jeux de données	Précision	Rappel	F1
Elephant Island 2014	0, 16	0, 71	0, 248
Elephant Island 2013	0, 13	0, 45	0, 18
Greenwich 2015	0, 07	0, 741	0, 129

Les paramètres utilisés pour cette partie sont dans le tableau ci-dessous.

TABLE 4.2 – Paramètres utilisés pour l'approche NMF few-shot

Réglage	Valeur	**Remarques**
Apprentissage		
k_{shots}	5	Nombre d'annotations
NMF		
$K_{\text{tar}}, K_{\text{bg}}$	$\{20, 60\}$	Rangs NMF (cibles / autres)
β	0	Divergence Itakura–Saito
Algorithme	MU	Mises à jour multiplicatives
Itérations max.	1000	critère d'arrêt : tol $= 10^{-15}$
Détection		
Pénalité PELT	0	-
Quantile détection	90%	-

En résumé, la NMFewShot capte bien des événements, et les faux positifs peuvent être *substantiellement* réduits en jouant avec les paramètres. Le prochain chapitre abordera la séparation des sources.

Chapitre 5

Séparation de sources par NMF

5.1 Principe

L'objectif de ce chapitre est de présenter un outil simple, qui s'inspire en partie de [4], basé sur la NMF, permettant de séparer les composantes biologiques (chants, etc.) des autres sons présents dans un audio (navires, vent, etc.).

Le principe repose sur la décomposition du spectrogramme de l'audio en une somme de composantes élémentaires. Certaines de ces composantes peuvent être interprétées comme d'origine biologique, tandis que d'autres sont associées aux bruits ou aux sons anthropiques.

5.2 Formulation mathématique

Soit un audio $y(t)$, de fréquence d'échantillonnage sr. Sa transformée de Fourier à court terme (STFT) a déjà été introduite en §3.1.1. On note $V \in \mathbb{R}_+^{F \times T}$ le spectrogramme associé à $y(t)$.

La NMF, aussi introduite en §3.2, approxime V par l'équation suivante :

$$V_{f,t} \approx W_{f,1}\, H_{1,t} + W_{f,2}\, H_{2,t} + \ldots + W_{f,K}\, H_{K,t}. \tag{5.1}$$

Pour chaque composante k, on définit un filtre de masquage de Wiener :

$$M_{f,t}(k) = \frac{W_{f,k} H_{k,t}}{\sum_{j=1}^{K} W_{f,j} H_{j,t}}. \tag{5.2}$$

Le spectrogramme associé à la composante est alors :

$$S_{f,t}(k) = M_{f,t}(k)\, S_{f,t}, \tag{5.3}$$

Pour chaque spectrogramme reconstruit $\tilde{S}_k$, on extrait quelques descripteurs spectraux (centroïde, énergie, etc.) sur la bande fréquentielle du source ciblée, puis on les agrège en un vecteur ϕ_k. Un score heuristique s_k est ensuite calculé à partir de ϕ_k, par combinaison linéaire, afin d'estimer si la composante est d'origine biologique.

Enfin, l'audio est reconstruit par transformée de Fourier inverse :

$$y_k(t) = \mathrm{iSTFT}\big(S_k\big). \tag{5.4}$$

5.3 Algorithme

Algorithm 5: Séparation de sources par NMF

Entrée: fichier audio, rang K
Sortie: audios séparés $y_k(t)|_{k=1}^{K}$ et scores biologiques $s_k|_{k=1}^{K}$

Charger $y(t)$
Calculer $S \leftarrow \text{STFT}(y)$, $V \leftarrow |S|$
Calculer une NMF $V \approx WH$ // $W \in \mathbb{R}_+^{F\times K}$, $H \in \mathbb{R}_+^{K\times T}$
for $k \leftarrow 1$ **to** K **do**
    $V_k \leftarrow W_{.,k}\, H_{k,.}$
    $M_k \leftarrow \dfrac{V_k}{\sum_{j=1}^{K} V_j}$ // masque de Wiener
    $\tilde{S}_k \leftarrow M_k \odot S$
    Extraire des caractéristiques de $\tilde{S}_k$ (énergie, etc.)
    Calculer un score heuristique s_k (biologique vs autre)
    $y_k(t) \leftarrow \text{iSTFT}(\tilde{S}_k)$
Save/visualiser y_k

5.4 Résultats qualitatifs

Pour valider l'approche, nous avons simulé un mélange simple composé de :
— un chant de baleine bleu
— des sons de navire provenant de [12],
— des bruits diffus de type vent sous-marin [12].
Après application de la NMF avec $K = 6$:
— une composante est clairement associée au chant de baleine,
— plusieurs composantes représentent les bruits,
La séparation est presque parfaite auditivement. Cependant, une validation quantitative reste nécessaire, en calculant par exemple des métriques telles que le rapport signal sur bruit (SNR) ou le Scale-Invariant Signal-to-Distortion Ratio (SI-SDR).

Chapitre 6

Perspectives

Ce travail a permis de mettre en évidence un certain intérêt de la NMF et de ses variantes (séquentielle, few-shot) pour l'analyse bioacoustique en milieu marin. Plusieurs pistes d'amélioration et d'extension peuvent être envisagées.

6.1 Filtrage adapté à l'espèce étudiée

Une première piste consiste à ajouter une étape de pré-filtrage, spécifiquement adaptée à l'espèce étudiée. Par exemple, on peut restreindre la bande de fréquences analysée à celle correspondant aux appels connus de cette espèce, ce qui pourrait améliorer la séparation dans la NMF en éliminant une partie du bruit hors bande. De plus, une adaptation automatique des seuils de détection en fonction du contexte (selon l'heure, avec activité diurne/nocturne, ou selon les conditions environnementales) permettrait de réduire les erreurs de détection.

6.2 Algorithmes de détection

Il serait intéressant d'affiner les algorithmes de détection de ruptures. On pourrait tester d'autres méthodes de détection (par exemple en utilisant des chaînes de Markov cachées pour la détection de changements) afin de mieux repérer le début et la fin des chants. De plus, on peut intégrer des contraintes dans la NMF, par exemple en imposant de la parcimonie sur la matrice d'activations H afin de favoriser des solutions plus sparses.

6.3 Équilibre des annotations en apprentissage few-shot

Dans le cas de l'apprentissage avec peu d'exemples (approche few-shot), il est fondamental que les différents labels de chants soient représentés de manière équilibrée dans la partie entraînement. Or, certains jeux de données présentent un fort déséquilibre entre les types de chants annotés(cf. §2), ce qui limite la capacité du dictionnaire appris (W_{train}) à généraliser aux nouvelles données. Une piste d'amélioration consiste à utiliser des techniques de rééchantillonnage afin de mieux couvrir la variété des chants présents. Ainsi, la NMFew-shot pourrait apprendre un dictionnaire représentatif de l'ensemble des vocalisations d'intérêt.

6.4 NMF few-shot pour la séparation de sources

Une autre perspective porte sur l'utilisation du dictionnaire appris via NMF few-shot dans un cadre de séparation de sources. En particulier, si les k-exemples fournis à l'entraînement correspondent à des extraits de chants de baleines, la factorisation obtenue $W_{\text{tg}}H_{\text{tg}}$ peut être interprétée comme représentant spécifiquement la partie biophonique du signal (c'est-à-dire les chants de baleines appris). Dans cette optique, on pourrait ensuite utiliser ce dictionnaire pour isoler les composantes correspondantes aux chants dans des audios mixtes contenant du bruit et d'autres sources sonores. Cette approche ouvrirait la voie à une détection et une séparation simultanées, en utilisant la même approche de NMF pour à la fois extraire les chants ciblés et rejeter les autres composantes sonores.

6.5 Comparaison avec des approches de l'état de l'art

Enfin, il serait intéressant de comparer les performances des méthodes abordées dans ce travail avec celles d'approches plus récentes issues de l'état de l'art du traitement du signal n'impliquant pas les réseaux neuronaux. Évaluer la NMF (et ses variantes) face à ces approches alternatives permettrait de situer sa performance. Cette comparaison donnerait un cadre objectif pour mesurer les gains apportés par la NMF par rapport aux autres méthodes conventionnelles.

En résumé, ces perspectives couvrent autant des améliorations techniques (prétraitements adaptés, gestion du déséquilibre) que des extensions des méthodes vers de nouvelles finalités scientifiques (séparation des sources). Elles ouvrent la voie à des applications plus fiables pour le suivi acoustique des écosystèmes marins.

Conclusions

Ce travail a proposé et évalué un pipeline complet pour la détection d'événements sonores, appliqué à des audios de baleines. En s'appuyant sur des représentations temps-fréquence (STFT, LTSA) et la factorisation en matrices non négatives (NMF), nous avons montré qu'il est possible d'extraire des motifs spectro-temporels interprétables et de suivre leurs activations temporelles. L'algorithme PELT a ensuite permis de détecter des ruptures correspondant à l'apparition et à la disparition de vocalisations.

Deux extensions majeures ont été exploitées. D'une part, une approche inspirée de few-shot learning, où des bases de la NMF sont initialisées avec des exemples connus, permettant d'orienter la détection même en présence de peu de données annotées. D'autre part, l'adaptation de la NMF à la séparation de sources, en combinant la factorisation avec un filtre de Wiener, a montré la faisabilité d'isoler les sons biologiques des bruits.

Dans l'ensemble, ce rapport met en évidence l'intérêt de méthodes sobres et interprétables, qui permettent à la fois la détection et la séparation de vocalisations. Ces résultats ouvrent plusieurs perspectives : régulariser la NMF, tester d'autres algorithmes de détection de ruptures.

Annexes

Pipeline de la détection basée sur la NMF standard

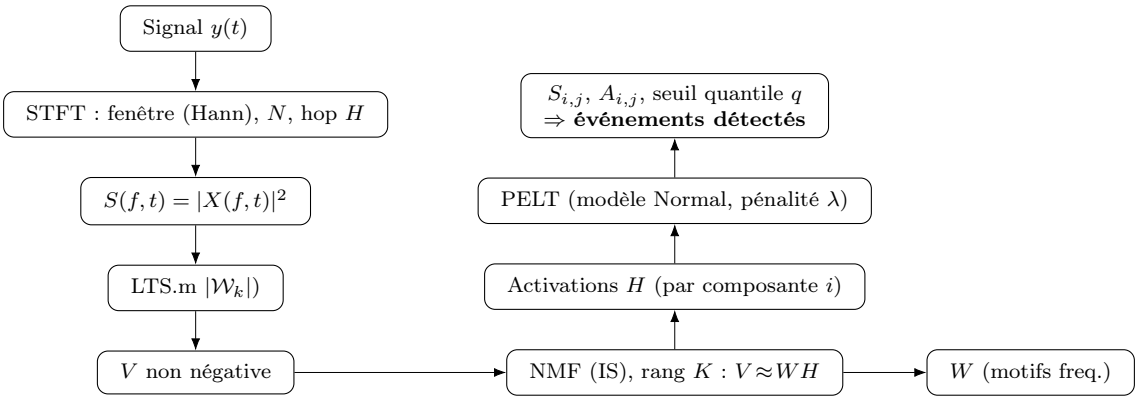
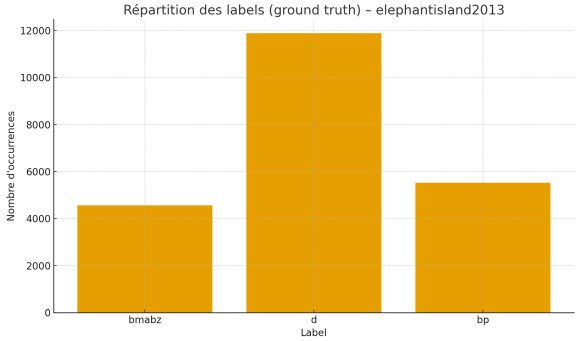
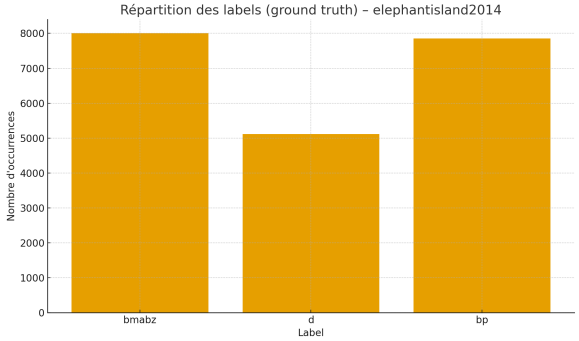


FIGURE 1 – Pipeline de la détection NMF-based

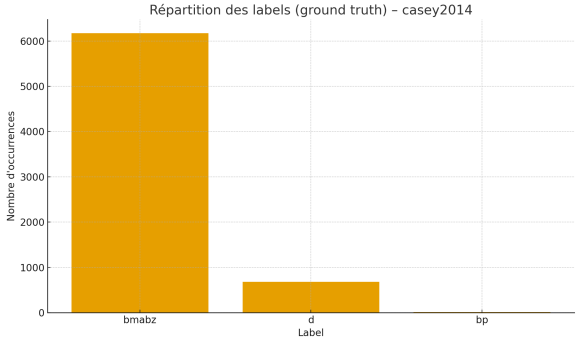
Distribution des labels selon les jeux de données



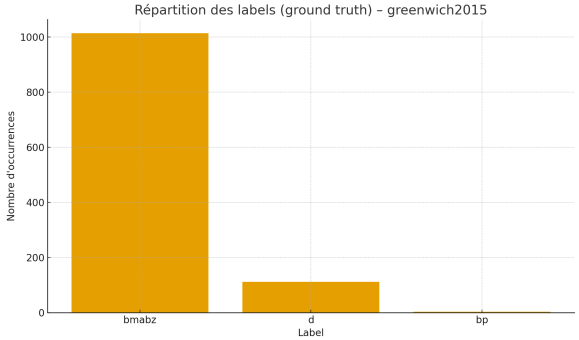
(a) Elephant Island 2013



(b) Elephant Island 2014



(c) Casey 2014



(d) Greenwich 2015

FIGURE 2 – Distribution des labels dans les différents jeux de données

Conversion des segments

Secondes $\leftrightarrow$ trames **STFT** (sr, hop H) :

$$\alpha = \left\lfloor \frac{a\,sr}{H} \right\rfloor, \quad \beta = \left\lceil \frac{b\,sr}{H} \right\rceil, \quad t \;\mapsto\; \frac{t\,H}{sr}\ \text{(s)}.$$

LTSA (pas Δt centré) :

$$\alpha = \left\lfloor \frac{a - \Delta t/2}{\Delta t} \right\rfloor, \quad \beta = \left\lceil \frac{b - \Delta t/2}{\Delta t} \right\rceil, \quad t \ \mapsto \ t\,\Delta t + \frac{\Delta t}{2},$$

où $\Delta t = |\mathcal{W}_k|$

Les méthodes abordées dans ce travail sont accessibles dans ce dépôt : https://github.com/mahamat9/internship_methods.

Bibliographie

[1] Few-shot bioacoustic event detection. Challenge DCASE 2024, Task 5.

[2] Supervised detection of strongly-labelled whale calls, 2025. Challenge DCASE 2025, Task 2.

[3] Léa Bouffaut, Richard Dréo, Valérie Labat, Abdel-O. Boudraa, and Guilhem Barruol. Passive stochastic matched filter for antarctic blue whale call detection. *Journal of the Acoustical Society of America*, 144(2) :955–965, 2018.

[4] Nafissa Dia, Julie Fontecave-Jallon, Pierre-Yves Guméry, and Bertrand Rivet. Application de la factorisation non-négative des matrices (nmf) pour le débruitage des signaux phonocardiographiques. In *GRETSI 2017 - XXVIème Colloque francophone de traitement du signal et des images*, September 2017.

[5] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura–saito divergence : With application to music analysis. *Neural Computation*, 21(3) :793–830, 2009.

[6] Cédric Févotte, Emmanuel Vincent, and Alexey Ozerov. Single-channel audio source separation with nmf : divergences, constraints and algorithms. *IEEE Signal Processing Magazine*, 31(3) :72–82, 2014.

[7] Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500) :1590–1598, 2012.

[8] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791, oct 1999.

[9] Tzu-Hao Lin, Shih-Hua Fang, and Yu Tsao. Improving biodiversity assessment via unsupervised separation of biological sounds from long-duration recordings. *Scientific Reports*, 7 :4547, 2017.

[10] D. K. Mellinger. A comparison of methods for detecting right whale calls. *Canadian Acoustics*, 32(2) :55–65, 2004.

[11] Brian S. Miller, The IWC-SORP/SOOS Acoustic Trends Working Group, Naysa Balcazar, Sharon Nieukirk, Emmanuelle C. Leroy, Meghan Aulich, Fannie W. Shabangu, Robert P. Dziak, Won Sang Lee, Jong Kuk Hong, and Danielle Harris. An open access dataset for developing automated detectors of antarctic baleen whale sounds and performance evaluation of two commonly used detectors. *Scientific Reports*, 11(1) :806, 2021.

[12] NOAA SanctSound. Sanctsound data portal - ship noise sample. https://sanctsound.portal.axds.co/#sanctsound/compare/prop-model.

[13] Charles Truong, Laurent Oudre, and Nicolas Vayatis. selective review of change point detection methods with a level-set perspective and the ruptures library. *Journal of Machine Learning Research*, 21(200) :1–57, 2020.