Classification de vocalises de mammifères marins en environnement sismique

Thomas Guilment

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

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

THESE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, image, vision*

Par
Thomas GUILMENT

Classification de vocalises de mammifères marins en environnement sismique

Thèse présentée et soutenue à Brest, le 21 juin 2018
Unité de recherche :
Thèse N° : 2018IMTA0080

Composition du Jury :

Président :	Jean-Yves ROYER	Directeur de recherche CNRS, UBO-IUEM
Rapporteurs :	Olivier ADAM	Professeur, Institut d'Alembert UMR 7190
	Jérôme MARS	Professeur, GIPSA-LAB Grenoble INP
Examinateurs :	Barbara NICOLAS	Chargée de recherche CNRS, CREATIS
	Jean-Yves ROYER	Directeur de recherche CNRS, UBO-IUEM
Directeur de thèse :	Dominique PASTOR	Professeur, IMT Atlantique, Campus de Brest
Encadrant :	François-Xavier SOCHELEAU	Maître de conférences, IMT Atlantique, Campus de Brest
Invité		
Simon VALLEZ	Docteur, ingénieur de recherche, SERCEL Brest	

N° d'ordre : 2018IMTA0080

Sous le sceau de l'Université Bretagne Loire

IMT Atlantique
Bretagne-Pays de la Loire

En accréditation conjointe avec l'Ecole Doctorale MathSTIC

CLASSIFICATION DE VOCALISES DE MAMMIFÈRES MARINS EN ENVIRONNEMENT SISMIQUE

Thèse de Doctorat

Mention : Signal, Image, Vision

Présentée par **Thomas Guilment**

Département : Signal et Communications

Laboratoire : Lab-STICC Pôle : CID

Directeur de thèse : Dominique PASTOR
Encadrant : François-Xavier SOCHELEAU

Soutenue le 21 juin 2018

Jury :

M. Olivier Adam – Professeur, Institut d'Alembert UMR 7190 (Rapporteur)
M. Jérôme Mars – Professeur, GIPSA-LAB Grenoble INP (Rapporteur)
M. Dominique Pastor – Professeur, IMT Atlantique Bretagne-Pays de la Loire, Campus de Brest (Directeur de thèse)
M. François-Xavier Socheleau – Maître de conférences, IMT Atlantique Bretagne-Pays de la Loire, Campus de Brest (Encadrant)
Mme Barbara Nicolas – Chargée de recherche CNRS, CREATIS (Examinateur)
M. Jean-Yves Royer – Directeur de recherche CNRS, UBO-IUEM (Examinateur)
M. Simon Vallez – Docteur, ingénieur de recherche, SERCEL Brest (Invité)

« *Le monde ne sera heureux que quand tous les hommes auront des âmes d'artistes,*
c'est-à-dire quand tous prendront plaisir à leur tâche. »
— Extrait du testament d'Auguste Rodin, 1911

Cette thèse a été préparée à Télécom-Bretagne au sein du département Signal et Communications (UMR CNRS 3192 Lab-STICC) en partenariat avec l'entreprise SERCEL (antenne de Brest) du 20 avril 2015 au 19 avril 2018. Le Laboratoire du Domaine Océanique a participé à ce travail en fournissant des données réelles exploitées dans ce manuscrit.

Remerciements

Je me souviendrai toute ma vie de mon entretien pour la thèse. Je partais de Lannion pour aller à Brest, en plein mois de janvier, la pluie, la nuit, le froid et la route, cette unique N12 me séparant de mon lieu de rendez-vous qui ce jour-là était barrée... par des échalotes ! Je suis finalement arrivée à passer l'entretien et après ces 3 années d'aventure, me voici en train de penser à toutes ces personnes qu'alors je ne connaissais pas et que j'ai eu la chance de rencontrer.

Avant toutes choses, je tiens à remercier vivement les personnes qui m'ont fait confiance et qui m'ont accompagné pour ces 3 ans de recherche : **François-Xavier Socheleau**, **Dominique Pastor** de l'IMT et **Simon**, **Laurent**, **Christophe** et **Stéphane** de Sercel.

Dominique, cher directeur, je suis heureux d'avoir pu travailler avec un personnage tel que toi. Je dis bien un personnage car tu es un monsieur du Sud venu travailler en Bretagne qui déteste la pluie ! Mathématicien, musicien et karatéka, merci de m'avoir transmis ta « philosophie de la recherche ». C'était un super sujet de travailler sur les méthodes de reconnaissance et c'est grâce à toi.

Un merci très spécial à toi, **François-Xavier**, cher encadrant qui m'a toujours dit ne pas exister officiellement. Pour moi, tu as été et tu es bien réel ! Tu as su justement m'encadrer, me cadrer et m'apprendre à avancer dans la recherche en toute humilité (un caillou de plus sur la montagne). Je penserai à toi dès que je prendrai mon papier et mon crayon (un porte-mine de préférence) pour faire mes futures recherches. Je sais qu'avec quelques corrections mineures on peut obtenir de beaux résultats. Merci pour m'avoir toujours donné confiance en moi. Tu es quelqu'un de vraiment inspirant et un jour j'espère bien venir t'écouter jouer un peu de guitare.

J'ai bien sûr une pensée pour l'entreprise Sercel de Brest. Tout d'abord à **Laurent** avec qui le projet a commencé. Tu as toujours été bienveillant. Merci ensuite à **Christophe**. Tu m'auras enseigné beaucoup sur les exigences d'une entreprise. Je pense notamment aux spécifications (et à l'ensemble des tests qui vont avec !) pour concrétiser le cahier des charges. Je remercie **Stéphane** pour sa participation aux réunions en partageant son vécu sur les situations rencontrées sur le terrain (avec ta tasse Pamguard à la main). Aussi, je remercie beaucoup **Simon** qui m'a toujours fait confiance. Merci d'avoir toujours pris le temps pour moi et de m'avoir bien accueilli au sein de Sercel. J'ai passé de bons moments dans l'entreprise et nous avons pu bien partager. Bien sûr, il y eu les rapports techniques, les codes, les prototypes de Sercel mais aussi la culture cinématographique et puis parler de la vie tout simplement.

Ensuite, je remercie vivement **Jean-Yves Royer** du LDO pour m'avoir fourni les données que j'ai pu exploiter tout au long de ma thèse. J'en profite pour remercier les personnes de l'ENSTA qui m'ont bien aidé. Merci à toi **Flore** pour les cours de biologie sur les baleines et pour avoir toujours été bienveillante avec moi. Merci beaucoup à toi **Manue** pour ton cours sur l'utilisation d'XBAT, pour toutes les informations à savoir sur les données et pour ta générosité.

A présent, voici le moment d'exprimer mes plus sincères remerciements à toutes ces personnes que j'ai pu côtoyer durant ces trois années. Pour commencer, je remercie **Malek**, mon collègue de bureau qui était déjà en 3^{e} année à mon arrivée. Je me suis senti à l'aise tout de suite,

merci d'avoir été présent pour moi (pour installer l'imprimante ou pour savoir qui contacter), pour ta bonne humeur permanente (toujours bien présente!), ton humour sans oublier nos parties de basket (en vraie comme dans le bureau!). Dans ce même bureau, je pense bien à toi, mon ami indonésien, mon cher **Budhi** pour toutes ces journées où on était simplement heureux de se retrouver et de rire sur les scènes extraordinaires des films de combats (16 secondes!). Enfin, merci à toi **Kiên**, mon voisin de bureau (le plus proche), pour ton esprit farceur et pour ces parties de badminton où tu m'as toujours mis des raclées avec le sourire et en jouant en jean (c'est peut-être ça le secret)!

J'aimerais d'ailleurs exprimer un merci particulier aux vietnamiens pour votre accueil dès mon arrivée. Merci pour toutes ces soirées, ces anniversaires, ces barbecues, pour les matchs de badminton et pour m'avoir partagé votre langue (bon, je ne sais toujours pas prononcer « ma » mais ça viendra), votre culture (c'est un autre monde!) et aussi votre gastronomie (les rouleaux de printemps c'est super bon). Merci **Khan**, qu'est-ce qu'on s'est bien amusé avec nos discussions sur le français, sur les français et merci aussi à ta femme **Chan** pour m'avoir appris à faire les nems. **Thanh**, ce fut toujours un grand plaisir que d'aller te retrouver dans ton bureau à l'incubateur. Tu as toujours su trouver les mots pour me motiver et me donner un beau reflet de ma vie, vraiment **MERCI**! Je pense aussi à ta femme, **Huong**, vous êtes des parents formidables. Cher **Phuong** (je ne t'appelerai pas Alex), toi le vietnamien français, je te remercie beaucoup pour ton aide à la fin de la thèse et pour tous ces repas partagés les soirs de travail au RAK. A cette même table venait manger **Duong**, mon élève qui est devenu mon collègue doctorant et surtout mon ami. Sache que tu n'es pas obligé de continuer de m'appeler « Professeur Thomas » (même si ça me fait toujours très plaisir). **Trang Hui** et **Loc**, merci aussi d'avoir été présents. Pour tous les autres, pardonnez-moi d'avance si j'orthographie mal vos noms (pour certains d'entre vous je ne les ai jamais vu écris), de même si j'en oublie, je vous remercie énormément pour tous ces moments passés ensembles : **Minh Tan**, **Chi Hieu**, **Anh**, **Tran Vu**, **Kinh**, **Uyen**, **Nhat Quang**, **Thuy**, **Thanh Nhan**, **Hien**, **Thoai My**, **Son**, **Minh**, les **Nam** et **Phi**.

Un merci particulier à **Aurélien** et **Luis**, du bureau d'à côté. Mon ami **Luis**, merci pour tout ce que tu m'as apporté, tes magnifiques photos, tes partages sur le deep learning, tes astuces linux, ton amour du jazz que tu chantes très bien, ta guitare, ton Thursday cake au labo où tu nous ramenais de bons gâteaux, ta sauce moutarde-miel, tes soirées pizzas où « il faut un peu de piment quand même car sinon ça n'a pas de goût » (sans parler de tes bonbons au piment et de ton cocktail mexicain bien épicé de ton pot de départ que seul **Khanh** appréciait avec le sourire). **Aurélien**, merci pour ton caractère bien trempé qui nous aura toujours fait sourire lors de nos parties de jeux de société.

A propos de jeu de société, je pense forcément à **Sébastien**. Merci pour tout ce que tu m'auras fait découvrir du monde des jeux de société (la peinture de figurines, les jeux de rôles, les festivals, les sites d'occasion), pour tous nos duels épiques de dungeon twister, pour toutes tes aventures (le geocaching, l'escalade, le canoé, les voyages en campingcar...) et merci pour avoir été bien présent pendant ces trois années! J'en profite pour remercier tous les joueurs, **Jean-Marie** pour tes jeux de développement, **Vincent** pour tes jeux de négociation et aussi pour avoir pris le temps de me partager tes conseils pour l'après-thèse, **Caroline** avec tes jeux familiaux, **Marie-Dominique** et **Sébastien** pour être venu jouer, et **Nicolas**, toi mon voisin doctorant d'en face et surtout mon ami avec qui nous avons désamorcé des bombes mais pas seulement. Tu as toujours été disponible pour moi et pour me partager généreusement tous tes conseils précieux, que ce soit pour le travail avec tes outils (tu es devenu le responsable non officiel des lames de calcul), tes programmes (les objets en matlab, lateX et TikZ), tes applications (Tortoise que je compte bien utiliser), les actualités scientifiques (les batteries diamants) ou pour profiter de la Bretagne avec tes bonnes adresses (merci pour le cocktail d'Han Solo en avant-première) et des sorties insolites (c'était la première fois que je mangeais africain!). J'ai adoré partager tes histoires de Krav-maga avec tes

entraînements de ouf et ton instructeur qui vous explique qu'on peut « découper efficacement » une personne même avec une toute petite lame. Bon, il faudra quand même penser à se calmer.

Avec **Nico**, nous avons aussi fait de la musique ensemble avec le groupe des musiciens du département que je ne saurais nommé (« 35 min d'impro » ? « Poupou et ses Boréliens »…?). En tout cas, je remercie tous les musiciens. J'ai toujours pu compter sur vous pendant la thèse et c'était génial de jouer ensemble. Quel bonheur de se défouler le temps d'une soirée ou d'un midi en improvisant de la musique en fonction du moment. Merci **Anto** pour ta musique funk modern free reggae blues folk rock jazz endiablée et tes solos de fou à la gratte ! Toi qui m'as toujours dit ne pas savoir trop lire les notes et juste te débrouiller, sache que j'aimerais beaucoup savoir me débrouiller comme toi. Je repense à tes discussions philosophiques et politiques, les soirées chez toi et à nos concerts dont un qui nous a même mené jusqu'à jouer de la « musique liquide » ! Faut le faire ! Merci **Manuel** pour ta présence toujours très chaleureuse, ton énergie débordante à la batterie, ta sublime barbe et tout ce vocabulaire espagnol très fleuri ! Merci **Aurel** pour ton groove, tes walking bass, pour la complicité que j'ai pu avoir avec toi lorsqu'on je jouais des contrechants sur ta partie (« I shot the sheriff » ♫ ♩) mais aussi pour la complicité au quotidien. **Stéphanie** (notre super chanteuse !) et toi m'avez hébergé et c'était génial (les origamis, la guitare, les échecs et skyrim), vraiment merci beaucoup !

Et puis il y a tous les autres doctorants que j'ai eu la chance de rencontrer. Merci à **Safa** pour m'avoir tout de suite impliqué dans la vie du laboratoire dès mon arrivée en me parlant de B-doc, d'IEEE et de ce qu'il était possible de réaliser en tant que doctorant comme la mobilité. J'ai une pensée particulière pour **Redouane** qui a toujours été très actif au laboratoire. Tu as organisé la journée des doctorants quasiment seul (avec l'aide de **Monique** quand même !), tu nous as montré comment créer une page personnelle et tu as toujours pris le temps de partager ton travail. Je me souviendrais de ta philosophie sur les deadlines. Un merci particulier à mon autre ami **Nicolas**, grâce à toi j'ai passé mon baptême de plongée en Bretagne ! Merci surtout pour avoir été d'un grand soutien, tu m'as à chaque fois reçu avec bonne humeur dans ton bureau et tu m'as toujours fait croire que j'étais capable de faire une bonne thèse.

En trois ans, mes collègues de bureaux ont changé et j'ai eu le plaisir de vous avoir avec moi : **Sabrina**, **Nacceredine** et **Paul**. **Sabrina**, la reine du thé, des tisanes et autres infusions, merci pour ta générosité et pour m'avoir partager la culture du Maroc, un jour je compte bien aller visiter. C'était sympa d'avoir les mêmes encadrants et d'être dans la même équipe de recherche ! **Paul**, même si ce fut assez bref, j'ai été vraiment heureux d'impoviser quelques notes avec toi au piano un midi et de discuter des vagues scélérates ! Enfin, un merci très particulier à toi **Nacceredine**, nous nous sommes bien entendu très rapidement et ça ne s'est pas arrêté. Merci de m'avoir partagé ton enfance en Algérie (les jeux vidéos d'enfances et comment ça se passait pour le « sport » à l'école). Merci pour nos conversations quotidiennes, toutes ces histoires qui font voyager (manger des pains aux chocolats à 2h du matin !). Je pense également à nos sorties en dehors de la thèse, merci à **Amina** et toi pour le cinéma, le cirque et pour tous ces moments passés ensemble.

Pour leur bonne humeur et au plaisir de se revoir, je remercie **Zahran**, **Marza**, **Carole**, **Rémi**, **Yassine**, **Guillaume B.**, **Guillaume A.**, **Mohamed**, **Fangping**, **Erwan**, **Zakaria**, **Alma**, **Mathieu**, **Said**, **Morgane**, **Romain**, **Ewen** et **Danielle**.

Bien sûr, la thèse n'aurait pas été la même sans tous les autres membres du département SC avec qui tout s'est toujours très bien passé. Tout d'abord, **Monique**, notre gestionnaire du labo, merci de m'avoir accompagner pendant ces trois ans, que ce soit dans toutes les démarches administratives, les cagnottes des cadeaux pour les pots de départs et les projets des doctorants. Tu as toujours été une personne sur qui j'ai pu compter et tu vas pouvoir prendre une retraite bien méritée. J'ai une pensée pour **Samir**, notre chef de département, merci d'avoir été à l'écoute et de

ne nous avoir jamais empêché de réaliser les projets que nous voulions faire. Merci notamment pour m'avoir toujours encouragé lorsque j'étais responsable des doctorants, d'être venu intervenir auprès des doctorants et pour avoir pris le temps pour moi à la fin de la thèse pour me partager ton expérience. J'ai une pensé pour **Pierre** et son beau papillon de Lorentz, j'ai eu le grand plaisir de travailler avec toi sur les TPs de dernière année. Merci d'avoir toujours pris le temps de me partager tes recherches et de m'expliquer ce que je connaissais/comprenais pas. Merci aussi à **Thierry C.** pour notre travaille ensemble et pour m'avoir fait découvrir ipython. **Christophe**, merci pour ton humour, ton point de vue très pragmatique qui m'a toujours ouvert les yeux et merci d'avoir accepté d'être mon chauffeur le temps d'une journée quand j'en ai eu besoin. Même si nous ne sommes pas beaucoup parlés pendant ces trois, merci à toi **Raphaël** pour tes conseils. Et pour tous nos échanges, les repas, les rire des pauses de 16h et tant d'autres choses, je remercie **Elsa**, **Zeina**, **Vincent** (un jour j'espère bien qu'on jouera de la musique ensemble), **Djalil** (j'entends toujours ton rire et je penserai bien à tes conseils de voyage!), **Frédéric** (merci pour ta curiosité musicale), **Ronan** (merci pour tes conseils et ton partage en traitement d'image), **Lucas** (ce fut court mais sympa) et **Jean-Marc** (merci pour les conseils sur la raspberry et pour m'avoir toujours aidé quand j'en avais besoin).

Un merci très particulier à toi **Thierry**, notre ingénieur première classe, sache que pour moi tu es bien plus que ça. Je ne sais comment te remercier. Tu as toujours pris le temps de me partager ton savoir-faire, tes expériences personnelles et professionnelles (grâce à toi j'ai tant appris!). J'ai hâte d'appliquer toutes tes bonnes pratiques dans mon futur emploi, merci pour tout!

J'ai également une pensée pour les autres personnes que j'ai rencontrée hors du département SC, merci à **François** pour les TPs d'image et pour son aide à la fin de la thèse. Je remercie **Jean-Philippe** pour la visite du département micro-onde et son humour. Merci à **Aimee** pour ses cours d'anglais, **Nicolas F.** pour la musique, **Laurent** pour nos discussions sur les échecs et la pédagogie. Merci à toi mon ami **Vincent H.** pour notre voyage sur l'île de Sein, notre initiation à l'aïkido, nos discussions et tes partages sur le chiffrement homomorphe. Je remercie aussi **Clément**, notre bibliothécaire, pour ta curiosité scientifique (merci d'être venu à ma soutenance) et pour ta bonne humeur!

Je remercie aussi le personnel de l'administration pour nos partages et avec qui tout s'est toujours bien passé : **Linda**, **Viviane**, **Martine**, **Fabienne** et **Priscilla**.

Après avoir remercié les personnes qui m'ont permis de vivre cette aventure, je remercie ceux qui m'ont permis d'y mettre un point final. Merci à nouveau à **Jean-Yves Royer** pour avoir accepté de présider mon jury, pour s'être intéressé à mon travail et pour avoir pris le temps de corriger ce manuscrit. Merci aussi à **Barbara Nicolas** d'avoir accepté de faire partie de mon jury et d'avoir posé des questions auxquelles on n'est pas obligé de répondre. Un grand merci à mes deux examinateurs : **Jérôme Mars** et **Olivier Adam** pour la lecture de mon manuscrit et pour leurs retours constructifs, d'une part sur le traitement du signal et d'autre part en bioacoustique. **Olivier**, je te remercie également d'être présent après la thèse, de nous avoir partagé comment la baleine à bosse saute hors de l'eau, comment le cachalot « ferme ses oreilles » et surtout, grâce à toi, j'ai la chance de poursuivre la recherche sur les baleines en Louisiane, pour ça un grand **MERCI!**

Je n'aurais jamais pu accomplir tout ce chemin sans le soutien inconditionnel de ma famille : Mon **Papa**, ma **Maman**, **Jacky**, **DD**, **Mathos** et **Clairette**, les **grands-parents**, les **tatas** et **tontons** (Merci **Marie** et **Renzo** pour notre tour en Bretagne, et merci Tonton **Jean** pour tes conseils), les **cousines** et les **cousins** (merci **Pedro** d'être toujours présent!) et à toute la famille. Merci pour vos encouragements, pour votre présence (quelle chance de vous avoir si nombreux à ma soutenance!) et d'avoir toujours cru en moi.

J'exprime un IMMENSE MERCI à ma **Séverine**, pour m'avoir suivi jusqu'en Bretagne, pour m'avoir toujours supporté et pour m'avoir toujours soutenu dans ce projet fou qu'est la thèse. Tu as été pour moi une « maison de lumière » qui m'a toujours aidé à faire le point. Grâce à toi, je sais que tout est possible, vraiment **MERCI!**

A mes amis de toujours, mes deux capitaines **Flore** et **Thomas**, malgré les tempêtes de la thèse (et de la vie!) vous êtes toujours là pour m'aider à garder le cap! **MERCI POUR TOUT** et à bientôt pour de nouvelles aventures!

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

Pour des raisons de lisibilité, le sens d'un acronyme n'est généralement défini qu'à sa première apparition dans le texte. Par ailleurs, nous avons choisi de conserver les acronymes dans leur forme la plus usuelle voici pourquoi il est fréquent que le terme anglais soit employé.

ACP Analyse en Composantes Principales. ix, 36, 41, 45
ADL Analyse Discriminante Lineaire. 36, 41, 45
ADN Acide DésoxyriboNucléique. 6, 10
CFAR *Constant False Alarm Rate*. 58
CGG Compagnie Générale de Géophysique. 6, 22
DDR3 Compression de DDR3 SDRAM définit comme la mémoire à accès direct synchrone à débit de données doublé de troisième génération, ou *Double Data Rate 3rd generation Synchronous Dynamic Random Access Memory*. 68
DSP Densité Spectrale de Puissance. x, 69, 70
EEG Électroencéphalographie. 56
EMD Modes de décomposition empiriques, ou *Empirical Mode Decomposition*. 34, 35
Ifremer L'Institut Français de Recherche pour l'Exploitation de la Mer. 6, 22
IMF *Intrinsic Mode Function*. 34, 35, 41
KNN K plus proche voisins, ou *K-Nearest-Neighbor*. 55
MAP Maximum A Posteriori. 65
OMP Orthogonale Matchgin Pursuit. 56, 68
PAM *Passive Acoustic Monitoring*. 23, 68
RAM Mémoire vive (ou à accès direct), ou *Random Access Memory*. 68
RTSDR Rapport du temps d'exécution sur la durée du signal, ou *Run-Time-to-Signal-Duration Ratio*. x, 68
SINR Rapport signal à bruit plus interférences ou *Signal to Interference plus Noise Ratio*. x, xi, 58, 63, 65, 66, 69, 72–77, 80–86, 90, 95, 104
SINR-SRC *Signal to Interference plus Noise Ratio combined with Sparse Representation-based Classification*. 7, 8, 58, 60, 63, 65, 66, 68, 69, 71, 77, 80, 81, 83–85, 87, 89, 102, 104, 105
SNR *Signal to Noise Ratio*. x, xi, 8, 41, 62, 80, 84, 87–95, 97, 98, 102, 105, 106
SONAR *SOund Navigation And Ranging*. 5, 6
SRC *Sparse Representation-based Classification*. xiii, 56–58, 62, 64, 66, 67, 80, 83, 95
THH Transformée de Hilbert Huang. 34, 35

Liste des notations

Nous avons regroupé ci-dessous les principales notations employées dans les différents chapitres du manuscrit. Nous nous sommes efforcés de conserver des notations cohérentes entre chaque chapitre.

$\mathscr{D}$	Ensemble
$\{(.,.)\}$	Ensemble de paires d'éléments
N	Entier positif
$\hat{(.)}$	Estimation
f	Fonction
δ	Symbole de Kronecker
$\mathbf{D}$	Matrice
$\|.\|_2$	Norme euclidienne
$\|.\|_F$	Norme de Fröbenius
$\|< .,. >\|$	Produit scalaire
$\|.\|_0$	Pseudo-norme ℓ_0
$(.)^T$	Transposée
$\|.\|$	Valeur absolue ou module
$\boldsymbol{w}$	Vecteur

Introduction générale

« *Va prendre tes leçons dans la nature, c'est là qu'est notre futur.* »

— Léonard de Vinci

Sommaire

Motivations scientifiques

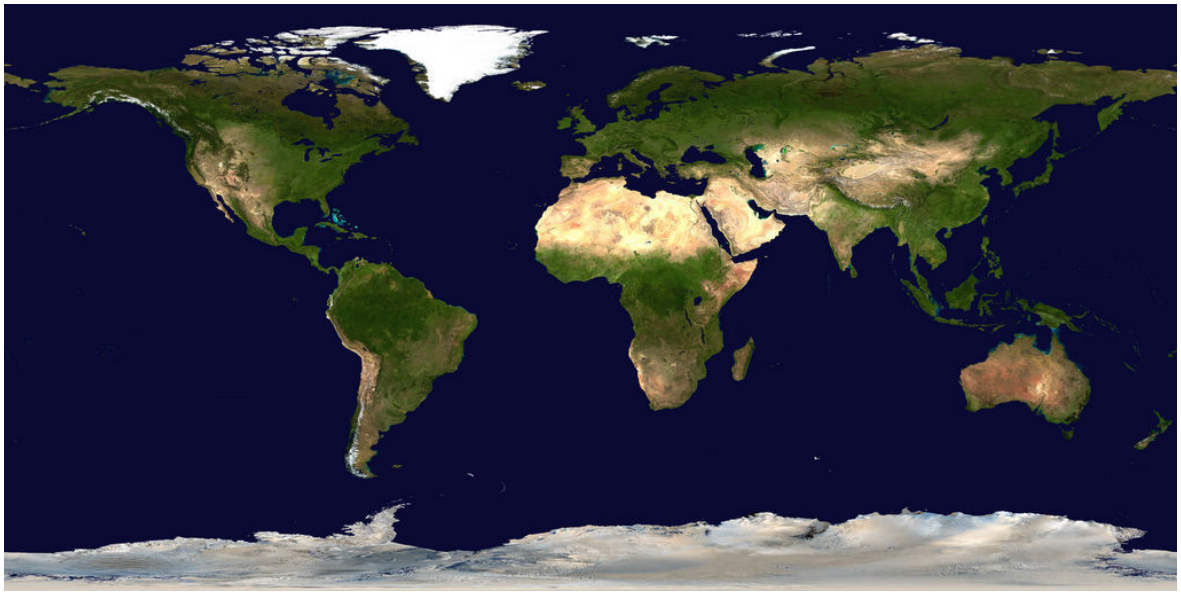


FIGURE 1 – Cette image spectaculaire « Blue Marble » est l'image en couleurs-vraies la plus détaillée de la Terre entière à ce jour (extraite et traduite de la NASA, 2004, plus de détails sur https://svs.gsfc.nasa.gov/2915)

Depuis leur apparition, il y a environ 40 millions d'années, les mammifères marins sont aujourd'hui de véritables trésors scientifiques. Leurs multiples évolutions dans le milieu marin font de ces animaux, à travers leurs gènes et leur histoire, des porteurs d'informations précieuses pour de nombreux domaines scientifiques, principalement liés au milieu marin. Telle l'observation du vol des oiseaux qui conduisit l'Homme à développer l'aviation, ou l'observation du glissement des gouttes d'eau sur une fleur de lotus qui permit l'élaboration de tissus hydrophobes, l'observation des mammifères marins favorise le développement de nombreuses applications et avancées utiles à l'Homme. Nous pensons, à titre d'exemple, à la navigation et la communication sous-marines (les baleines migrent sur de longues distances, arrivent à se repérer sous l'eau et communiquent entres elles par des sons), la détection et la localisation de sources sous-marines (les baleines détectent et localisent leurs prédateurs et leurs proies sous l'eau), la cartographie des fonds marins à travers les ondes acoustiques (les dauphins utilisent les ultrasons comme un SONAR), la conception de drones sous-marins (la peau des mammifères marins résiste à la corrosion de l'eau de mer et certains cachalots peuvent plonger en apnée jusqu'à 3000 mètres de profondeur), l'étude du vieillissement chez les mammifères (la baleine du Groënland vit jusqu'à 200 ans)… Cependant, même si l'étude de ces animaux emblématiques de la vie sous-marine est une source d'informations précieuses, il est généralement difficile de les observer visuellement, en particulier les cétacés (les baleines et les dauphins) qui passent la majeure partie de leur vie sous l'eau.

Même si l'Océan (ou océan mondial) recouvre environ 70 % de la surface de notre planète bleue (*cf.* figure 1), il reste un « univers » méconnu et difficilement accessible. Que ce soit par une description chimique (l'Océan comme solution électrolytique) ou par une description physique (l'Océan comme modèle d'état de la salinité, de la température et de la pression), sa composition est telle que la pression augmente très rapidement avec la profondeur, les ions formant la salinité sont corrosifs et surtout les ondes électromagnétiques sont rapidement absorbées en plus de subir des effets dissipatifs dans ce milieu. Il est en conséquence impossible d'appliquer à l'Océan l'ensemble des technologies utilisées pour l'exploration terrestre, aérienne et spatiale qui utilisent majoritairement des instruments de mesure basés sur des ondes électromagnétiques.

Néanmoins, sa densité et son élasticité font de l'Océan un milieu privilégié pour l'utilisation

des ondes acoustiques. En utilisant ces propriétés, il devient possible d'« observer » le paysage sonore de l'Océan, soit par l'*acoustique active*, c'est-à-dire en émettant des ondes acoustiques et en étudiant leurs réflexions (principe du SONAR), soit par l'*acoustique passive*, c'est-à-dire en « écoutant » les ondes acoustiques à l'aide d'hydrophones par exemple.

Contexte d'étude

Dans ce travail de thèse, en partenariat avec l'entreprise SERCEL, nous participons au contexte général d'étude des mammifères marins en nous focalisant sur la **reconnaissance de formes** appliquée à l'identification des sons produits par les cétacés, en particulier les **mysticètes** (baleines à fanons), avec une prise en compte particulière, celle de travailler sur des signaux issus d'**environnement sismique**.

La **reconnaissance de formes** est une sous-branche de l'*intelligence artificielle*. L'objectif de cette discipline est de développer des méthodes capables de reconnaître automatiquement des « formes » dans des signaux complexes (paroles, images, texte manuscrit, séquence d'ADN, etc.). Dans notre contexte, les formes à reconnaître sont associées aux sons produits par les mysticètes qui communiquent par des cris (ou vocalises) sous l'Océan.

L'**environnement sismique** correspond au contexte d'application de l'entreprise SERCEL, une société spécialisée dans la conception et l'utilisation d'outils géophysiques. Cette entreprise effectue notamment des prospections pétrolières marines, c'est-à-dire qu'elle utilise l'*acoustique active*, en l'occurrence des sources *sismiques* de forte amplitude sonore ou *tirs sismiques*, afin d'obtenir des données (généralement traitées par d'autres sociétés comme Ifremer ou la CGG par exemple) qui permettent de cartographier les différentes strates océaniques. En cas de présence de pétrole, ces données peuvent ensuite être exploitées par les groupes pétroliers.

En raison de la forte puissance sonore produite par les tirs de canon à air (installés sur les *bateaux sismiques*), des réglementations ont été mises en place à des fins préventives pour la protection des mammifères marins. Cela se traduit par l'application de procédures concrètes lors de campagnes sismiques. En particulier, il convient de vérifier qu'il n'y a pas de mammifères marins présents dans un rayon de 500 mètres autour du canon à air pendant la production d'ondes sismiques. Pour répondre aux besoins d'observer les mammifères marins, SERCEL a développé un logiciel de surveillance acoustique passive : QuietSea. Ce logiciel fonctionne en adéquation avec un réseau d'hydrophones et permet de réaliser la détection et la localisation des cétacés.

Objectifs

Afin d'anticiper les futures contraintes réglementaires, SERCEL cherche à étendre le périmètre fonctionnel de QuietSea en y intégrant des modules offrant la possibilité de faire de la reconnaissance. Cette anticipation devient notamment nécessaire, car certaines réglementations différencient d'ores et déjà les actions à réaliser en fonction des espèces. De plus, l'utilisation de bateaux sismiques a un coût très élevé pour l'entreprise. Celle-ci veut avoir une grande confiance dans les détections de potentiels cétacés. Voici pourquoi avoir un module de reconnaissance pourrait également permettre de confirmer ou d'infirmer les éventuelles détections. C'est dans ce contexte de continuité du logiciel QuietSea que s'inscrit cette thèse en lien avec SERCEL. Nous avons borné notre cadre d'étude en nous focalisant sur les mysticètes. C'est la raison pour laquelle ce manuscrit ne contiendra pratiquement que des informations relatives à ces baleines sans nécessairement considérer le cas des autres mammifères marins. Notre problématique générale est la suivante, il s'agit de :

« Mettre en œuvre un algorithme de reconnaissance des vocalises de mysticètes en environnement sismique. »

En résumé, l'objectif de ce travail de thèse est de proposer une méthode de reconnaissance des vocalises de mammifères marins en portant une attention particulière à la gestion des fausses alarmes, c'est-à-dire à la façon dont la méthode permet de confirmer ou d'infirmer les détections éventuelles de mammifères marins proposées par le logiciel QuietSea.

Organisation du manuscrit

La manuscrit de thèse est organisé en 4 chapitres :

1. Contexte et problématique
2. La classification en bioacoustique
3. SINR-SRC
4. Une extension de SINR-SRC : le détecteur multiclasses

Le **premier chapitre** commence par définir la reconnaissance de formes, la bioacoustique et les sources sonores présentes dans l'Océan. Ensuite, nous présentons les mammifères marins. En particulier, nous nous focalisons sur les mysticètes avec pour objectif de détailler l'ensemble des informations qui sont susceptibles d'être exploitées pour réaliser la reconnaissance des sons de ces baleines. Puis, nous introduisons notre partenariat avec l'entreprise SERCEL ce qui nous permet de définir précisément notre problématique. Enfin, nous exposons les contraintes inhérentes à notre problématique et la démarche scientifique adoptée par rapport à ces contraintes afin de fixer au mieux notre cadre d'étude.

Le **deuxième chapitre** établit ce qu'est la classification, tout d'abord à un niveau général par rapport à sa finalité, puis comment celle-ci est utilisée dans la littérature bioacoustique. Nous séparons ainsi les méthodes de reconnaissance en deux approches principales. La première approche est basée sur les méthodes de **réduction de dimensions** ou de **partitionnement**. L'idée est de construire un système de reconnaissance qui va projeter les données d'entrées dans un sous-espace partitionné en autant de « régions », ou dimensions, qu'il y a de classes. De cette façon, la position géométrique d'une donnée d'entrée dans l'espace d'arrivée détermine à quelle classe elle est apparentée. La seconde approche utilise une **mesure de similarité** qui compare la donnée d'entrée aux données apprises par le système. L'identification de la classe est déterminée par la classe apprise la plus proche, au sens de la mesure considérée, de la donnée observée. Enfin, nous discutons des problématiques de validation des algorithmes de reconnaissance.

Les deux chapitres suivants représentent nos contributions.

Le **troisième chapitre** présente la mise en œuvre et l'évaluation de notre méthode de reconnaissance dénommée SINR-SRC. Dans cette situation, les données d'entrée du système sont considérées étant la sortie d'un détecteur. A partir de la notion de mesure de similarité, nous introduisons l'utilisation des représentations parcimonieuses et l'apprentissage de dictionnaire pour représenter les données et faire de la reconnaissance. Nous supposons dès lors que cris ou vocalises de mysticètes se trouvent dans un sous-espace linéaire décrit par une représentation basée sur un dictionnaire. SINR-SRC tient compte de la gestion des fausses alarmes (et plus généralement des classes inconnues) en refusant d'assigner le signal observé à une classe donnée s'il n'est pas inclus dans le sous-espace linéaire couvert par les dictionnaires de vocalises de mysticètes. Le rejet du bruit (ou plus généralement des classes inconnues) est réalisé sans apprentissage de descripteurs. De plus, la méthode proposée est modulaire au sens où les classes de vocalises peuvent être ajoutées ou retirées du système de reconnaissance sans nécessiter de « re-apprentissage ». Le classifieur est facile à concevoir puisqu'il repose sur quelques paramètres. Des résultats expérimentaux sur cinq types de vocalises de mysticètes sont présentés. Ils comprennent des cris en Z

ou Z-call de baleine bleue Antarctique, deux types de cris de baleine bleue pygmée de Madagascar, des 20Hz-pulse de rorqual commun et des D-call de baleine bleue de l'Atlantique Nord. Sur cet ensemble de données, contenant 2 385 cris et 15 000 échantillons de bruit, un rappel moyen de 92,1 % est obtenu et 97,3 % des données de bruit (persistantes et transitoires) sont correctement rejetées par le classifieur. Pour finir nous illustrons également la possibilité de prendre en compte les effets de la dispersion modale dans l'apprentissage de dictionnaire. Pour finir, des preuves de concepts sont présentées afin de démontrer les possibilités d'utilisation de notre méthode vers de l'apprentissage incrémental semi-supervisé.

Le **dernier chapitre** propose de développer une extension de l'algorithme SINR-SRC afin qu'il puisse réaliser la détection et la reconnaissance conjointement. C'est-à-dire que les données d'entrée ne sont plus traitées à la sortie d'un détecteur mais, au contraire, le signal est balayé et traité. La mise en œuvre de cette méthode, dénommée le *détecteur multiclasses*, est détaillée. Puis les performances sont mesurées à travers une base de données similaire à la base de données utilisée pour SINR-SRC. Avec cette nouvelle architecture, le détecteur multiclasses permet, au contraire de SINR-SRC, de traiter plusieurs classes simultanément. Le détecteur multiclasses est capable de détecter et de reconnaître une vocalise d'intérêt. La détection ainsi obtenue permet d'annoter automatiquement les vocalises d'intérêt en donnant un temps de début et un temps de fin de détection, une plage de fréquence, une estimation du SNR et un coefficient de confiance. Les résultats obtenus nous indiquent des pistes d'amélioration. L'apprentissage de certaines classes, comme les D-call, doit être plus représentatif. La reconnaissance des 20Hz-pulse doit être rendue plus robuste aux environnements sismiques.

Contributions scientifiques

Ce travail de thèse a donné lieu aux contributions scientifiques suivantes :

- T. Guilment, F.-X. Socheleau, D. Pastor et S. Vallez, « Reconnaissance de vocalises de mysticètes par apprentissage de dictionnaire et représentations parcimonieuses », *Sea Tech Week - SERENADE*, le 12 octobre 2016 à Brest. (Présentation orale)
- T. Guilment, F.-X. Socheleau, D. Pastor et S. Vallez, « Sparse Representation-Based Classification of Mysticete Calls », submitted to *The Journal of the Acoustical Society of America*, April 2018.
- T. Guilment, F.-X. Socheleau, D. Pastor et S. Vallez, « Joint detection-classification of baleen whale sounds using sparse representations », submitted to *DCLDE 2018*, June 2018, Paris. (Présentation orale)

Chapitre 1

Contexte et problématique

« *Dans la vie, rien n'est à craindre, tout est à comprendre.* »

— Marie Curie

Sommaire

Le contexte d'étude se situe dans le domaine de la *bioacoustique* en se focalisant principalement sur la *reconnaissance de formes* ici appliquée à l'identification de vocalises de mammifères marins. Dans ce chapitre, après avoir rappelé quelques principes de la reconnaissance de formes, nous revenons sur la définition de la bioacoustique. Ensuite, nous décrivons le paysage sonore de l'Océan. Enfin, nous précisons les frontières de notre problématique notamment dans le contexte de la prospection sismique dans le cadre de notre partenariat avec l'entreprise SERCEL.

1.1 La reconnaissance de formes

La reconnaissance de formes est une sous-branche de l'*intelligence artificielle*. Comme son nom l'indique, elle consiste à automatiquement identifier des « formes » (« structures » ou « motifs ») dans des signaux complexes (paroles, images, séquence d'ADN, texte manuscrit, etc.). La reconnaissance de formes, et particulièrement la notion de « formes » fait appel à des concepts difficiles à définir de façon purement formelle mathématiquement et se définit surtout par son contexte applicatif. Il s'agit alors de mettre en œuvre des méthodes ou algorithmes qui vont reproduire de façon automatique des tâches de reconnaissance ou d'interprétation que l'Homme effectue assez facilement, mais dont la façon de procéder n'est pas toujours explicite.

1.2 La bioacoustique

Comme définie très clairement par la société française d'acoustique dans « Le livre blanc de l'acoustique en France 2010 p. 94 » :
« *La bioacoustique est une branche de la science apparentée à d'autres disciplines scientifiques qui étudie la production et la réception des signaux acoustiques chez les animaux. À ce titre elle s'intéresse aux organes acoustiques et aux appareils de production et de réception des sons, aussi bien qu'aux processus physiologiques et neurophysiologiques par lesquels ceux-ci sont produits, reçus et traités. La bioacoustique tente de comprendre le lien entre les caractéristiques des sons produits par un animal et le type d'environnement dans lequel il les utilise, ainsi que les fonctions pour lesquels ils sont conçus. [...] D'autre part elle s'intéresse et concerne les effets sur les animaux des nuisances sonores produites par l'homme. Mais cette discipline intègre également les récents développements biomédicaux qui permettent l'exploration des réponses électrophysiologiques face à des stimuli acoustiques (potentiels évoqués auditifs par exemple), et pour détecter, entre autres, les lésions du système auditif dérivées de l'exposition à des sources sonores anthropogéniques.* »

Comme vue plus haut, la reconnaissance de formes a toute sa place dans ce contexte de la bioacoustique appliquée aux mammifères marins car elle permet d'automatiser des tâches d'identification faites par l'homme et en particulier celle d'identifier les mammifères marins à partir de leurs vocalises. Nous verrons plus en détails les sons produits par les mammifères marins dans la section 1.4 ci-dessous.

1.3 Le paysage sonore de l'Océan : quand nos oreilles deviennent nos yeux

La bioacoustique prend tout son sens lorsqu'il s'agit d'identifier des animaux difficiles à observer visuellement. Ces difficultés d'observation peuvent être environnementales (forêt très dense, haute montagne difficilement accessible, désert, profondeurs de l'Océan, etc.), météorologiques (forte pluie, brouillard, tempêtes, etc.) ou encore liées à la façon de vivre de l'animal à observer (animal migratoire, nocturne, etc.). Dans ces conditions, que ce soit pour l'homme qui souhaite réaliser des observations de ces milieux ou pour les animaux qui vivent dans ces environnements, la vie se passe « à l'aveugle ». Il est alors nécessaire de s'adapter en usant d'autres sens que la vue

pour se repérer et/ou communiquer.

Dans le cas des mammifères marins, le milieu océanique est tel que les signaux électromagnétiques (comme la lumière) et les signaux chimiques (comme les odeurs) ne se propagent pas très loin (de l'ordre de centaines de mètres pour la lumière et pour les odeurs) et pas très bien. Néanmoins, sa densité et son élasticité font de l'Océan un milieu privilégié pour les ondes acoustiques qui se propagent 4 à 5 fois plus rapidement que dans l'air [1] et sont également moins absorbées. Ainsi, il est possible d'« observer » le paysage sonore de l'Océan en « écoutant » les ondes acoustiques.

Ce paysage sonore est constitué d'une infinité de sons (*cf.* figures 1.1) regroupés généralement en trois catégories : la *biophonie*, correspondant aux sons émis par les être vivants (des sons de petits organismes, comme le benthos ou le pélagos, aux vocalises de grandes baleines), la *géophonie*, composée des sons naturels non-biologiques comme la météorologie (pluie, vent, éclairs, etc.) ou les phénomènes géologiques (séismes, éruptions volcaniques sous-marines, etc.) et l'*anthropophonie*, formée par l'ensemble des sons produits par l'Homme et ses activités [KRAUSE, 2008].

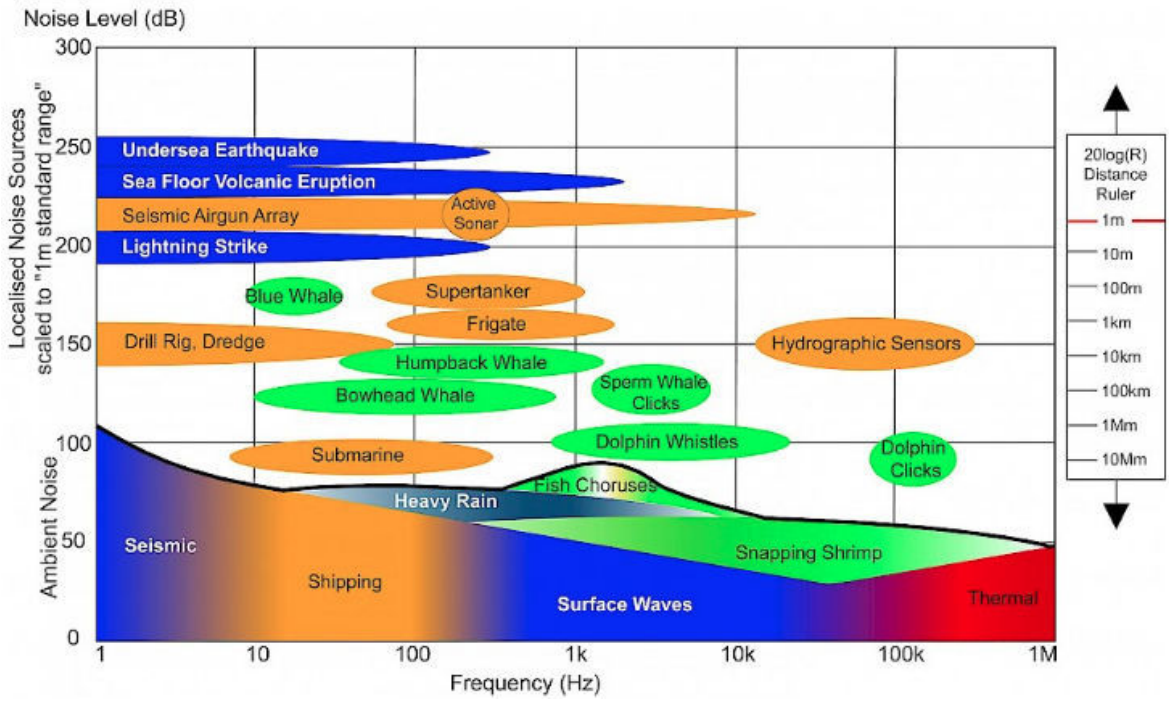


FIGURE 1.1 – Spectre du niveau de bruit (« Intensité acoustique par Hertz ») en fonction de la fréquence (mesurée en Hertz) présentant les niveaux et fréquences des sources sonores anthropiques (en orange) et naturelles (géophonie en bleu et rouge, et biophonie en vert et jaune) présentes en milieu marin. (légende inspirée et image extraite du site de la commission [OSPAR, 2018]

Afin de poursuivre avec l'ensemble des sons existants sous l'océan, nous présentons ci-après la famille des mammifères marins puis nous poursuivons avec la présentation de notre partenariat industriel qui participe notamment à l'antropophonie.

1.4 La grande famille des mammifères marins

Il existe quatre grands groupes de mammifères marins (*cf.* figure 1.2) : les mysticètes, les odontocètes, les pinnipèdes et les siréniens. Ces quatre familles peuvent être divisées en deux groupes :

1. La célérité moyenne du son dans l'eau est de l'ordre de 1500 m/s

les mammifères non inféodés à l'eau pour les siréniens et les mammifères inféodés [2] à l'eau pour les 3 groupes restants. Notre travail s'est focalisé sur les mysticètes. Voici pourquoi ce manuscrit contient principalement des informations relatives à ces baleines sans considérer systématiquement le cas des mammifères marins en général.

1. Baleine à bosse

2. Dauphin

3. Lion de mer

4. Lamentin

FIGURE 1.2 – Représentants des 4 groupes de mammifères marins : 1. Mysticètes, 2. Odontocètes, 3. Pinnipèdes et 4. Siréniens (images extraites de www.wikipédia.org)

Comme nous le verrons plus précisément lors de la définition de notre problématique (*cf.* section 1.6), le premier pas vers notre futur algorithme de reconnaissance des vocalises des mysticètes est tout d'abord d'identifier les différents mysticètes. De plus, leur localisation à travers le globe peut peut-être fournir des informations discriminantes pour la reconnaissance. Pour obtenir la liste des mysticètes ainsi que leur distribution géographique nous nous sommes principalement inspirés des sites WoRMS (**WO**RLD **R**EGISTER OF **M**ARINE **S**PECIES), CBI (**C**OMMISSION **B**ALEINIÈRES **I**NTERNATIONALE) [2018] et IUCN (**U**NION **I**NTERNATIONALE POUR LA **C**ONSERVATION DE LA **N**ATURE) [2018] qui sont des sites officiels réunissant de multiples sources relatives à chaque espèce.

1.4.1 Liste des mysticètes

Avant d'énoncer la liste des mysticètes, il faut préciser que certaines espèces sont peu documentées pour avoir la preuve de leur existence actuelle (ex. : baleine franche pygmée). De la même façon, certaines espèces peuvent être sujettes à controverse dans la communauté des biologistes qui n'est pas toujours unanime quant à l'appellation de l'ensemble des espèces. Cette liste constitue donc une proposition et non une référence (ex. :le rorqual tropical *Balaenoptera edeni*). Nous avons choisi ici de garder la hiérarchie utilisée en taxonomie [3] (*cf.* modèle figure 1.3) avec le sous-ordre en français, la famille en latin, le genre en latin et le nom de l'espèce en français (*cf.* figure 1.4). Pour une classification plus complète avec les appellations latines, anglaises et françaises, le lecteur pourra se référer à l'annexe A.

FIGURE 1.3 – Modèle taxonomique

1.4.2 Répartition géographique

Les informations sur la distribution des espèces à travers le globe sont globales et non exhaustives. Nous mettons l'accent ici sur les grandes différences de répartition des espèces avec deux cartes représentatives. Nous ne prenons pas en compte, dans les cartes suivantes, ni les périodes

2. Une espèce inféodée à un organisme ou à un milieu est une espèce qui est liée très fortement à cet organisme ou ce milieu et qui peut difficilement vivre sans celui-ci.

3. *Systema Naturae* (1758) généralise le système de nomenclature binominale (système linéen [Carl Von Linné (1707-1778)])

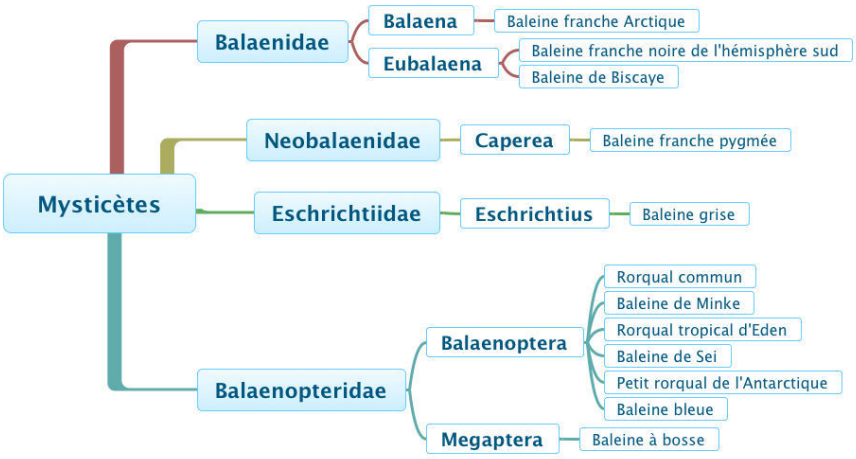


FIGURE 1.4 – Taxonomie (simplifiée) des mysticètes

migratoires ni les périodes de reproduction. Pour plus de cartes concernant une partie des autres espèces, le lecteur pourra se référer à l'annexe B.

FIGURE 1.5 – Répartition de la baleine grise

Comme l'indique la répartition de la baleine grise (*cf.* figure 1.5), il est intéressant de noter que cette espèce est très localisée. Cette information est alors très utile pour la reconnaissance. Par exemple, pour une mission en Atlantique, si une détection affirme la présence de vocalises de baleine grise, nous saurons que c'est une fausse alarme.

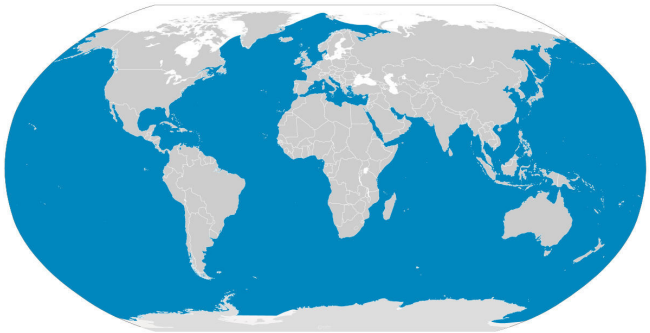


FIGURE 1.6 – Répartition du rorqual commun

Au contraire des espèces très localisées, l'information géographique du rorqual commun (*cf.* figure 1.6) ne permet pas de confirmer ou d'infirmer sa présence suite à une éventuelle détection.

1.4.3 La communication chez les cétacés

Avant de présenter plus en détail les caractéristiques des vocalises émises par les mysticètes, il convient de se demander dans quel but ces baleines communiquent-elles. Nous savons que, parce que l'Océan est un milieu aux conditions difficiles (augmentation rapide de la pression avec la profondeur, corrosion, absorption des ondes électromagnétiques...), alors les ondes sonores sont une alternative intéressante comme vecteur d'informations. Néanmoins, ce n'est pas une raison suffisante pour justifier la production sonore des cétacés sachant que toutes les espèces marines n'émettent pas toutes des sons. La réponse à cette problématique reste une question ouverte et la communauté biologiste n'a pas fini d'interpréter l'ensemble des communications produites par les mammifères marins. Cependant, dans le cas des cétacés, il est possible de décrire plusieurs fonctions associées à ces communications. C'est ce que propose le chapitre 11 du livre de [MANN, 2000] (*cf.* figure 1.7).

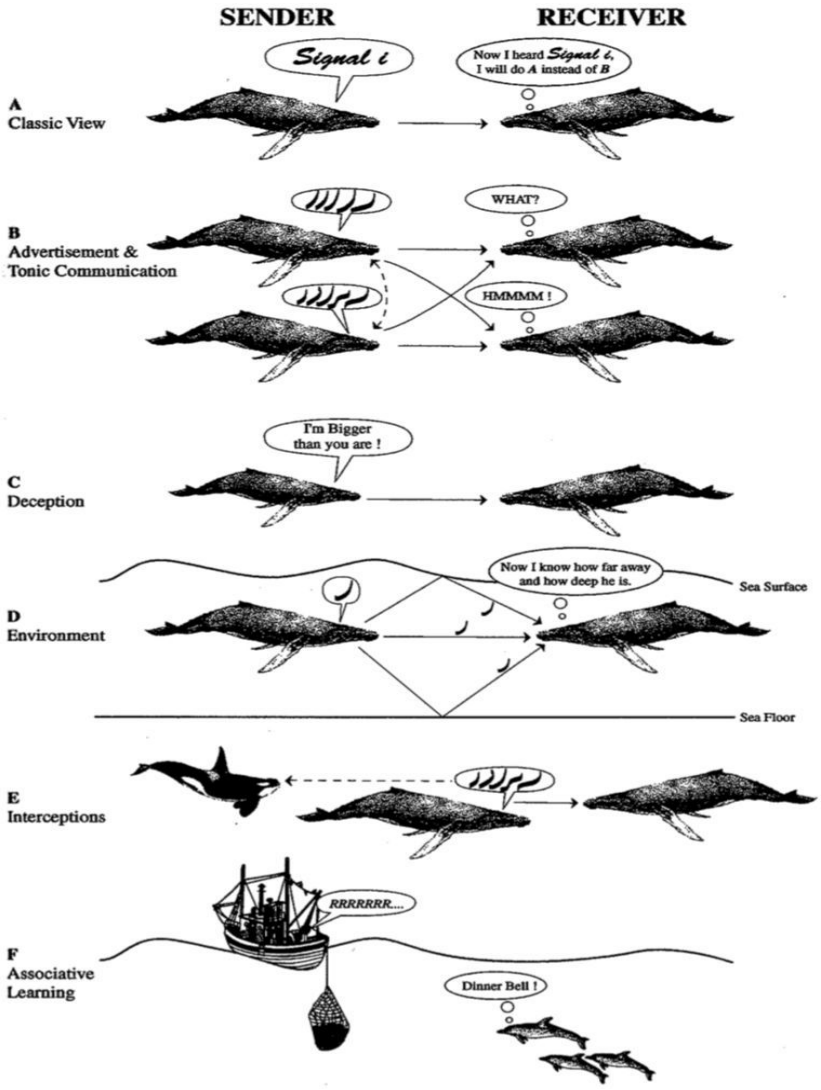


FIGURE 1.7 – Image extraite et traduite de [MANN, 2000] sur la communication des cétacés. A est la vision « classique » de la communication et de B à F la vision est défini comme « écologique » par l'auteur. B : Annonce (ou publicité) et communication tonique. C : Tromperie. D : Environnement. E : Interceptions. F : Apprentissage associatif.

Dans cette représentation, l'animal émetteur (*sender* en anglais) communique une information à un ou plusieurs « destinataires » (*receiver* en anglais). Dans la situation A, l'auteur présente une situation « classique » où l'émetteur renseigne le destinataire en transmettant une information dont ce dernier a besoin et/ou qu'il ne connaît pas. L'auteur parle de « réduction d'incerti-

tude ». Les situations suivantes viennent contraster cette vision du partage de connaissances en utilisant la communication à d'autres fins, l'auteur parle d'une vision « écologique ». Ce point de vue inclut les concepts d'annonce [4] (ou publicité) et de communication tonique [5] (*advertisement & Tonic communication* de la situation B) : l'idée est de considérer la communication comme un processus potentiellement manipulateur, impliquant potentiellement de nombreuses parties et de nombreux signaux. Cette vision ne suppose pas que tous les signaux sont « honnêtes », mais il envisage la possibilité d'une tromperie (*deception* situation C). Ce point de vue reconnaît que les signaux changent au fur et à mesure qu'ils traversent l'environnement, et que cela peut informer le destinataire à la fois sur le signal et sur l'environnement physique (situation D). Il peut y avoir plusieurs destinataires d'un signal, et d'autres peuvent intercepter le signal, en utilisant l'information à leur propre avantage (situation E). Les destinataires involontaires peuvent même apprendre des réactions bénéfiques aux signaux d'autres espèces (situation F), comme lorsque des dauphins apparaissent soudainement sur un bateau de pêche lorsqu'un treuil qui libère du poisson est allumé. Du point de vue de la reconnaissance, il est également pertinent de prendre conscience de l'organisation sociale qui peut exister chez les cétacés comme par exemple chez les baleines à bosse [CLAPHAM, 1996]. Dans notre contexte, cela permet d'anticiper que les données seront probablement liées à plusieurs vocalises de natures différentes (associées à une fonction particulière comme la recherche de nourriture, la chasse, la reproduction, etc.) et provenant de plusieurs individus. Le fait que certains signaux sont associés à une fonction peut également participer à une localisation plus précise de l'espèce considérée. Par exemple, si on connaît par avance le son produit par une baleine quand elle va manger et l'endroit où elle trouve sa nourriture alors la détection de cette baleine n'en sera que plus accessible.

1.4.4 Vocalises

Nous proposons ici d'illustrer la diversité des sons émis par les mysticètes par des spectrogrammes de la littérature. Nous pourrons ensuite détailler la nature générale de ces vocalises.

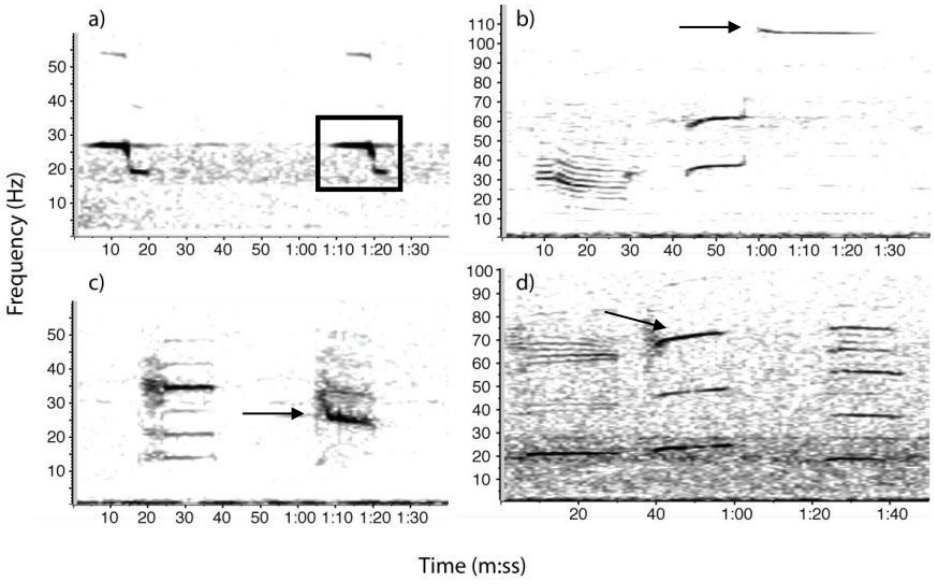


FIGURE 1.8 – Spectrogrammes de cris de différentes sous-espèces et sous-populations des baleines bleues détectés par les hydrophones du réseau OHASISBIO **a)** 2 vocalises de baleine bleue Antarctique, **b)** 1 vocalise de baleine bleue pygmée de type Sri Lanka, **c)** 1 vocalise de baleine bleue pygmée de type Madagascar, **d)** 1 vocalise de baleine bleue pygmée de type Australie. Les flèches et la boîte représentent la partie du signal utilisé pour réaliser le *template* lors de la détection automatique. Légende traduite et image extraite de [SAMARAN et GUINET, 2012])

4. L'idée est d'influencer la décision du destinataire, éventuellement à son détriment
5. L'auteur défini la communication tonique comme une communication qui implique une réponse immédiate.

Comme le montre la figure 1.8, extraite de [SAMARAN et GUINET, 2012], sur les sous-espèces de baleine bleue, chaque sous-espèce a véritablement son panel de vocalises qui lui est propre. En conséquence, vouloir reconnaître la baleine bleue n'a plus vraiment de sens. Il s'agit en réalité d'identifier toutes les sous-espèces de baleine bleue. De plus, malgré un aspect bien stéréotypé de ces vocalises, le tableau (*cf.* figure 1.9) extrait de [MCDONALD et collab., 2009] montre que la fréquence moyenne des baleines bleues a tendance à diminuer au fil des années. En conséquence, une méthode capable de bien identifier les vocalises de baleine bleue aujourd'hui sera peut-être incapable de le faire demain.

Song type (region)	Duration (s)	Initial			Final			Change yr^{-1}		Relative
		Year	Freq.	dB	Year	Freq.	dB	dB	Density (%)	density (%)
NE Pacific	19	1960	22.2	188.4	2003	15.9	185.5	0.067	1.8	100
SW Pacific	6 + 12	1964	30.8/25.3	190.7	1998	25.8/20.1	188.8	0.027	0.8	47
NW Pacific	12 + 12	1968	25/23	187.1	2001	19.45/17.9	184.9	0.066	1.8	126
N Atlantic	8	1959	23	196.3	2004	17.6	193.9	0.053	1.4	11
S Ocean	10	1995	28.5	196.2	2005	26.9	195.7	0.050	1.3	7
N Indian Ocean	27	1984	116	199.8	2002	106	199.0	0.044	1.2	3
SE Indian Ocean	20	1993	19.5	186.9	2000	19.0	186.6	0.043	1.2	82

FIGURE 1.9 – Evolution des fréquences moyennes des baleines bleues extraite de [MCDONALD et collab., 2009]

En plus d'une grande diversité des vocalises au sein même d'une espèce (avec ses sous-espèces), il peut exister une grande variabilité au sein même de la vocalise considérée. C'est ce que démontre la figure 1.10, extraite de [TRYGONIS et collab., 2013], qui compare des vocalises présentées par des couples de spectrogrammes qui sont générés avec les mêmes paramètres et qui pourtant présentent des caractéristiques différentes (nombre d'harmoniques, répartition d'énergie, durée, ...) pour chaque même type de vocalise (*cf.* légende de la figure 1.10).

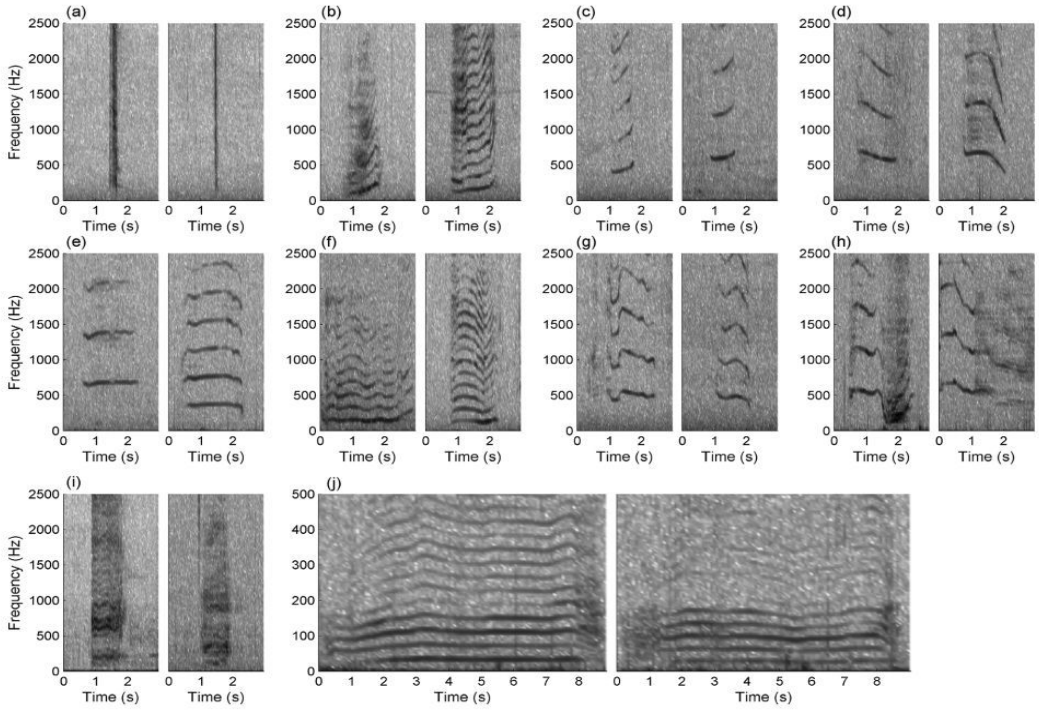


FIG. 2. North Atlantic right whale call types, showing two exemplars per case to illustrate the variability within categories: (a) gunshot, (b) upcall, (c) upcall high, (d) downcall, (e) constant, (f) tonal low, (g) modulated, (h) hybrid, (i) pulsive, (j) foghorn. Spectrogram parameters for call types (a)–(i) are 2048 samples for FFT and hamming window, 50% overlap. Window size for (j) is 8192 samples (hamming, 50% overlap), displayed at a frequency scale of 0–500 Hz. Note the absence of contamination by flow or boat noise that allowed the reliable identification of the fundamental harmonic for feature extraction.

FIGURE 1.10 – Vocalises de baleine franche de l'Atlantique Nord (légende et image extraites de [TRYGONIS et collab., 2013])

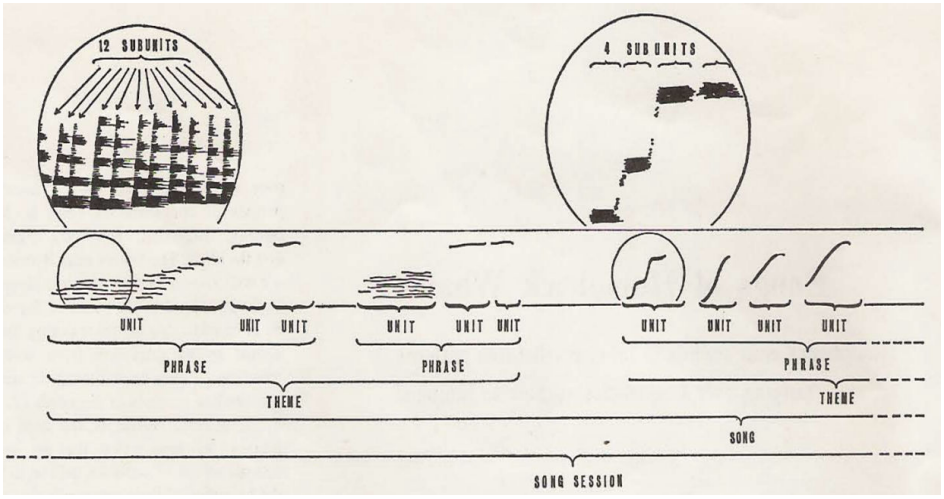


FIGURE 1.11 – Structure d'une chanson par [PAYNE et MCVAY, 1971]

La figure 1.11, extraite de [PAYNE et MCVAY, 1971], décrit la structure de chanson produite par les mysticètes suivants : les baleines bleues, le rorqual commun, la baleine de Minke, la baleine à bosse et la baleine du Groënland (baleine franche Arctique). Cette structure de chanson montre qu'une vocalise ne se suffit pas toujours à elle-même et peut appartenir à un contexte bien spécifique.

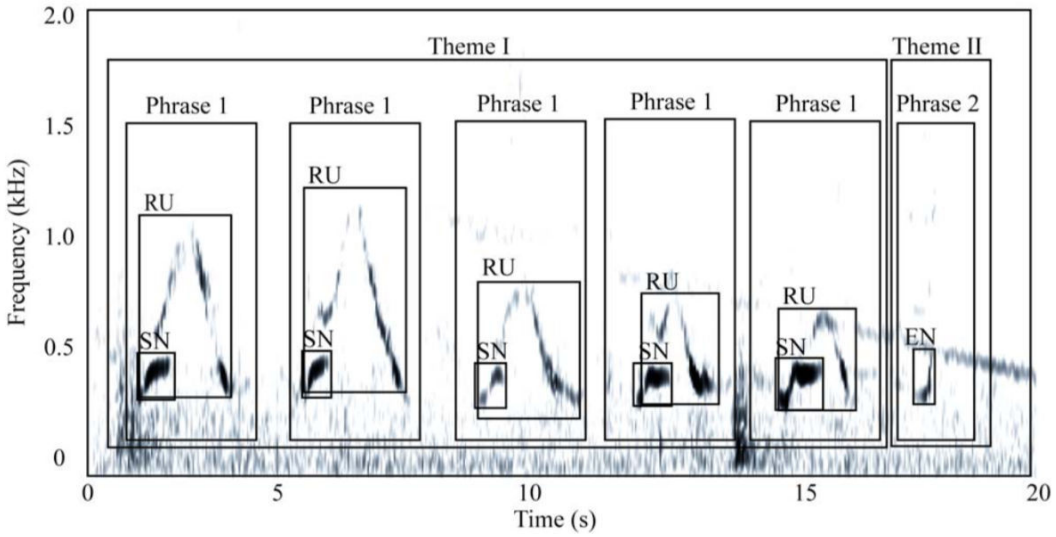


FIGURE 1.12 – Systématique créée par [PAYNE et MCVAY, 1971] (*cf.* figure 1.11) pour la chanson de baleine à bosse appliquée à la chanson de la baleine du Groënland en 2008 enregistrée dans la baie de Disko extrait de [TERVO O.M., 2011]. Cette chanson est composée de deux thèmes. Le premier thème I consiste en cinq répétitions de phrase 1 tandis que le thème II comprend un exemple de phrase 2. Les notes de chansons sont étiquetées avec des majuscules. Bien que non vérifiés pour cette chanson particulière, nous pouvons remarquer la production simultanée de notes de chanson RU et SN.

Les figures 1.12 et 1.13, extraites de [TERVO O.M., 2011], présentent le cas particulier des baleines franches de l'Arctique capables de modifier leur chanson. Là où les autres mysticètes génèrent des sons assez stéréotypés, cette baleine a la capacité de créer des vocalises nouvelles d'une année à l'autre. A notre connaissance, il existe cinq espèces de mysticètes connues pour produire de telles chansons : la baleine à bosse *Megatera novaeangliae* [PARSONS et collab., 2008; PAYNE et MCVAY, 1971], la baleine bleue *Balaenoptera musculus* [CUMMINGS, 1971], le rorqual commun *Balaenoptera physalus* [WATKINS et collab., 1987], la baleine de Minke *Balaenoptera acutorostrata* [GEDAMKE et collab., 2001; MELLINGER et collab., 2000] et la baleine du Groënland [LJUNGBLAD

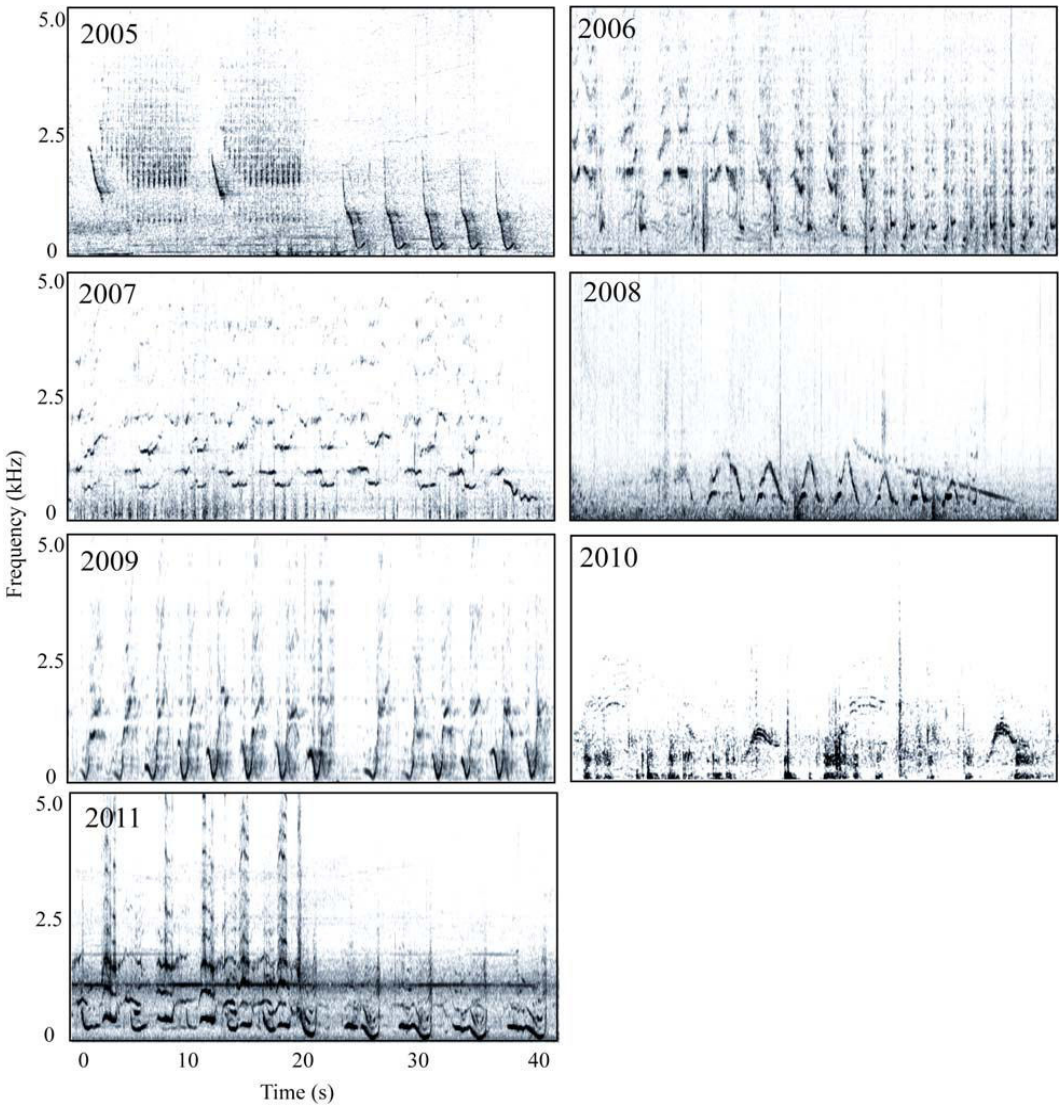


FIGURE 1.13 – Evolution par année de la chanson de la baleine du Groënland extrait de [TERVO O.M., 2011]

et collab., 1982; STAFFORD et collab., 2008] *Balaena mysticetus.*

1.4.5 Les vocalises à la réception

Nous nous intéressons dans cette partie aux transformations (considérations physiques et/ou interférences) que peuvent subir les ondes acoustiques entre leur émission et leur réception par un hydrophone. L'intérêt est de prendre connaissance de ce contexte représentatif des conditions réelles lié à l'acquisition des données afin d'anticiper les éventuelles modifications que subissent les vocalises produites par les mammifères marins depuis leur production jusqu'à leur réception (enregistrement). Nous séparons les éventuelles déformations en une partie liée à la « physique des ondes » à savoir le cas du *multitrajet* (*cf.* figure 1.16) et le cas de la *dispersion modale* (*cf.* figure 1.15). Ensuite, nous présentons un exemple de signaux produits par les mysticètes, mais que nous ne considérons pas comme une vocalise précise : le *chorus* de la baleine bleue (*cf.* figure 1.17).

La célérité des ondes acoustiques

Dans l'océan, la célérité c du son dépend des trois variables d'état : la température, la pression hydrostatique (souvent traduit par la profondeur) et la salinité. Il existe une formule couramment

utilisée pour estimer la célérité du son en fonction de ces trois grandeurs [CLAY et MEDWIN, 1977] :

$$c = 1449.2 + 4.6\mathrm{T} - 0.055\mathrm{T}^2 + 0.00029\mathrm{T}^3 + (1.34 - 0.010\mathrm{T})(\mathrm{S} - 35) + 0.016z$$

où c est la célérité du son en $\mathrm{m.s^{-1}}$, T est la température en °C, S est la salinité en PSU (*Pratical salinity units*) ou ‰ et z est la profondeur en mètres. Comme l'exprime cette formule, il n'est pas simple de déterminer précisément la célérité du son dans un milieu d'eau de mer. Nous pouvons noter que c augmente avec l'augmentation des trois grandeurs. De façon générale, il est raisonnable de considérer que la vitesse de propagation du son se trouve généralement entre 1450 et 1550 $\mathrm{m.s^{-1}}$.

Multi-trajet et dispersion modale

Une propriété intéressante est que le son a tendance à rester « piégé » dans l'eau de mer. Ce sont aux interfaces de ce milieu comme par exemple, la surface, les différentes couches sédimentaires du fond marin, mais également au sein même du milieu comme entre les différents courants et/ou différentes « couches » d'eau de mer aux propriétés différentes que vont apparaître des propriétés acoustiques spécifiques comme le multitrajet et la dispersion modale.

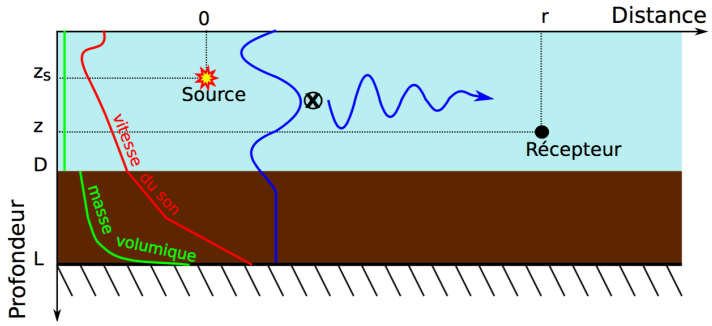


FIGURE 1.14 – Schéma extrait de [BONNEL, 2010]. Illustration de la propagation modale dans un guide d'onde quelconque entre une source impulsionnelle et un unique capteur.

Afin d'illustrer la propagation modale, nous reprenons l'exemple proposé dans [BONNEL, 2010], représenté par la figure 1.14. Dans ce modèle, l'océan est considéré comme un guide d'onde en deux dimensions avec un axe z pour la profondeur et un axe r pour la distance à la source, ici une source impulsionnelle. Cette source, en fonction de ses longueurs d'onde va générer plusieurs *modes* qui vont se propager (courbe bleue horizontale) jusqu'au récepteur. Par rapport à la structure émise à la source, il est important de noter que ce phénomène physique va faire évoluer cette structure. C'est ce que nous pouvons observer sur la figure 1.15.

Comme nous pouvons l'observer, plus la distance à la source est grande et plus l'effet de la propagation modale sur les vocalises est important. Comme expliqué dans [BONNEL, 2010], plus précisément nous sommes en présence de dispersion intermodale (la séparation des modes augmente avec la distance ainsi que lorsque la fréquence diminue) et intramodale (deux fréquences ont des vitesses de groupe différentes et cet effet tend à étaler les signaux en temps) de façon simultanée.

Dans le cas du multitrajet présenté figure 1.16, nous nous trouvons en *grand fond*, c'est-à-dire que la profondeur de l'eau est de l'ordre de plusieurs kilomètres, les ondes acoustiques sont réfléchies entre la surface et le fond marin ce qui génère un phénomène d'échos à la réception. Comme pour la propagation modale, la hauteur d'eau et la longueur d'onde de la source considérées ont un rôle dans le nombre de trajets générés. Néanmoins, à la différence de la dispersion modale, la structure n'est pas changée en fréquence, mais est simplement répétée. En pratique, ce phénomène peut devenir problématique si nous souhaitons par exemple compter le nombre d'individus en identifiant le nombre de vocalises sans nécessairement connaître la localisation de la source. C'est notamment le cas lorsque les vocalises sont assez courtes comme celles du rorquals communs présentées à la figure (b) de la figure 1.16.

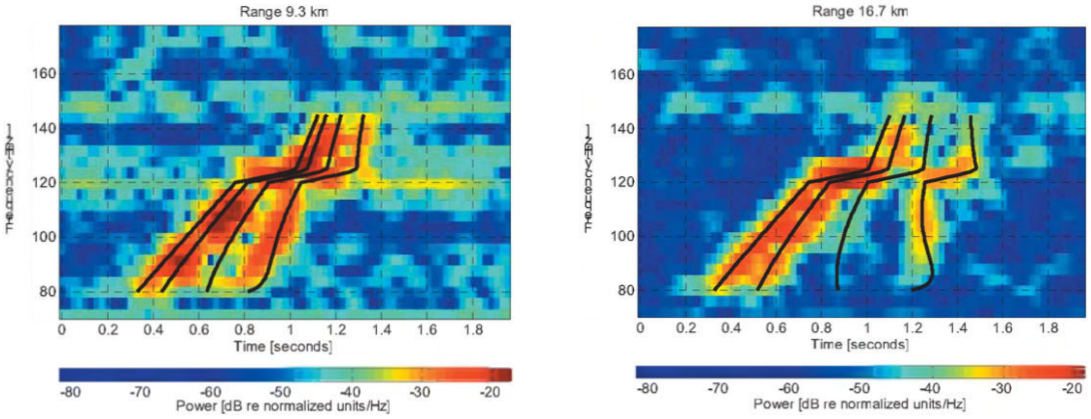


FIGURE 1.15 – Spectrogrammes extraits de [WIGGINS et collab., 2004] d'un cri de baleine franche avec en lignes noires la superposition des modes estimée par le modèle des auteurs à deux distances différentes de la source.

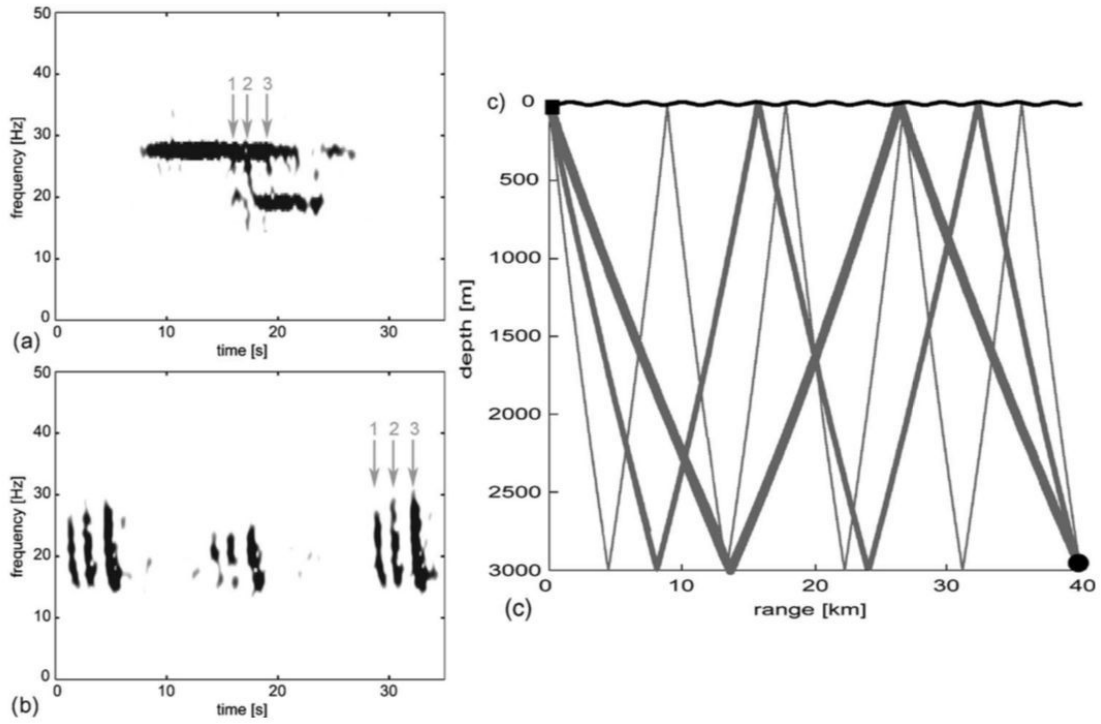


FIGURE 1.16 – Effet du multitrajet extrait de [ŠIROVIĆ et collab., 2007]. Cris enregistrés de baleine bleue a) et de rorqual commun b) au large de la péninsule Antarctique occidentale, montrant les arrivées par trajets multiples. Dans les deux exemples, les chemins montrés étaient les premier, deuxième et troisième rebonds marqués respectivement 1, 2 et 3 ; le chemin direct n'est pas visible. Les distances calculées étaient de 33 km pour le rorqual bleu et de 40 km pour le rorqual commun. Les rebonds théoriques contribuant à l'arrivée des vocalises de rorquals communs sont illustrés dans la partie c), la ligne épaisse représentant le premier rebond, la ligne d'épaisseur moyenne pour le deuxième rebond et la ligne mince pour le troisième rebond. L'emplacement de l'appel de baleine est indiqué par un carré noir et l'emplacement de réception de l'ARP (*Acoustic Recording Packages*) est indiqué par un cercle noir.

Le chorus

Du point de vue du signal et de la reconnaissance de formes, les interférences sont des signaux qui sont structurés, mais qui ne représentent pas de classes ou de formes à identifier. Ils ne sont pas produits par des mammifères marins. Néanmoins, cela ne signifie pas pour autant que tous les signaux produits par les mammifères marins correspondent à une classe bien définie. C'est le

cas notamment des baleines bleues, qui, par la multiplicité de vocalises génèrent un *chorus*, c'est-à-dire un fond d'énergie dans une bande de fréquence bien définie où il est difficile d'identifier une vocalise particulière (*cf.* figure 1.17).

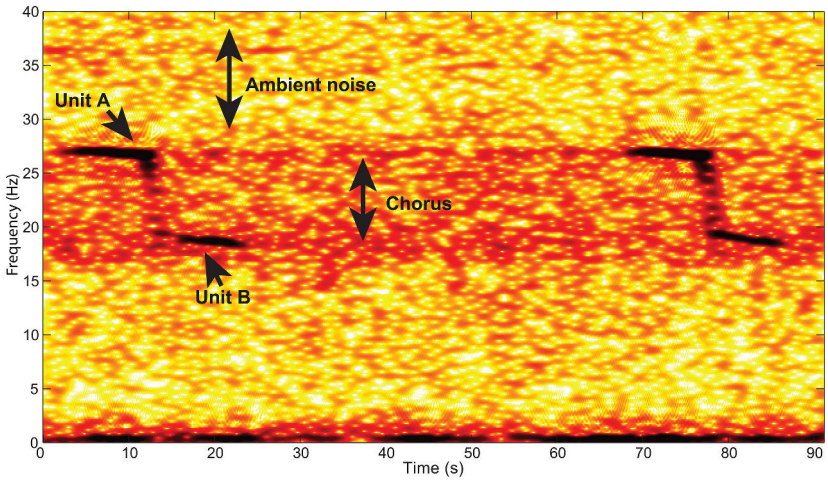


FIGURE 1.17 – Spectrogramme de cri en Z ou *Z-call* de baleine bleue extraite de [LEROY et collab., 2016] illustrant la présence d'un chorus superposé aux vocalises considérées. Il est également possible d'observer le bruit ambiant ainsi qu'un découpage de la vocalise en Z en deux unités A et B.

Les vocalises produites par les baleines sortent alors de la classe de signaux et constituent un véritable « paysage sonore ».

1.4.6 Conclusion sur les mysticètes

Nous pouvons raisonnablement classer l'ensemble des sons produits par les mysticètes comme des sons structurés en énergie et de durée finie. Ces structures sont décrites comme des modulations de fréquence (moans, glissement, ...) et les modulations d'amplitude (clique, pulse, burst, ...). A cela vient s'ajouter la structure rythmique et répétitive des vocalises, pouvant mener pour certaines espèces à une véritable chanson. De plus, chaque espèce produit ses sons avec un rythme spécifique. On parle d'ICI (*Inter-call interval* en anglais), c'est-à-dire d'intervalle inter-vocalises. Ensuite, la propagation des ondes sonores en milieu marin peut donner lieu à de la dispersion modale ou du multitrajet, ce qui, pour les vocalises, peut créer de réelles différences entre le son produit par un mammifère et le son reçu par l'hydrophone. Pour finir, les sons émis par les mysticètes sont dans les basses fréquences et, par conséquent, se propagent sur de longues distances (de l'ordre de quelques kilomètres à plusieurs centaines de kilomètres).

1.5 L'entreprise SERCEL et l'environnement sismique

Maintenant que les mammifères marins ont été présentés, nous présentons ci-dessous notre contexte industriel par la biais de notre partenariat avec SERCEL. L'entreprise SERCEL est une société spécialisée dans la conception et l'utilisation d'outils géophysiques. La définition donnée par l'encyclopédie UNIVERSALIS [2017] définit parfaitement ce contexte : *« La géophysique utilise les méthodes de la physique pour étudier la Terre et son environnement. [...]* (Elle) *permet de connaître à distance les propriétés physiques des matériaux en analysant des signaux détectables depuis la surface du sol ou dans l'espace. »*

Plus précisément, SERCEL utilise une sous-branche de la géophysique, les méthodes dites sismiques. *« Les ondes sismiques générées par des séismes, naturels ou artificiels, constituent un puissant moyen d'investigation des propriétés élastiques des matériaux terrestres. En géophysique, les méthodes sismiques sont celles qui permettent d'imager les milieux profonds avec le plus de finesse. Parmi les techniques de prospection industrielle, celles qui utilisent la réflexion d'ondes sismiques*

à incidence quasi verticale sont très efficaces pour imager les limites des formations sédimentaires piégeant les hydrocarbures. Leur principe est proche de celui du sonar embarqué sur les navires pour mesurer la profondeur des fonds marins, mais les dispositifs d'émission et de réception des signaux sismiques sont beaucoup plus complexes et lourds à mettre en œuvre. Les méthodes de réflexion sismique ont donné lieu à des développements considérables, tant pour les dispositifs de sources et de capteurs que pour le traitement numérique des signaux recueillis. En mer, les capteurs sont constitués de flûtes d'hydrophones remorquées et déployées sur plusieurs kilomètres à l'arrière de navires-laboratoires. »

La prospection marine par acquisition sismique se schématise de la façon suivante (*cf.* figure 1.18) :

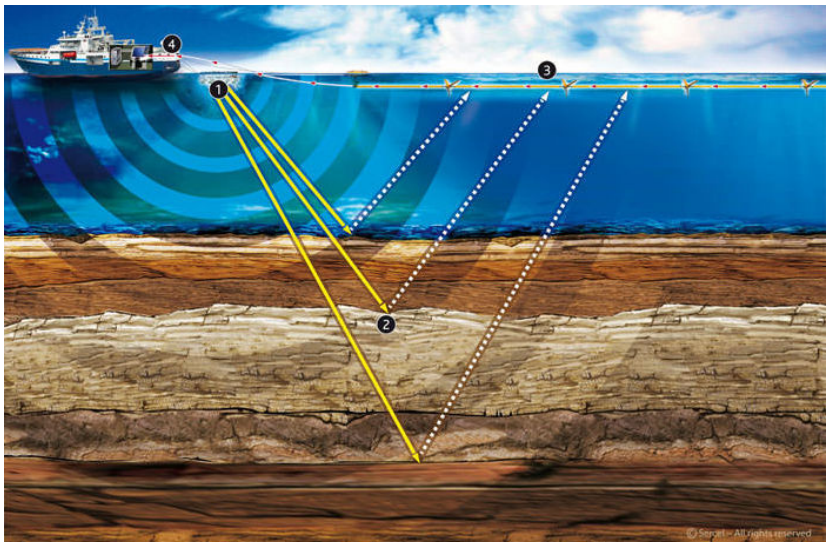


FIGURE 1.18 – Principe de la prospection marine (image provenant de l'entreprise SERCEL)

❶ Emission d'une énergie acoustique contrôlée et générée par un canon à air compressé (source sismique de forte amplitude sonore).

❷ L'énergie sismique se propage jusqu'au fond pour être ensuite réfléchie par les différentes couches géologiques.

❸ L'énergie réfléchie est détectée par l'ensemble des hydrophones ou accéléromètres.

❹ Le système d'acquisition de SERCEL enregistre les données obtenues qui sont généralement traitées par une autre société comme CGG, Ifremer ou autre pour être ensuite exploitées par les groupes pétroliers.

C'est dans cet environnement que se pose la question des conséquences éventuelles des tirs sismiques sur la faune et la flore marine. Un large éventail d'effets sur la faune marine exposée à des niveaux sonores accrus a été documenté [BOYD et collab., 2008; FORTESCUE et collab., 2005; SOUTHALL et collab., 2008; UNEP (UNITED NATIONS ENVIRONMENT PROGRAMM)]. Ces effets peuvent varier d'une légère modification du comportement à la diminution ou la perte des capacités auditives, et dans certains cas, aller jusqu'à des blessures graves ou la mort.

Pour limiter ces effets, en plus de la pollution et de la chasse, différentes réglementations [SOS GRAND BLEU, 2018] ont été mises en place pour préserver les cétacés. Cela se traduit par l'application de procédures concrètes lors de campagnes sismiques dont voici ci-dessous un exemple extrait de la campagne ANTITHESIS (ANTILLES THERMICITÉ SISMOGENÈSE). Cet exemple nous servira de cas générique et s'applique parfaitement au contexte des bateaux sismiques utilisés par SERCEL :

— *Nous ne pénétrerons jamais avec la source dans le sanctuaire des Mammifères Marins.*

— *Une zone d'exclusion d'un rayon de 500 m autour du navire*[6] *sera respectée.*

6. Plus précisément, la zone d'exclusion est autour de la source sonore et non du bateau

— *Toute incursion d'un mammifère à proximité de la zone d'exclusion de 500 m entraînera l'arrêt immédiat des tirs.*

— *4 Observateurs de Mammifères Marins se relaieront en permanence pour scruter l'océan depuis un point élevé (> 20m) permettant un horizon dégagé bien au-delà de la zone d'exclusion.*

— *Un système d'écoute passive PAM (Passive Acoustic Monitoring) sera actif 24h/24 pour aider à la détection d'éventuels mammifères.*

— *Une procédure de Ramp up (augmentation très progressive de la puissance des tirs) sera mise en œuvre à chaque démarrage de la source sismique pour prévenir les mammifères et leur laisser le temps de s'éloigner.*

— *1 seul bateau*[7] *naviguera en ligne droite sur des profils de plus de 200 km : éviter les effets de piège dus à plusieurs bateaux et faciliter l'évitement des mammifères.*

L'utilisation de bateaux sismiques a un coût très élevé pour l'entreprise. En conséquence, celle-ci veut avoir une grande confiance dans les détections de potentiels cétacés. Même si des observateurs sont présents pour scruter la surface de l'eau, ils ne peuvent travailler que le jour. De plus, même en plein jour, les conditions météorologiques peuvent limiter la visibilité et donc la détection d'éventuels mammifères marins. Un système PAM pour « écouter » les sons produits dans l'océan pourra alors aider à la détection d'éventuels mammifères marins indépendamment des conditions de visibilité.

La société SERCEL a développé son propre logiciel PAM : QuietSea. Il permet de voir la position des hydrophones par rapport à la position du bateau et, surtout, il réalise la détection et la localisation des cétacés (*cf.* figure 1.19) en étant complètement intégré à l'architecture multicapteurs de SERCEL (*cf.* figure 1.20). Le réseau de capteurs transmet ses données au bateau qui sont ensuite traitées par les algorithmes de QuietSea.

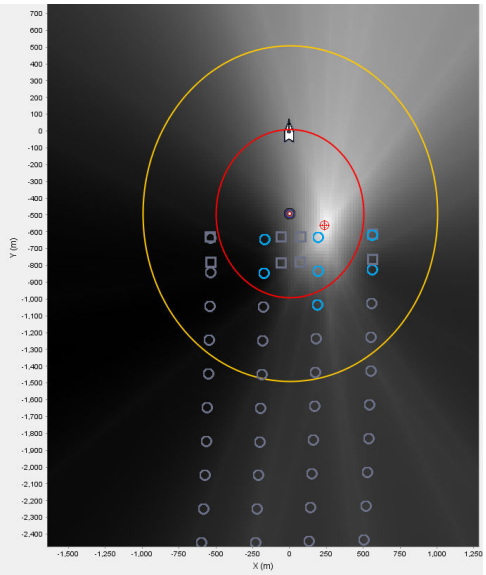


FIGURE 1.19 – Localisation de sources sonores avec QuietSea

FIGURE 1.20 – Architectures multicapteurs

Localisation d'un cétacé par QuietSea avec le système d'acquisitions multicapteurs de SERCEL (images provenant de l'entreprise SERCEL).

Sur la figure 1.19, après activation de la détection (capteurs activés en bleu), la source (cible rouge) est localisée avec une représentation de la probabilité de présence (rayons en niveau de gris). La zone de réglementation (500 m autour du canon à air) est représentée en rouge et la zone orange (1 km autour du canon à air) représente une zone d'alerte.

7. Cette dernière condition n'est pas toujours envisageable et il arrive que plusieurs bateaux soient présents en même temps pour la campagne.

Il faut noter que, comme les données sont traitées « *in sea* », alors les données basses fréquences sont transmises complètement alors que les données hautes fréquences doivent être compressées pour être transmises au bateau en un temps raisonnable (question d'échantillonnage et de bande passante). Au niveau des cétacés, cela signifie que les sons basses fréquences comme les vocalises produites par les mysticètes (données brutes pouvant être comprises entre 0 et 2000 Hz) seront plus faciles à traiter que les données hautes fréquences comme les sifflements des odontocètes (données compressées).

1.6 Problématique et démarche scientifique adoptée

1.6.1 Problématique

Afin d'anticiper les futures contraintes de réglementation, SERCEL cherche à étendre le périmètre fonctionnel de QuietSea en y intégrant des modules offrant la possibilité de faire de la reconnaissance. Cette anticipation devient nécessaire car certaines réglementations différencient d'ores et déjà les actions à réaliser en fonction des espèces. C'est dans ce contexte de continuité du logiciel QuietSea que s'inscrit cette thèse en lien avec SERCEL. Nous nous sommes principalement focalisés sur le cas d'étude des mysticètes. C'est la raison pour laquelle ce manuscrit ne contiendra pratiquement que des informations relatives à ces baleines sans nécessairement considérer le cas des odontocètes.

La problématique de la thèse peut être énoncée de la façon suivante :

> *« Mettre en œuvre un algorithme de reconnaissance des vocalises de mysticètes en environnement sismique. »*

En résumé, notre objectif principal est de créer un algorithme qui va permettre d'identifier les espèces à partir de leurs vocalises. Pour répondre à cet objectif, nous avons déjà présenté les mysticètes et leurs vocalises. A présent, nous allons explorer les différentes notions connexes à la problématique :

— La *reconnaissance* : plus particulièrement, quelles sont les méthodes pour faire de la reconnaissance de vocalises de mammifères marins aujourd'hui ? Quels sont également les verrous technologiques ?

— La *mise en œuvre d'un algorithme de reconnaissance* : il convient de définir une méthode qui soit la mieux adaptée à notre problématique, c'est-à-dire que nous devons mettre en avant l'ensemble des avantages et inconvénients des méthodes existantes pour développer la meilleure possible dans notre contexte.

1.6.2 La démarche scientifique adoptée

Nous proposons ici une discussion autour des relations entre les contraintes inhérentes à notre problématique et la démarche scientifique adoptée. Après avoir établi un état de l'art des méthodes de reconnaissance et après avoir identifié le nombre de classes (types de signaux) à reconnaître, nous aurons nécessairement besoin d'une base de données pour effectuer nos tests. Comme nous le verrons au chapitre 2, les données sont le fondement des méthodes de reconnaissance et de leur mise en œuvre. En particulier, les données sont « l'expérience du terrain » que nous allons transmettre au système. Même si la modélisation mathématique a fait de grandes avancées concernant la simulation de données, dans notre contexte, il ne peut y avoir de validation d'une méthode sans tests sur des données réelles.

En effet, un partenariat industriel implique la garantie d'un développement d'une méthode « qui produit des résultats » et donc, un suivi évolutif vers un objectif concret. Mathématiquement, nous devons pouvoir relier l'ensemble des paramètres de la méthode à des entités mathématiques mesurables afin que l'ensemble des traitements effectués soient facilement interprétables.

Financièrement parlant, les fausses alarmes, c'est-à-dire les cas où la méthode de reconnaissance retourne l'identification d'une structure dans le signal d'observation comme étant vraie

alors qu'en réalité il n'y a pas de vocalises et donc pas de mammifères marins, coûtent cher à l'entreprise.

Comme vue dans la partie 1.5, une fausse alarme implique, pour respecter la réglementation, d'arrêter les tirs de canon pendant 30 minutes (phase d'écoute), puis passé ce délai, de revenir au niveau de puissance du canon à air en augmentant l'intensité pendant une demi-heure encore. Cela représente 1h de dépenses (équipage, carburant, etc.) à cause d'une erreur du système (mauvaise détection et/ou mauvaise identification). C'est donc une composante majeure des contraintes liées à l'entreprise qu'il faudra obligatoirement prendre en compte. Par rapport à la reconnaissance de formes, cela implique les considérations suivantes :

- L'ensemble des motifs à reconnaître dans les signaux, c'est-à-dire la définition des classes, sera vu comme un espace dont les frontières sont floues. La notion de représentativité d'une classe sera définie de façon statistique. Cela revient à dire que la question « à partir de quels changements une vocalise n'est plus une vocalise ? » est un problème mal posé qui n'a pas de solution unique, mais peut être considéré comme un ensemble de solutions statistiques directement reliées au contexte applicatif. Dans notre cas, nous considérons qu'une classe est définie, de façon générale, par l'ensemble des informations structurelles observées dans les données annotées par des experts (informations intrinsèques comme la provenance des données et les observations physiques comme la durée moyenne, les évolutions fréquentielles, etc.) et des évolutions présentées dans la littérature (informations extrinsèques comme les informations géographiques et saisonnières de présence de mammifères marins, les évolutions à long-terme des vocalises).
- La gestion de la fausse alarme se fait en considérant le « complémentaire » du point précédent, à savoir que l'ensemble des fausses alarmes représente les mauvaises détections que la méthode identifie comme de vraies vocalises. Cela revient à considérer un ensemble infini de signaux indésirables ou inconnus qui représente exactement le complémentaire de l'union des ensembles des classes des signaux souhaités. Nous souhaitons alors pouvoir placer une frontière (un seuil) qui, associé à une mesure, nous permettra de séparer cet ensemble de signaux indésirables ou inconnus de l'ensemble des classes. Cela revient à construire des frontières, soit basées sur l'ensemble des classes à reconnaître, soit basées sur l'ensemble des signaux inconnus, soit les deux. Mathématiquement, l'identification de ces deux ensembles devra être bien définie. Il faut alors vérifier avant toute mise en œuvre que les classes à identifier sont différentes entre elles et également différentes de l'ensemble des signaux inconnus donc que du point de vue mathématique, il existe une frontière entre chaque ensemble et que ceux-ci soient « suffisamment disjoints ».
- Le même raisonnement que précédemment est à prendre en compte dans le cas de bonnes détections.

La combinaison des points précédents implique une méthode dynamique. En particulier, vu les propriétés acoustiques des océans, nous savons que l'ensemble des sons d'environnement est infini. Notre méthode doit donc s'adapter aux changements de base de données. Pour l'industriel, cela signifie que les garanties de résultats obtenues à une étape du développement de la méthode doivent être valables aussi à l'étape suivante, et ce, à chaque évolution de la méthode. La mise en œuvre de la méthode sera alors incrémentale.

En parallèle des considérations industrielles, il est nécessaire de considérer l'aspect académique de la thèse et notamment les contributions à la communauté bioacoustique. Pour ce faire, nous devons réellement prendre en compte et effectuer une analyse critique des méthodes employées afin de produire un travail qui puisse être également applicable aux problématiques bioacoustiques. De plus, il faudra s'efforcer de rendre notre démarche la plus reproductible possible afin de pouvoir fournir un moyen de comparaison par rapport aux futures méthodes proposées par la communauté.

Chapitre 2

La classification en bioacoustique

« S'il a été remarqué que quelques dauphins pouvaient reconnaître jusqu'à cinquante mots de notre langue, aucun humain n'a jamais pu comprendre un seul mot de la leur. »

— Carl Sagan

Sommaire

Afin d'éviter toute confusion, commençons par préciser que la classification (*clustering* en anglais) est souvent associée, par abus de langage ou par anglicisme, à la reconnaissance (*classification* en anglais). La tendance actuelle tend à dénommer l'ensemble des méthodes de classification et de reconnaissance comme appartenant à l'apprentissage automatique (*machine learning* en anglais). Dans ce chapitre, nous parlerons de *classification non supervisée* ou *apprentissage non supervisé* pour les méthodes de classification (ou méthodes descriptives), de *classification supervisée* ou *apprentissage supervisé* pour les méthodes de reconnaissance (ou méthodes prédictives), et de *classification semi-supervisée* pour les méthodes dites hybrides. En particulier, après avoir introduit de façon intuitive la notion générale de la classification, nous traduisons plus formellement notre problématique. Ensuite, nous présentons les avantages et inconvénients des approches proposées dans la littérature bioacoustique dans le cas de la reconnaissance des sons des cétacés. Enfin, nous mettons l'accent sur les problématiques dédiées aux données réelles et à la validation des algorithmes de classification supervisée.

2.1 La classification en général

Aujourd'hui il nous est encore difficile de définir globalement la classification notamment, car chaque communauté associe le terme « classification » au vocabulaire de son contexte scientifique. Nous trouvons ainsi aujourd'hui que la classification est liée aux méthodes d'apprentissage (*machine learning* et *deep learning* pour apprentissage automatique et apprentissage profond en anglais), à l'intelligence artificielle, à la vision par ordinateur (*computer vision* en anglais), à la reconnaissance de formes (*pattern recognition* en anglais), à l'optimisation, à la topologie mathématique, à la reconnaissance, aux modèles de prédiction, aux méthodes de segmentation, etc. Néanmoins, il est possible de séparer la manière de faire de la classification en deux approches principales en fonction du niveau d'informations disponibles sur les données et de l'objectif recherché.

2.1.1 La classification supervisée

La première approche consiste à vouloir remplacer (ou du moins seconder) un expert ou un superviseur. Par exemple, dans notre contexte, l'objectif est de réaliser automatiquement les tâches de reconnaissance et/ou de détection en identifiant chaque nouvelle observation, comme l'aurait fait un ou plusieurs experts. On parle alors de *classification supervisée* et plus largement d'*analyse discriminante*. Pour la validation de ces méthodes de classification, il est nécessaire d'avoir une base de données annotées. En résumé, l'idée principale est d'utiliser l'expérience, ou *vérité terrain*, donnée par un ou plusieurs experts du domaine considéré pour construire et valider l'algorithme de reconnaissance. La construction consiste à « apprendre » cette vérité terrain, c'est la phase d'apprentissage. La validation consiste à comparer les résultats proposés par l'algorithme avec la vérité terrain, c'est la phase de test. Nous reviendrons plus loin (*cf.* section 2.7) sur les questions de validation de méthodes d'apprentissage et sur les problématiques inhérentes aux bases de données.

2.1.2 La classification non supervisée

La seconde approche est plus complexe à résumer. Mathématiquement, on cherche à interpréter les données de façon géométrique (ou topologique) en se basant sur la notion de distance ou *similarité*. Cette notion de distance peut s'appliquer entre les données directement ou s'utilise entre des « sous-espaces » de données qu'on appelle l'ensemble de descripteurs ou *features* (pour caractéristiques en anglais). La distance ainsi définie va alors permettre de *partitionner* les données par rapport aux descripteurs choisis. On parle de classification non supervisée (*clustering* en anglais) ou plus largement d'analyse descriptive ou exploratoire. Nous proposons ici quelques problématiques liées à ces méthodes : la recherche de sous-espaces représentatifs (descripteurs pertinents par rapport au contexte) afin de réaliser, par exemple, de la segmentation (sélection de

zones d'intérêt) ; la visualisation (représentation schématique simple) et/ou l'accès rapide aux informations voulues par un ou plusieurs utilisateurs dans un contexte donné ; la génération d'hypothèses, c'est-à-dire le fait de proposer un comportement général à partir d'observations faites sur les données ; la simulation de données observées (problématique d'incrustation dans de l'image ou de la vidéo comme par exemple la synthèse de texture). L'idée ici est que les données ne sont généralement pas labellisées ni « classées » par un expert. Les enjeux sont alors, soit d'observer des données à des fins statistiques (proportion, comptage, etc.), soit de construire des classes pertinentes par rapport au contexte applicatif dans l'idée de faire de la classification supervisée.

2.1.3 La classification semi-supervisée

Pour finir, lorsque la base de données est partiellement annotée et qu'il n'y a pas d'expert pour terminer le travail d'annotation, la classification non supervisée peut être utilisée afin d'aider à compléter les annotations manquantes. On parle alors de *classification semi-supervisée* voire de détection ou d'apprentissage de nouveauté. Dans la pratique, qu'elle soit manuelle ou non, l'utilisation de la classification non supervisée, c'est-à-dire le fait de « regarder » les données, est fondamentale. Cette étape permet éventuellement de confirmer ou d'infirmer les propositions faites par un expert sur les annotations des données, en plus de vérifier si la tâche demandée est réalisable manuellement par un être humain.

2.2 Formulation mathématique de notre problématique

Dans notre contexte d'acoustique passive, comme nous savons *a priori* quelles données nous souhaitons reconnaître, notre méthode de reconnaissance correspond à de la classification supervisée. L'objectif est de donner de la *connaissance* (phase d'apprentissage) à un système pour qu'il devienne capable de faire de la *re-connaissance*. De façon plus générale, nous cherchons à transformer un ensemble de données en informations pertinentes. Ces informations sont pertinentes lorsqu'elles permettent de séparer les données de classes différentes et/ou de rassembler les données d'une même classe afin de mieux les identifier.

Formellement, à partir d'un ensemble de N couples de données labellisées $\mathscr{D} = \{(x_i, \ell_i)\}_{1 \leq i \leq \mathrm{N}}$ avec $x_1, x_2, \ldots, x_\mathrm{N}$ les données elles-mêmes ou un ensemble de descripteurs, et $\ell_1, \ell_2, \ldots, \ell_\mathrm{N}$ les labels associés (ou noms des classes de signaux), nous cherchons une fonction (un classifieur) f qui, après avoir effectué une phase d'apprentissage (construction de f) sur une base d'apprentissage $\mathscr{D}_\mathrm{A}$ (sous-ensemble de $\mathscr{D}$), soit capable de retourner une estimation du label ℓ_y (ou classe) associé à l'entrée (ou l'observation) y soit :

$$f(y|\mathscr{D}_\mathrm{A}) = \hat{\ell}_y \tag{2.1}$$

Ici, le « sachant $\mathscr{D}_\mathrm{A}$ » représente l'expérience ou la connaissance apportée au système pour qu'il puisse réaliser de façon automatique l'estimation du label ℓ_y associé à l'observation y. L'idée est toujours de vouloir remplacer l'expert du domaine. C'est cette expertise qui, associée à l'architecture du système, détermine la façon de représenter chaque classe.

2.3 Vue d'ensemble générale des étapes de la reconnaissance

Avant de rentrer dans les détails de la reconnaissance en bioacoustique sous-marine, nous proposons de décrire le processus général effectué sur les données (ici des signaux). Tout d'abord, les données brutes $\mathscr{D}$ sont séparées en deux ensembles disjoints, la base d'apprentissage $\mathscr{D}_\mathrm{A}$ et la base de test $\mathscr{D}_\mathrm{T}$. Les données d'apprentissage $\mathscr{D}_\mathrm{A}$ sont ensuite « transformées » ou *projetées* dans un espace représentatif des signaux d'intérêts (signaux sélectionnés par un expert). Dans la littérature, il s'agit principalement d'espaces temps-fréquence ou temps-échelle (*cf.* plus loin la section 2.5). A partir de cette *représentation* des données vient l'*extraction de descripteurs* ou *caractéristiques* ou *attributs*. Ces descripteurs peuvent être vus comme des « nouvelles variables d'observation » qui permettent de discriminer les signaux d'intérêt, par exemple des informations temps-fréquence.

Les données d'apprentissage $\mathscr{D}_\mathrm{A}$ résident alors dans un sous-espace $\mathscr{X}_\mathrm{A}$ uniquement constitué des valeurs (qualitative et/ou quantitative) des descripteurs des signaux d'intérêts. A partir de ces descripteurs et de la vérité terrain, la phase d'apprentissage prend fin avec le choix et/ou la construction de f (*cf.* plus loin section 2.7). Il existe deux façons générales de considérer f (*cf.* plus de détails dans la partie 2.7).

La première approche consiste à avoir f reposant sur un comparateur ou *une mesure de similarité*. C'est-à-dire que le système de reconnaissance a en mémoire un ou plusieurs représentants de chaque classe. Ces représentants peuvent prendre en compte les connaissances *a priori* des données ou se baser directement sur des données brutes. Le choix de f et la construction de ces représentants correspondent à la phase d'apprentissage. Comme nous le verrons plus loin, la reconnaissance correspond alors à identifier le représentant le plus ressemblant de l'observation courante (entrée du système). La procédure est alors liée à des problématiques d'estimation, l'idée est que le modèle de chaque classe soit le plus représentatif possible.

La seconde approche consiste à contraindre directement l'espace de représentation des données. Il s'agit de réduire l'espace d'entrée jusqu'à un espace de sortie qui soit généralement de dimension égale au nombre de classes à identifier. La construction de f consiste alors à résoudre un problème de minimisation d'une fonction dite *fonction de coût*. Cette fonction de coût permet d'identifier les erreurs commises par le système afin de le corriger ou de le mettre à jour. Dans ce cas, f est un système dont les sorties correspondent au nombre de classes attendues. Pour chaque élément x de $\mathscr{X}_\mathrm{A}$, f se met à jour pour converger vers le résultat voulu. Une fois l'apprentissage terminé, l'espace de sortie représente alors une *partition* (l'espace est partitionné c'est-à-dire, qu'il a des frontières associées à chaque classe) qui permet d'identifier chaque élément par sa position géométrique dans cet espace.

Enfin, vient la phase de test, la base de test $\mathscr{D}_\mathrm{T}$ subit les mêmes traitements que les données d'apprentissage pour devenir un ensemble de descripteurs $\mathscr{X}_\mathrm{T}$ de même nature que $\mathscr{X}_\mathrm{A}$. De cette façon, les performances de la méthode peuvent être évaluées en comparant les résultats proposés par f avec la vérité terrain donnée par l'expert.

Pour la suite, nous proposons de commencer par donner du sens à la définition de ce qu'est une *classe*. Puis, nous présentons les représentations des données d'observation proposées dans la littérature bioacoustique pour introduire les descripteurs utilisés. Ensuite, nous discuterons de l'architecture des systèmes utilisés, notamment leur mise en œuvre, à savoir si les méthodes considérées sont basées sur un ensemble de projections (par exemple les réseaux de neurones) ou plutôt sur une mesure de similarité. Enfin, nous discuterons des problématiques inhérentes aux bases de données réelles et à la validation des méthodes de reconnaissance.

2.4 La notion de classes de signaux

C'est l'expertise de l'homme et le contexte applicatif qui détermine la définition des classes. Néanmoins, les différentes façons dont les êtres humains (les experts du domaine considéré) procèdent pour identifier un élément d'une classe ne sont pas toujours identiques, en plus de ne pas être explicites. Mathématiquement, la définition d'une classe est alors considérée comme un problème mal posé, c'est-à-dire qu'il n'existe pas de solution ou que celle-ci n'est pas unique. Concrètement, dans notre contexte, nous proposons raisonnablement de définir nos classes de façon assez générale, comme étant l'ensemble des signaux qui sont de même nature et qui sont produits par les mêmes espèces de mysticètes (par exemple, la classe « Z-call » correspond aux signaux nommés Z-call par les bioacousticiens et désigne l'ensemble de tous les Z-call produits par les baleines bleues de l'Antarctique). Certes, nous pouvons constater que cette définition ne permet pas de définir clairement les frontières de nos classes, mais cela n'enlève en rien la validité des données labellisées par l'expert. Ainsi, même s'il nous est impossible de déterminer avec exactitude à partir de quelle(s) modification(s) un élément d'une classe n'en devient plus un, l'objectif reste de vouloir obtenir des résultats identiques à ceux proposés par les experts.

2.5 La représentation des données

Il existe de nombreuses représentations associées aux hypothèses faites sur le signal et aux informations recherchées. Une représentation peut être vue comme un espace dans lequel on souhaite décomposer ou projeter les données. D'un côté, la projection est susceptible de réduire la quantité globale d'informations (par rapport aux données brutes) et d'un autre côté elle permet de « mettre en valeur » les informations pertinentes contenues dans les données. Comme nous l'avons vu plus haut, dans notre contexte de classification supervisée, la représentation utilisée n'est pas une fin en soi. C'est généralement l'étape qui va précéder l'extraction des caractéristiques afin d'avoir, par la suite, un ensemble de descripteurs pertinents pour la mise en œuvre de la méthode de reconnaissance. Dans la littérature, l'ensemble des descripteurs est construit à partir de nombreuses représentations, nous présentons les plus courantes et les plus représentatives de celles utilisées en bioacoustique ci-dessous.

2.5.1 La représentation spectrographique

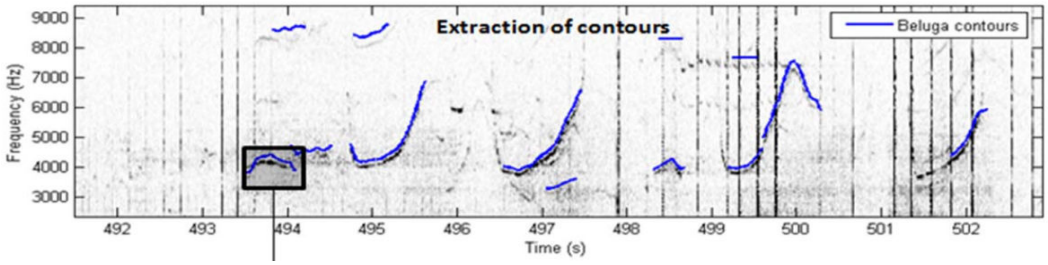


FIGURE 2.1 – Exemple d'utilisation du spectrogramme pour visualiser et extraire des sifflements de beluga extrait de [DELARUE et collab., 2010]

La représentation spectrographique (*cf.* figure 2.1) est la représentation de référence dans le milieu de la bioacoustique, nous pouvons la retrouver, entre autres, dans [ANDRÉ et collab., 2011; BINDER et HINES, 2012; MELLINGER et CLARK, 2000; MELLINGER et collab., 2011]. Comme vue au chapitre 1, cette représentation permet « d'accéder » aux informations « structurelles » des signaux considérés dans le plan temps-fréquence. Comme l'illustre la figure 2.1, les auteurs ont utilisé la représentation spectrographique de Fourier afin d'extraire des informations sur les sifflements de beluga (contours bleus sur le spectrogramme). Il est ainsi possible de conclure que le spectrogramme permet aussi bien de visualiser des sifflements très hautes fréquences (*cf.* figure 2.1) que basses fréquences (*cf.* spectrogrammes du chapitre 1). Nous pouvons résumer les avantages et les inconvénients d'utiliser la représentation spectrographique par rapport aux données brutes.

Avantages :

- ✓ Cette représentation est communément utilisée et connue du milieu bioacoustique.
- ✓ Le traitement du signal et de l'image sont envisageables sur le spectrogramme.
- ✓ Elle est intuitive au sens où elle est physiquement interprétable.
- ✓ Elle est raisonnablement robuste au bruit.

Inconvénients :

- ✗ Elle nécessite une calibration (paramètres du spectrogramme).
- ✗ La projection des données entraîne une perte d'informations.
- ✗ Elle est soumise au principe d'incertitude d'Heisenberg-Gabor, ce qui implique qu'il est impossible d'utiliser le même jeu de paramètres pour des signaux de nature différente.

2.5.2 La représentation cepstrographique

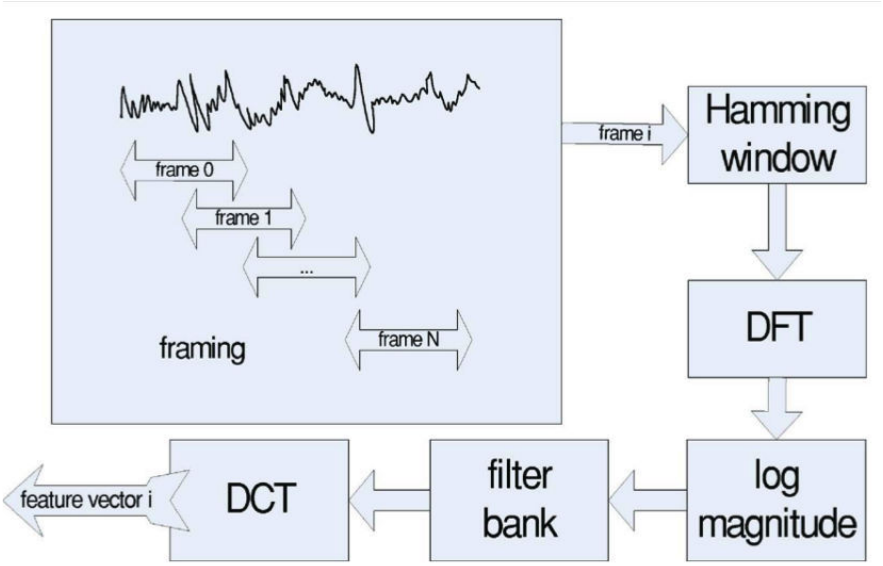


FIGURE 2.2 – Représentation cepstrographique extraite de [ROCH et collab., 2007]

La représentation cepstrographique est également assez présente dans le milieu bioacoustique. Nous pouvons notamment la retrouver dans [ROCH et collab., 2007], [ROCH et collab., 2011] et [HARLAND et ARMSTRONG, 2004]. Cette représentation se justifie par sa grande utilisation dans le traitement de la parole [FUKADA et collab., 1992]. Elle donne accès au signal source en effectuant une déconvolution et peut être utilisée par exemple pour identifier différents individus produisant les mêmes vocalises au sein d'une même espèce. De plus, il est possible d'utiliser les coefficients cepstraux directement comme descripteurs, ce qui permet de travailler dans un espace de dimension réduite par rapport aux données brutes. Cette compression des données est intéressante pour le traitement des signaux produits par les odontocètes et, par extension, aux traitements des phénomènes hautes fréquences qui impliquent de travailler avec une fréquence d'échantillonnage élevée et donc souvent sur des grandes bases de données. Même si ce contexte de vocalises d'odontocètes hautes fréquences n'est pas, *a priori*, lié à notre contexte de signaux basses fréquences des vocalises de mysticètes, cette représentation reste pertinente. Voici les avantages et inconvénients d'utiliser cette représentation.

Avantages :

- ✓ Cette représentation est déjà bien utilisée et connue dans le domaine du traitement de la parole.
- ✓ Elle permet de réduire l'ensemble des caractéristiques.
- ✓ Elle est raisonnablement robuste au bruit.

Inconvénients :

- ✗ Elle nécessite une calibration.
- ✗ La projection des données et la sélection de coefficients réduisent d'autant plus l'information.
- ✗ Elle ne suffit pas à représenter toutes les espèces en même temps.
- ✗ Elle est soumise au principe d'incertitude d'Heisenberg-Gabor.

2.5.3 La représentation en ondelette : le scalogramme

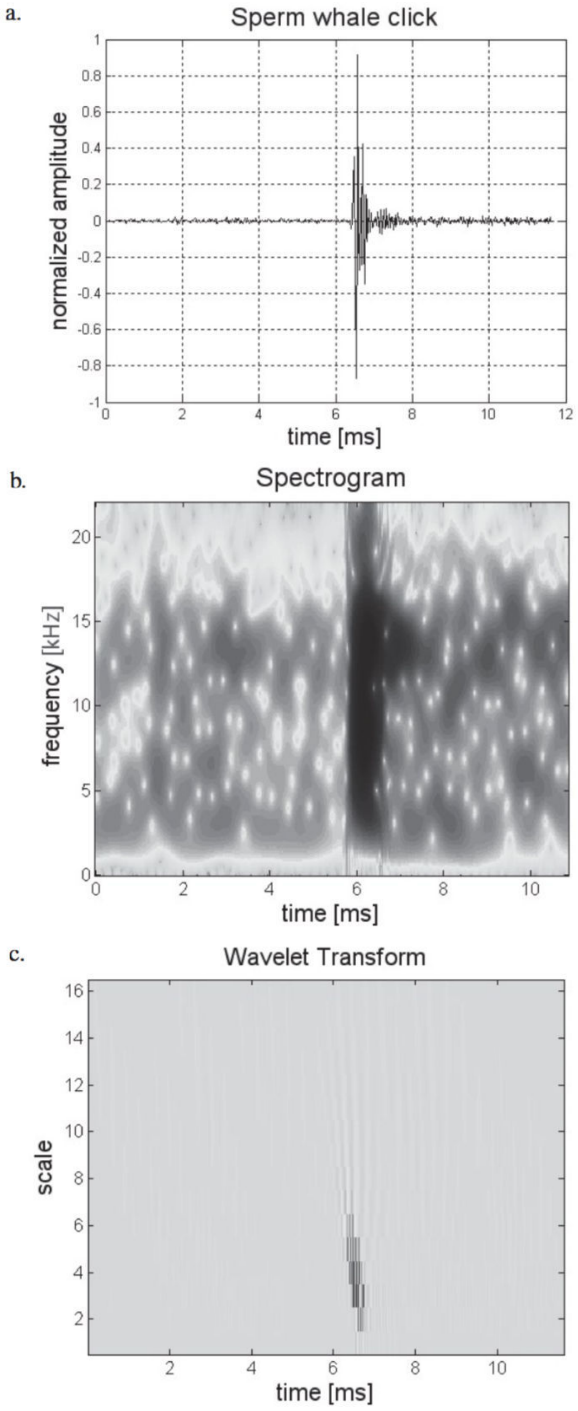


FIGURE 2.3 – Exemple d'utilisation du spectrogramme et des ondelettes (scalogramme) appliqués à la reconnaissance de clics de cachalot extrait de [LOPATKA et collab., 2005]. a. Représentation temporelle d'un clic de cachalot ; b. Représentation spectrographique ; et c. la transformée en ondelettes de ce clic.

Les ondelettes ont une place très importante dans le traitement du signal et des images depuis leur formalisation dans les années 1950. Comme mentionné brièvement plus haut, une lacune importante de la représentation spectrographique est d'être soumise au principe d'incertitude d'Heisenberg. En effet, ce principe empêche de choisir une résolution temps-fréquence qui soit aussi précise que l'on veut. En bioacoustique, cela se traduit par le fait que les paramètres utilisés pour observer des vocalises doivent s'adapter aux propriétés temporelles (aussi bien pour des vocalises « étalées » dans le temps que pour des vocalises impulsives) et/ou fréquentielle (aussi bien

pour des signaux très « localisés » en fréquence ou des signaux avec harmoniques que pour des signaux « épais » en fréquence). Afin de combler cette limitation, les ondelettes proposent d'utiliser plusieurs échelles d'observation qui vont permettre de faire de l'*analyse muti-résolution* en temps et en fréquence. Intuitivement, il est possible de « voir » les ondelettes comme une combinaison de plusieurs jeux de paramètres temps-fréquence qui vont ainsi représenter l'ensemble des propriétés du signal considéré avec plus de finesse qu'un unique spectrogramme. Prenons comme exemple la figure 2.3, qui concerne un clic de cachalot qui est très court dans le temps (quelques millisecondes visibles sur la représentation temporelle a.) tout en étant très étalé en fréquences (*cf.* représentation spectrographique b.). La représentation en ondelettes permet d'obtenir d'autres types de descripteurs. Par exemple, pour les petites échelles qui « capturent » des signaux très courts en temps alors ces échelles sont adaptées pour la reconnaissance de sons impulsifs (ici, l'énergie est structurée dans le carré temps-échelle des échelles 2 à 6 pour la figure 2.3). Comme dit dans [LOPATKA et collab., 2005], les ondelettes ont comme avantages d'être robustes aux bruits et adaptées aux applications temps-réel. Néanmoins, il est important de noter qu'une fois encore, utiliser une telle représentation implique de régler des paramètres supplémentaires par rapport aux données brutes. Voici ci-dessous les avantages et les inconvénients d'une telle représentation.

Avantages :

- ✓ Cette représentation a été bien étudiée en mathématiques ces dernières années.
- ✓ Elle est intuitive au sens du temps-échelle.
- ✓ Elle est raisonnablement robuste au bruit.
- ✓ Elle demande un faible coût en temps de calcul.

Inconvénients :

- ✗ Elle nécessite une calibration (choix des ondelettes, nombre d'échelles).
- ✗ La projection des données et la sélection de coefficients réduisent d'autant plus l'information par rapport aux données brutes.
- ✗ Elle est également soumise au principe d'incertitude d'Heisenberg-Gabor même si c'est dans une moindre mesure que la représentation spectrographique.

2.5.4 La représentation par la transformation de Hilbert-Huang

La représentation par la transformation de Hilbert-Huang est également utilisée en bioacoustique, mais de façon plus anecdotique, à notre connaissance. Elle est présente dans [ADAM, 2006b] et [ADAM, 2006a]. Cette transformée, proposée par [HUANG et collab., 1998], s'obtient en trois phases. Tout d'abord, les signaux sont décomposés en un nombre fini de signaux mono-composants appelés IMF (Intrinsic Mode Function) grâce à une méthode empirique appelée EMD (Empirical Mode Decomposition). Ensuite, la transformée de Hilbert est appliquée sur les IMFs et enfin, vient l'extraction de l'enveloppe complexe et de la fréquence instantanée. Ainsi, à l'aide de l'amplitude et la fréquence instantanées obtenues, il est possible d'avoir une représentation temps-fréquence (*cf.* figure 2.4). Comparativement aux représentations précédentes, la transformée de Hilbert-Huang (THH) est pilotée par les données et donc elle n'est pas conditionnée par un ensemble de paramètres. Elle n'est d'ailleurs pas soumise au principe d'incertitude d'Heisenberg-Gabor. La figure 2.4 nous permet d'apprécier les différentes résolutions des spectrogrammes de Fourier, d'ondelette et d'Hilbert-Huang. Comme pour les représentations précédentes, nous résumons ci-dessous les avantages et inconvénients de cette représentation.

Avantages :

- ✓ La THH est complètement adaptative (pilotée par les données) et ne nécessite pas de base particulière pour décomposer le signal considéré.
- ✓ La THH permet d'avoir une représentation temps-fréquence sans être conditionnée par un ensemble de paramètres (au contraire du spectrogramme de Fourier).
- ✓ Elle n'est pas soumise au principe d'incertitude d'Heisenberg.
- ✓ La possibilité d'utiliser peu d'IMFs permet d'envisager des applications en temps-réel.

Inconvénients :

- ✗ Même si la THH est bien adaptée pour des signaux non stationnaires, comme par exemple les signaux impulsifs, elle n'est généralement pas adaptée pour l'observation des harmoniques car elle ne sépare pas très bien des signaux proches fréquentiellement (comportement proche du comportement humain).
- ✗ Le signal considéré doit satisfaire des conditions de régularité, de variations lentes et de séparation fréquentielle. Sinon, il devient difficile de donner du sens aux IMFs.
- ✗ La THH est sensible au bruit, cela donne généralement lieu à des mélanges de mode (*mode mixing* en anglais). Les effets sont qu'une même caractéristique du signal observé peut se retrouver « partagée » entre plusieurs modes. De plus, le lissage des enveloppes des IMFs géré par l'EMD peut produire des artefacts ce qui rend alors difficile l'interprétation des IMFs.

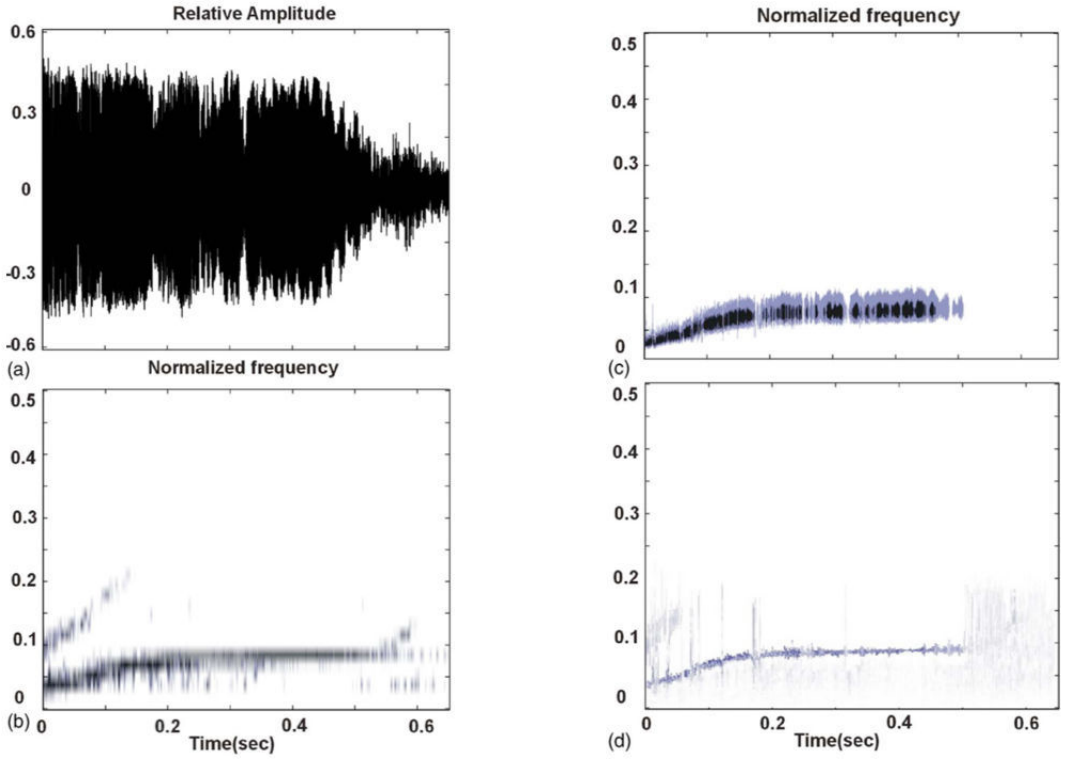


FIGURE 2.4 – Exemple de l'utilisation de la transformée de Hilbert-Huang pour une vocalise d'orques comparativement à d'autres représentations extrait de [ADAM, 2006a]. (a) Représentation temporelle du signal. (b) Spectrogramme de Fourier. (c) Spectrogramme par la transformée en ondelette de Morlet. (d) Spectrogramme par la transformée de Hilbert-Huang.

2.5.5 Les représentations abstraites

Les représentations que nous définissons comme « abstraites » sont pour nous les représentations qui ne sont pas reliées à des variables physiquement interprétables, mais plutôt des quantités mathématiques abstraites. Pour la reconnaissance, la réduction de l'espace de départ (données brutes) doit permettre de mieux séparer les données d'entrées (extraction de descripteurs) afin de les identifier automatiquement par leur position dans le nouvel espace (reconnaissance).

Dans le cas de [BINDER et HINES, 2012], la figure 2.5 permet de visualiser en deux dimensions un partitionnement (nouvelle représentation avec séparation des données) en couleurs obtenues par ACP (Analyse en Composantes Principales) et ADL (Analyse Discriminante Linéaire) à partir d'une base d'apprentissage (données validées par un ou plusieurs experts). Ici, sont considérés des signaux annotés et produits par un odontocète (le cachalot - *sperm whale* en anglais) et trois mysticètes (la baleine de minke - *Minke whale*, du Groënland - *Bowhead whale*, à bosse - *Humpback whale* et franche - *Right whale*) ainsi que les résultats de la reconnaissance par les cercles (les croix sont des centroïdes ou barycentres des « nuages de points colorés »). Les erreurs se voient par les cercles de couleurs qui ne sont pas dans la zone correspondante, par exemple des cercles rouges (de la partie « Humpback ») se retrouve dans la zone noire de « Bowhead ».

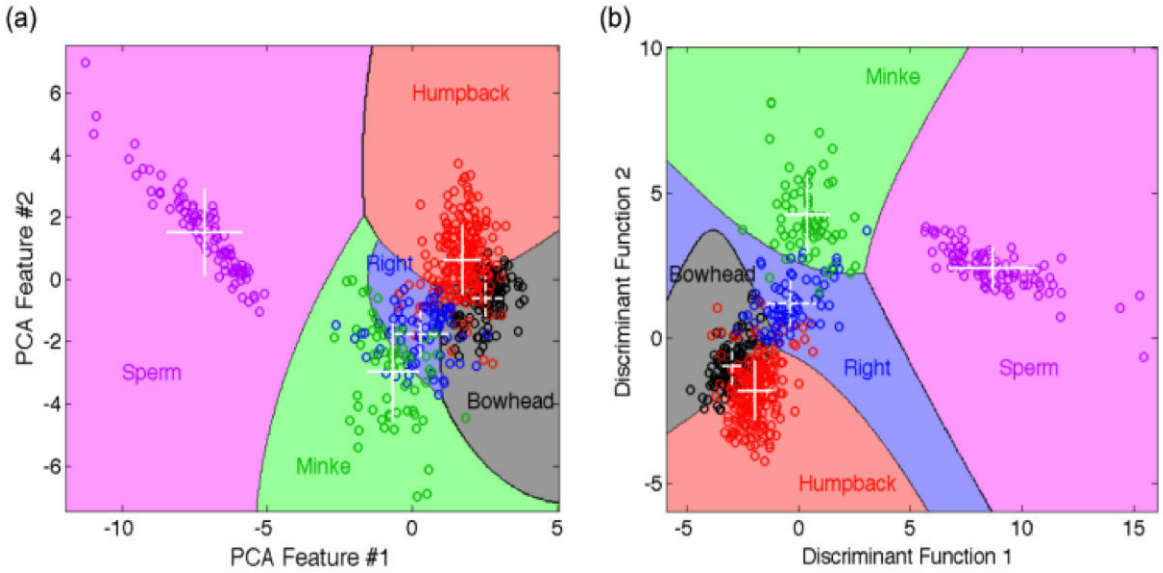


FIGURE 2.5 – Comparaison entre les représentations obtenues par ACP et par ADL extraite de [BINDER et HINES, 2012].

La lecture des deux graphiques de la figure 2.5 permet de voir que les axes sont des descripteurs abstraits (*features* 1 et 2) qui ne sont pas liés à une grandeur physiquement interprétable. Néanmoins, il est possible dans ces représentations, d'utiliser la géométrie (distance entre les points et les régions) pour noter que certaines vocalises partagent très certainement des propriétés structurelles communes. Cela semble être le cas des vocalises de la baleine à bosse (cercles rouges) avec ceux de la baleine du Groënland (cercles noirs) et ceux de la baleine franche (cercles bleus). De la même façon, nous pouvons noter que les signaux plutôt hautes fréquences produits par le cachalot (cercles magenta) se retrouvent bien séparés des signaux basses fréquences produits par les quatre mysticètes (cercles rouges, bleus, noirs et verts).

Une autre représentation abstraite présente dans la littérature est liée à l'utilisation des réseaux de neurones [ERBE, 2000; HALKIAS et collab., 2013; MELLINGER, 2004]. Afin de commencer à introduire les réseaux, dont nous parlons plus loin (*cf.* section 2.7), nous nous sommes inspirés du cours de [NG, 2011]. Commençons par présenter ce que représente un unique neurone artificiel (*cf.* figure 2.6). Un « neurone » est une unité de calcul qui prend comme entrée un vecteur $x =$

$[x_1, x_2, x_3, b]$ (avec b le *biais* égal à +1 ici) et en sortie $h_{W,b}(x) = g(W^T x)$ où $f : \mathbb{R} \mapsto \mathbb{R}$ est appelée la *fonction d'activation*. On trouve notamment la fonction sigmoïde pour la fonction g :

$$g(z) = \frac{1}{1 + exp(-z)}$$

Dans cet exemple, g correspond aux mêmes entrées et sorties définies par la régression logistique.

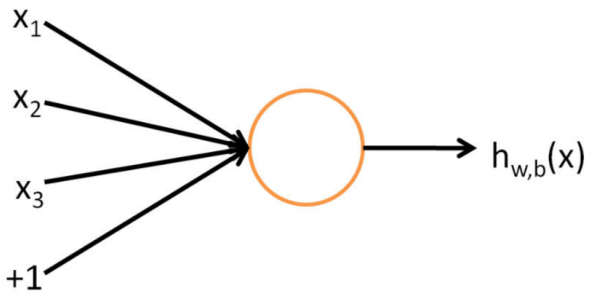


FIGURE 2.6 – Représentation d'un neurone artificiel extraite de [NG, 2011]

A présent, nous utilisons la figure 2.7 ci-dessous pour présenter le vocabulaire des réseaux de neurones.

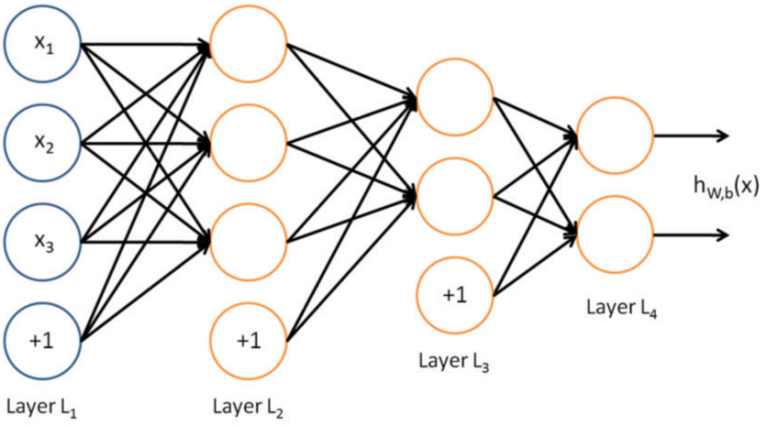


FIGURE 2.7 – Exemple d'un auto-encodeur extrait de [NG, 2011]

La couche (*layer* en anglais) d'entrée L_1 en bleue représente les données d'entrée. Une couche peut être vue comme un ensemble de neurones qui ne sont pas connectés entre eux. Même si sur le schéma, la couche L_1 est formée de cercles, elle ne représente pas des neurones mais, uniquement des entrées auxquelles s'ajoute un biais (ici +1). L_4 s'appelle la couche de sortie et les couches L_2 et L_3 sont des couches cachées. Entre les couches, sont visibles les connexions entre les neurones. Cet ensemble de connexions « inter-couches » correspond à un ensemble de poids ou coefficients. On retrouve ces coefficients dans la fonction $h_{W,b}$, où ils sont représentés par la matrice W.

Ce qu'il faut comprendre du point de vue des représentations, c'est que le passage d'une couche à une autre est une projection dans un espace de dimension égale au nombre de neurones de la couche considérée. Il est alors intéressant de contraindre l'espace d'entrée vers un espace plus petit qui va automatiquement séparer les données. Dans le cas de la reconnaissance, l'objectif est que la couche de sortie soit, comme pour les méthodes de la figure 2.5, une partition de l'espace d'entrée correspondant exactement au nombre de classes voulues. Dans ce cas, ce n'est pas la géométrie qui nous indique l'appartenance à une classe mais, la valeur la plus élevée de la couche de sortie. Le fait de contraindre un réseau s'appelle l'entraînement du réseau. Cet entraînement est possible sur une base d'apprentissage qui va permettre d'indiquer au réseau quelle est la sortie

souhaitée pour l'entrée correspondante. Ensuite, les poids de W sont mis à jour par minimisation d'une fonction de coût (algorithme de rétropropagation des erreurs du réseau) jusqu'à obtenir une séparation des classes satisfaisantes.

Du point de vue des représentations, chaque couche cachée est une nouvelle représentation abstraite des données d'entrées. Il est ainsi envisageable d'utiliser des données d'entrée venant déjà d'une autre représentation. C'est par exemple ce que propose [HALKIAS et collab., 2013] qui utilise en entrée du réseau des descripteurs extraits sur le spectrogramme de Fourier.

2.5.6 Discussion sur les représentations

Les représentations de type ondelettes, transformée de Hilbert-Huang... permettent d'extraire des informations de type temps-fréquence ou temps-échelle. Ce sont des « variantes » du spectrogramme de Fourier. Finalement l'objectif principal d'une représentation est d'accéder à un maximum d'informations non-observables directement (souvent avec un minimum de dimension et avec le moins de paramètres ou de calibrations possibles) suffisamment discriminantes pour faire de la reconnaissance. La représentation ainsi définie doit permettre d'accéder à des descripteurs pertinents. Nous avons vu plus haut que pour s'affranchir d'un certain nombre de paramètres, il faut privilégier les approches pilotées par les données. C'est d'ailleurs ce que propose les méthodes de réduction de dimension qui s'efforcent de séparer au mieux les données. Néanmoins, l'aspect empirique de ces approches rend souvent ardue l'interprétation des représentations obtenues (comme par exemple les couches cachées d'un réseau de neurones qui portent bien leur nom) et/ou des descripteurs associés à ces représentations. Dans ces conditions, c'est généralement l'expérimentation qui remplace l'impossibilité d'utiliser des justifications mathématiques solides.

Un autre constat est que, parmi l'ensemble des représentations utilisées en bioacoustique, aucune représentation ne suffit à elle seule à représenter la diversité des signaux bioacoustiques marins. En effet, nous avons envie de « mélanger » ces représentations et/ou de « concaténer » les bases de représentations afin de conserver les avantages de toutes les décompositions précédentes. Cela signifie qu'il faut peut-être quitter le côté « orthogonal » ou non-lié des bases, en conservant leur propriété génératrice, pour aller vers l'utilisation de *dictionnaires* ou de *bases surcomplètes*. Même si ces dictionnaires sont redondants et que la décomposition dans ces représentations n'est généralement pas unique, il est envisageable d'utiliser de telles représentations pour décomposer les signaux de vocalises de mysticètes, comme nous le verrons au chapitre 3.

2.6 Les descripteurs

Comme leur nom l'indique, les *descripteurs* vont permettre de *décrire* les données. Le concept de descripteurs est lié à la classification non supervisée. L'objectif est de trouver des sous-espaces de représentations des données dans lesquels les informations pertinentes sont accessibles. Cet objectif est très lié à la problématique de la *segmentation*, c'est-à-dire à la *détection* des zones où le signal d'intérêt ou un motif structuré est présent. Afin de présenter ces problématiques de segmentation et d'extraction de descripteurs de façon concrète, nous proposons de reprendre les représentations vues précédemment.

2.6.1 Segmentation et extraction de descripteurs à partir d'une représentation temps-fréquence

Nous commençons avec l'utilisation du spectrogramme de Fourier. Comme nous l'avons vu plus haut, cette représentation temps-fréquence permet d'extraire des descripteurs facilement interprétables, principalement liés à trois grandeurs physiques : le temps, la fréquence et l'énergie. Nous illustrons la segmentation et l'extraction de descripteurs à partir du spectrogramme avec deux exemples généraux [GILLESPIE, 2004] (*cf.* figure 2.8) et [BAUMGARTNER et MUSSOLINE, 2011] (*cf.* figure 2.10).

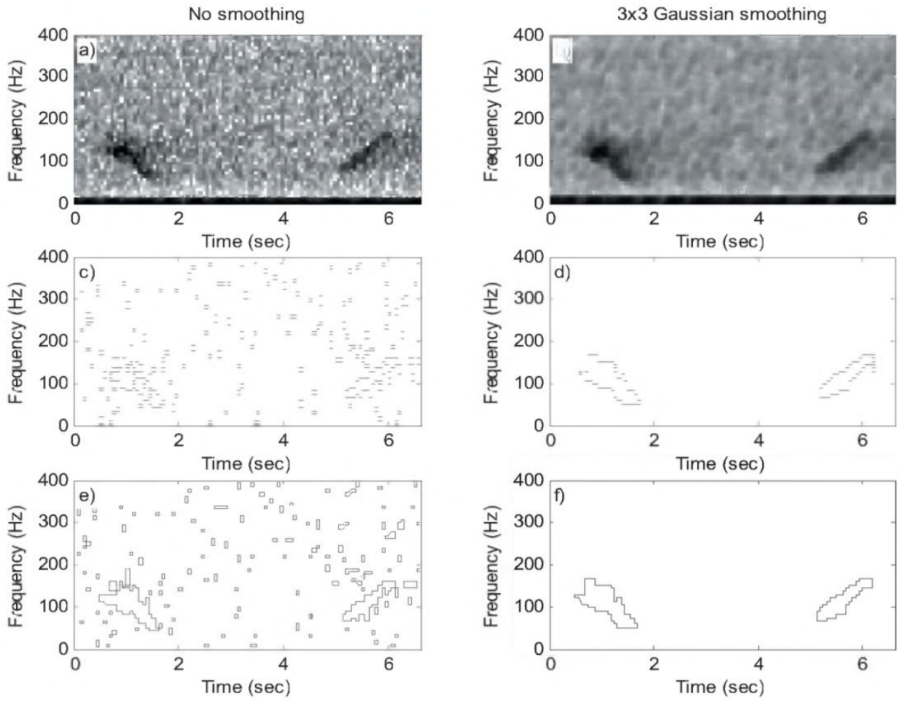


FIGURE 2.8 – Exemple de segmentation extrait de [GILLESPIE, 2004]. Détection des sons d'intérêt : a) Spectrogramme contenant deux cris dont l'un est un *upsweep* de baleine Franche ; b) Spectrogramme après un lissage gaussien ; c) Détection de contours sans le lissage gaussien ; d) Détection de contours avec le lissage gaussien ; Profils des sons sans lissage gaussien ; f) Profils des sons avec lissage gaussien

Dans cette première situation, [GILLESPIE, 2004] propose de traiter le spectrogramme comme une image (filtrage gaussien) et considèrent les signaux d'intérêt comme des contours. La recherche de contours revient à détecter les « discontinuités » de l'amplitude par rapport aux dimensions spatiales. Une discontinuité peut être vue comme une variation (un gradient) suffisamment élevée par rapport à l'ensemble des variations de l'image. Dans le cas présent, la segmentation a pour objectif de détecter et situer des régions ou « contours fermés » correspondants aux vocalises recherchées. Une fois cette étape réalisée, il devient alors possible d'extraire des descripteurs (*cf.* figure 2.9).

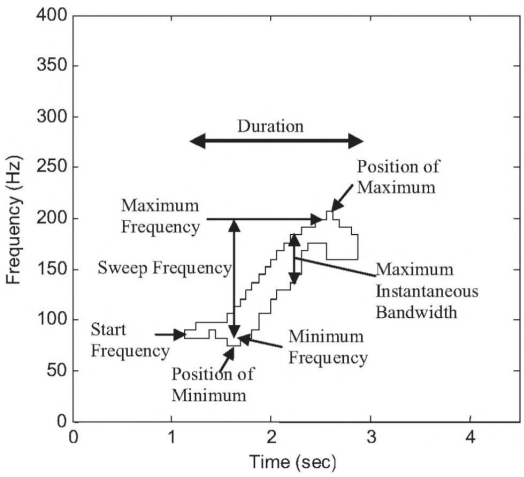


FIGURE 2.9 – Exemple de descripteurs extrait de [GILLESPIE, 2004]

Comme nous pouvons l'observer sur la figure 2.9, les descripteurs dépendent du temps (durée totale de la région segmentée) et de la fréquence (fréquence de départ, « épaisseur » de la région segmentée en fréquence, plage de fréquence entre le minimum et le maximum atteint). Comme

nous le verrons dans la section 2.7 suivante, ces descripteurs peuvent ensuite être utilisés pour apprendre un modèle représentatif de la classe de signaux considérés.

A présent, poursuivons avec un ensemble de traitements similaires proposés par [BAUMGARTNER et MUSSOLINE, 2011] (*cf.* figure 2.10).

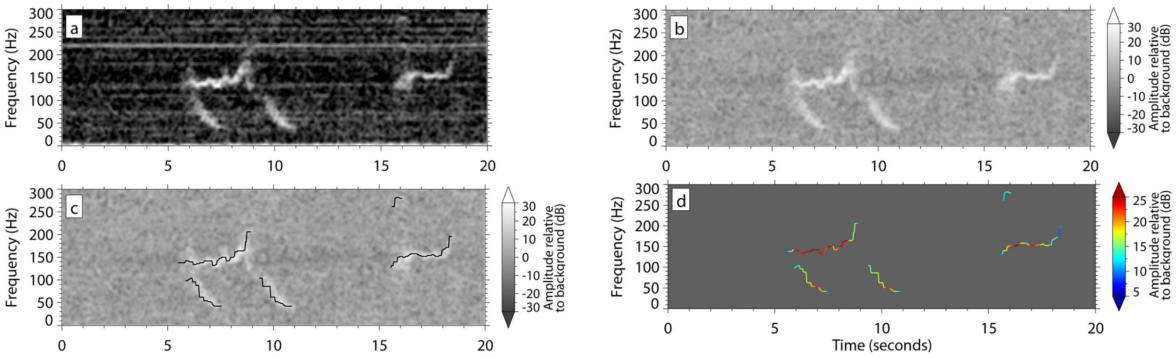


FIGURE 2.10 – Exemple d'utilisation du *pitch-tracking*. (a) Spectrogramme régularisé par un lissage gaussien (le même que celui utilisé par [GILLESPIE, 2004]). (b) Spectrogramme filtré par soustraction d'une moyenne glissante dans chaque bande de fréquence. (c) *pitch-track* ou « Poursuite de contours » (d) Même *pitch-track* avec les amplitudes représentées en couleurs.

La segmentation proposée par [BAUMGARTNER et MUSSOLINE, 2011] consiste à rechercher des « contours 1D » contrairement aux régions 2D proposées précédemment par [GILLESPIE, 2004]. Cette fois-ci, les descripteurs prennent également en compte l'énergie le long du contour extrait. De la même façon que les descripteurs vus plus haut, ces descripteurs peuvent ensuite être utilisés pour apprendre un modèle représentatif de la classe de signaux considérés.

Dans l'ensemble des représentations temps-fréquence, nous avons également vu la représentation en ondelettes. Cette représentation est plus robuste au bruit que le spectrogramme de Fourier [LOPATKA et collab., 2005] et permet d'utiliser les coefficients en ondelettes directement comme descripteurs. Dans ce cas, la segmentation est réalisée automatiquement avec la détection.

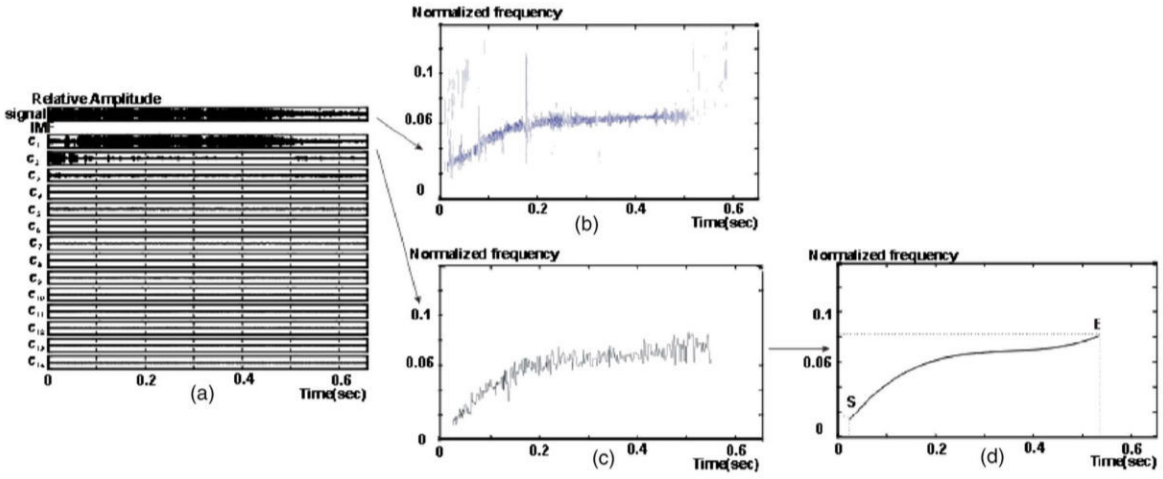


FIGURE 2.11 – Exemple de post-traitement permettant d'extraire des descripteurs pour les vocalises d'orques extrait de [ADAM, 2006a]. (a) Modes de décomposition empirique du signal original. (b) Représentation temps-fréquence du signal original. (c) Débruitage : représentation temps-fréquence du premier mode. (d) Extraction de descripteurs : interpolation polynomial d'ordre 3 (S et E : début - *start* et fin - *end* de la vocalise)

Pour poursuivre avec les descripteurs temps-fréquence, la transformée de Hilbert-Huang utilisée dans [ADAM, 2006a] et [ADAM, 2006b] est pertinente (*cf.* figure 2.11). Contrairement aux descripteurs proposés par [GILLESPIE, 2004] et [BAUMGARTNER et MUSSOLINE, 2011], ici les descripteurs permettent de décrire et de reconstruire complètement la courbe temps-fréquence observée. La segmentation dans ce cas est réalisée par un seuillage en fonction du SNR (*Signal to Noise Ratio*) observé. C'est en fait une détection qui va déterminer les zones temporelles de présence ou d'absence de la vocalise recherchée. Si, dans une autre situation, nous avions voulu utiliser les IMFs directement comme descripteurs, alors il est intéressant de noter qu'il n'y a pas de segmentation, les descripteurs sont alors extraits à partir du signal observé complet.

Pour finir avec les descripteurs temps-fréquence, il nous semble raisonnable de considérer les coefficients cesptraux. Nous présentons ci-dessous un spectrogramme de Fourier et les coefficients cesptraux correspondants (*cf.* figure 2.12).

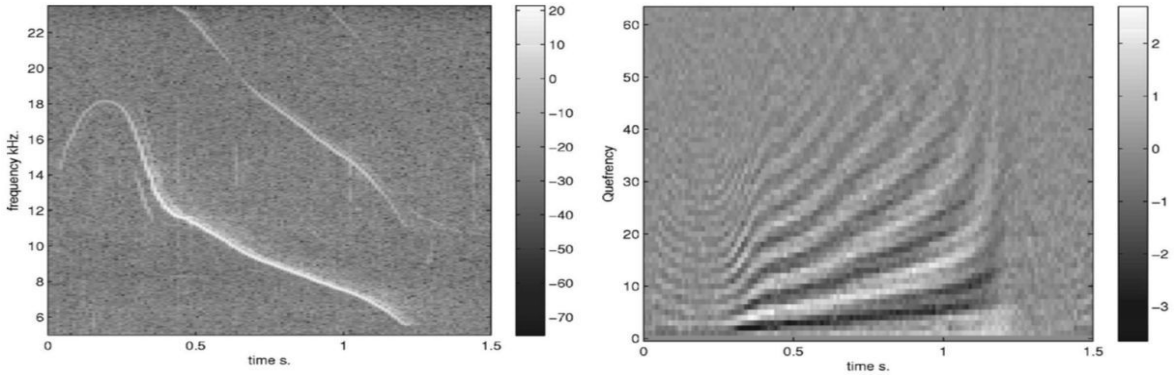


FIGURE 2.12 – Exemple d'un spectrogramme de Fourier de sifflements (figure de gauche) avec son équivalent avec 64 coefficients cesptraux (figure de droite) extrait de [ROCH et collab., 2007].

Dans cette situation, il est également intéressant de noter qu'il n'y a pas de segmentation, les descripteurs sont extraits du signal entier.

2.6.2 Segmentation et extraction de descripteurs à partir d'une représentation abstraite

Les représentations abstraites permettent de réduire l'espace considéré de façon à partitionner l'espace en régions correspondantes aux classes à identifier. Appliqué directement sur les données brutes labellisées, cela permet de s'affranchir de la segmentation. Néanmoins, en fonction des annotations (vérité terrain fourni par l'expert, généralement des carrés temps-fréquence correspondant aux zones de présence des vocalises considérées), il est parfois profitable d'utiliser des descripteurs extraits d'une représentation temps-fréquence ou temps-échelle afin d'améliorer la séparation des données de l'espace final. Dans ce cas, les représentations abstraites permettent alors d'affiner les descripteurs utilisés en sélectionnant et/ou en générant des descripteurs abstraits pertinents. Reprenons les outils mathématiques comme l'ACP et l'ADL utilisés par [BINDER et HINES, 2012]. Dans cet article, les auteurs ont au départ extrait 58 descripteurs. Ensuite, l'ACP permet de trouver des *composantes*, considérées comme des descripteurs abstraits qui vont permettre de séparer les données et ainsi obtenir le graphique de la figure 2.5. Enfin, un détecteur est utilisé pour réaliser la segmentation des zones temporelles d'intérêt.

Pour poursuivre avec les descripteurs abstraits, nous présentons le cas de l'auto-encodeur utilisé par [HALKIAS et collab., 2013] (*cf.* figure 2.13). Dans la configuration de ce réseau, l'objectif est d'approximer la fonction identité en cherchant à avoir la même sortie dans la couche L_3 que le vecteur d'entrée de la couche L_1, c'est-à-dire qu'on cherche à obtenir $h_{W,b}(x) \approx x$. La couche L_2 (couche cachée) va alors être la nouvelle représentation de la couche d'entrée. La fonction iden-

tité est généralement triviale, mais dans le cas de l'auto-encodeur, la couche cachée L_2, qui ne contient que peu de neurones, est une contrainte intéressante. En effet, elle « oblige » le réseau à devoir choisir quels éléments de la couche d'entrée sont les plus pertinents pour reconstruire l'entrée. En ce sens, le passage par la couche L_2 réalise une compression des données due à la réduction de la dimension. Ainsi, la couche L_2 est une représentation abstraite qui peut être utilisée directement comme descripteurs pertinents des données d'entrées. C'est ce qu'utilise [HALKIAS et collab., 2013] pour extraire des descripteurs pertinents sur des imagettes (ou patches) de spectrogrammes (*cf.* figure 2.14).

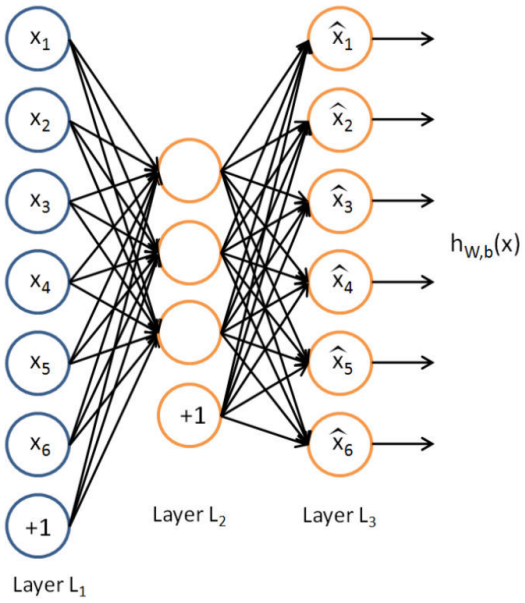


FIGURE 2.13 – Exemple d'un auto-encodeur extrait de [NG, 2011]

Nous pouvons constater qu'un premier ensemble de paramètres est choisi pour construire les spectrogrammes. Ensuite, est utilisé l'auto-encodeur pour générer des descripteurs abstraits associés à chaque ensemble d'imagettes de chaque classe. Enfin, ces descripteurs sont ensuite placés à l'entrée d'un réseau afin de réduire l'espace d'entrée jusqu'à obtenir en sortie le nombre de classes voulues.

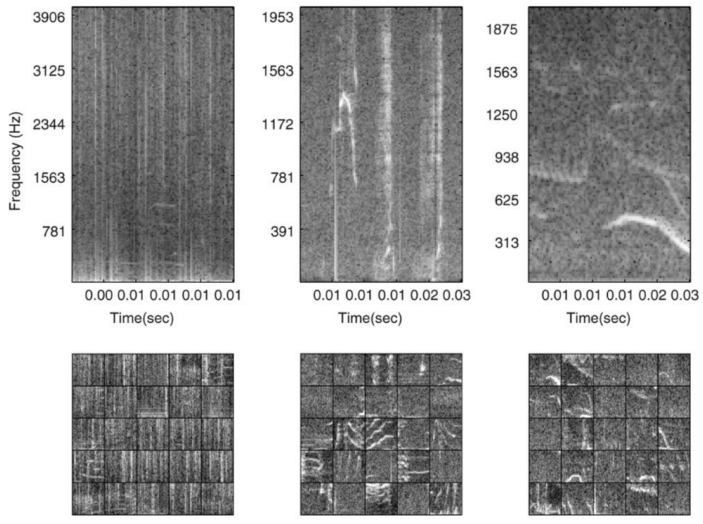


FIGURE 2.14 – Spectrogrammes contenant des signaux d'intérêts (rangée du haut) et 25 imagettes normalisées et mises à l'échelle par espèce (rangée du bas) extraits de [HALKIAS et collab., 2013]. De gauche à droite : Baleine franche de l'hémisphère sud, baleine à bosse et baleine du Groënland.

2.6.3 Discussion sur les descripteurs

Nous venons d'avoir un aperçu de l'univers conséquent des possibilités (en fait de taille infinie) entre le nombre de représentations des données avec leurs paramètres et la nature des descripteurs qui peuvent en être extraits. Comme nous le verrons plus loin, la représentation, la segmentation et les descripteurs sont déterminants pour mener à bien la reconnaissance. Dans la littérature, les descripteurs sont la résultante de plusieurs démarches souvent implicites et pas toujours bien justifiées. Tout d'abord, nous retrouvons la volonté d'imiter l'être humain, bien que le processus d'identification des signaux d'intérêt chez ce dernier ne soit pas explicitement défini et/ou démontré. Les descripteurs sont alors liés aux représentations temps-fréquence. Puis, il existe des descripteurs qui ont fait leurs preuves dans d'autres domaines, que ce soit par exemple l'utilisation des coefficients cepstraux en traitement de la parole ou les descripteurs des auto encodeurs en traitement d'image. Enfin, il existe des descripteurs qui sont la résultante de projections mathématiques abstraites. De notre point de vue, la notion de descripteur doit être reliée à la notion de classe et plus précisément à la notion de *représentativité* d'une classe. Quels que soient les descripteurs choisis, nous considérons que les caractéristiques extraites ne sont pas assez représentatives si elles ne permettent pas d'approximer, avec suffisamment peu d'erreurs, l'ensemble des signaux de la classe considérée. A nouveau, nous verrons plus loin, dans le chapitre 3, que les représentations parcimonieuses offrent de bonnes propriétés pour l'approximation de signaux structurés. Pour finir, rappelons que plus le nombre de paramètres et/ou d'hyper-paramètres est important, plus la validation de la méthode proposée devient complexe. En effet, une méthode qui nécessite n paramètres pouvant chacun prendre p valeurs impliquent de tester p^n situations et comme nous pouvons le constater, la plupart des couples « représentations-descripteurs » proposés par la littérature nécessitent un nombre de paramètres finalement assez nombreux. Cela nous incite à nous diriger vers des méthodes utilisant le moins de paramètres possibles. A présent, présentons ci-dessous la construction du classifieur.

2.7 Apprentissage et architecture des méthodes de reconnaissance

La construction ou le choix du classifieur f (*cf.* équation 2.1) s'appelle l'apprentissage. L'apprentissage se fait généralement sur une partie de la base de données qu'on décompose en deux parties : une base d'apprentissage et une base de tests. Une façon d'aborder l'apprentissage est de considérer la notion d'erreur de notre classifieur f, c'est-à-dire le nombre de fois où il se trompe de prédiction. L'erreur ainsi mesurée correspond à la probabilité d'avoir $\{f(y|\mathscr{D}_{\mathrm{A}}) \neq \hat{\ell}_y\}$. En ce sens, la construction du classifieur revient à un problème de minimisation. Néanmoins, la minimisation de l'erreur de prédiction sur l'ensemble d'apprentissage est un critère insuffisant, il faut également considérer la capacité d'apprentissage du système de reconnaissance. Dans le premier cas, si l'algorithme a trop de capacité d'apprentissage, on parle de sur-apprentissage (*overfitting* en anglais) et cela conduit à une mauvaise généralisation (mauvaise prédiction sur une nouvelle donnée). Plus simplement, c'est le cas où le système a « appris par cœur ». Cela veut dire que même si le système ne commet aucune erreur sur la base d'apprentissage (le problème de minimisation est résolu), face à une nouvelle donnée le système va très certainement « se tromper ». Il trouvera que la nouvelle donnée est très différente des données apprises. Une solution pour prévenir le sur-apprentissage est alors d'utiliser énormément de données afin de donner plus « d'expérience » au système. Dans le second cas, au contraire, si l'algorithme manque de capacité d'apprentissage, on parle de sous-apprentissage (*underfitting* en anglais) et cela conduira également à une mauvaise généralisation. Intuitivement, le système ne fait pas suffisamment la distinction entre les classes. Dans ce dernier cas, une solution consiste à augmenter la capacité d'apprentissage comme par exemple le nombre de descripteurs ou, dans le cas des réseaux de neurones, augmenter le nombre de neurones. Du point de vue de la validation, si la quantité de données le permet, il convient de faire intervenir un ensemble supplémentaire appelé *ensemble de validation croisée*, en plus de l'ensemble de test, pour travailler sur la généralisation de la méthode mise en œuvre (nous revenons sur cette problématique dans la section 2.8).

Commençons par expliciter l'apprentissage, c'est-à-dire l'obtention de f. C'est à cette étape de la reconnaissance qu'intervient la connaissance délivrée par l'expert, c'est-à-dire l'utilisation des annotations de la base d'apprentissage. L'apprentissage va alors dépendre à la fois de la représentation des données choisies ainsi que la nature de f. Nous proposons de séparer la démarche employée pour faire ce choix de f en bioacoustique en deux approches principales. La première approche consiste à faire de la reconnaissance basée sur une **mesure de similarité** tandis que la seconde approche consiste à faire de la **reconnaissance par réduction de dimension**.

2.7.1 Architecture basée sur une mesure de similarité

L'idée principale est de **comparer l'observation courante y à la connaissance $\mathscr{D}_A$ donnée au système**, en utilisant une mesure de similarité (par exemple la corrélation) ou une distance (ensemble des normes mathématiques). Dans cette configuration, f est fixée et l'apprentissage consiste alors à travailler uniquement sur les descripteurs des données d'entrée. La connaissance $\mathscr{D}_A$ donnée au système constitue l'ensemble des représentants des classes que l'on souhaite reconnaître. Ainsi, chaque classe de $\mathscr{D}_A$ peut être représentée par les données brutes, un ensemble de descripteurs moyens [BAUMGARTNER et MUSSOLINE, 2011] ou encore un modèle théorique [MELLINGER et CLARK, 2000] (*cf.* figure 2.15).

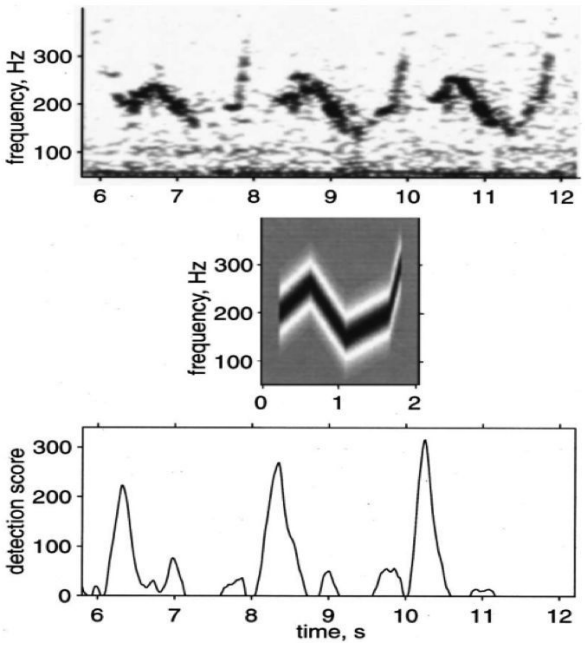


FIGURE 2.15 – Corrélation de spectrogramme extrait de [MELLINGER et CLARK, 2000].

Le principe est que l'observation et les classes apprises par le système soient dans un même espace de représentation afin d'être comparables. De cette façon, la distance entre l'observation y et le modèle (ou les données brutes) qui est la plus petite (ou la plus grande en fonction de la mesure utilisée) permet d'associer l'observation au label (à la classe) correspondant. Dans la littérature bioacoustique, les méthodes les plus utilisées sont les méthodes dites de corrélation de modèle ou d'extraction d'attributs [BITTLE et DUNCAN, 2013]. A titre d'exemple, nous présentons le modèle imaginé par [MELLINGER et CLARK, 2000] (*cf.* figure 2.15). Dans ce cas, les données entrantes sont comparées (distance ou corrélation) avec le modèle obtenu par l'étude des caractéristiques (méthodes de *Template matching* en anglais).

Voici ci-dessous, les avantages et les inconvénients de ce type de méthode :

Avantages :

- ✓ La méthodologie est généralement **simple à comprendre et à reproduire**.
- ✓ Les résultats pour un modèle ne dépendent *a priori* pas du nombre total de modèles choisis. Par construction, la méthode finale sera **dynamique et modulable** (il est en effet possible d'ajouter et d'enlever des classifieurs [modèle par classe] pour faire la reconnaissance).
- ✓ Elle est **intuitive** au sens où la reconnaissance par comparaison, notamment en décrivant les motifs à reconnaître, est proche du comportement humain et de ses différents niveaux d'expériences.
- ✓ La méthode discrimine les observations uniquement par rapport à ce qu'elle « connaît ». Ainsi, une observation « loin » (au sens de la distance choisie) de toutes les classes apprises correspond probablement à une observation inconnue. En conséquence, il est souvent possible de fixer un seuil sur la mesure (ou notre distance) choisie qui permet de **gérer les classes inconnues et plus particulièrement les classes de bruits**.

Inconvénients :

- ✗ Dans le cas d'une représentation choisie, la méthode **nécessite une calibration** (paramètres de la représentation) pour l'apprentissage comme pour le test.
- ✗ En plus de la représentation choisie, la construction d'une modèle par classe peut être **extrêmement fastidieux**.
- ✗ Dans le cas d'un modèle (ou un ensemble de descripteurs) par classe, celui-ci est généralement l'unique représentant de la classe considérée. C'est-à-dire que l'expérience donnée au système ne permet *a priori* pas de reconstruire l'observation, ce qui, par nature, est une preuve de **sous-apprentissage**.
- ✗ Un modèle n'est généralement pas dynamique et ne prend donc pas en compte les éventuels changements de structures des vocalises au fil des années. (C'est le cas par exemple du déplacement en fréquence des vocalises de baleine bleue [McDonald et collab., 2009])
- ✗ Même si les données brutes ont l'avantage de contenir le maximum d'informations, le fait de les avoir en mémoire dans le système augmente nécessairement la complexité de la méthode choisie.
  - ➔ Par exemple, si le système contient 3 classes à compter de 1000 représentants par classe, alors pour chaque observation il doit effectuer 3000 calculs de distance avant de pouvoir trouver la classe correspondante.

2.7.2 Architecture basée sur de la réduction de dimension

La seconde approche consiste à faire de la reconnaissance directement par la représentation utilisée. Plutôt, que de construire un modèle, l'idée est de **contraindre directement l'espace de représentation lui-même**. f devient alors l'architecture même du système utilisé, généralement un ensemble de représentations abstraites et sa construction revient à résoudre un problème d'optimisation. C'est le cas par exemple de l'ACP (idée de meilleures composantes discriminantes) et l'ADL (maximisation de la variance inter-classe et minimisation de la variance intra-classe) vues plus haut (*cf.* section 2.5 figure 2.5). Afin d'illustrer au mieux cette approche, nous développons le cas représentatif des réseaux de neurones (*cf.* figure 2.16), où f est une fonction de coût et où sa construction consiste à pénaliser un ensemble de coefficients de projection (on parle aussi de minimisation du risque bayésien empirique).

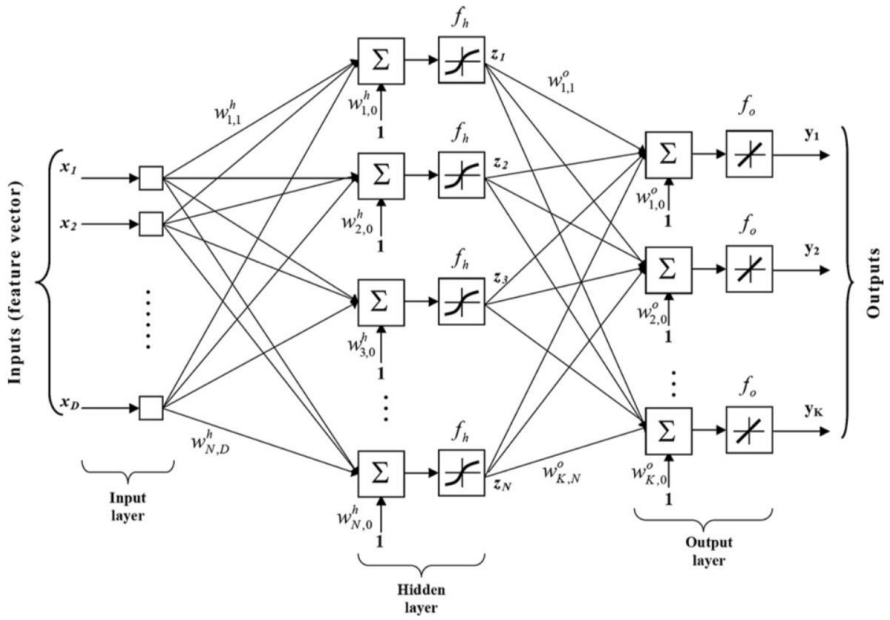


FIGURE 2.16 – Exemple d'un réseau de neurones (perceptron multi-couches) appliqué à la reconnaissance de vocalises de baleine bleue extrait de [BAHOURA et SIMARD, 2010]. Le réseau est caractérisé par D entrées, une couche cachée de N nœuds et K sorties. $w^h_{i,j} et w^0_{j,i}$ sont les connections pondérées de la couche cachée et de la couche de sortie respectivement.

A partir d'un nombre de sorties fixées (correspondant au nombre de classes à reconnaître), l'apprentissage est réalisé de la façon suivante, pour chaque entrée donnée au système, la sortie est comparée au résultat attendu (labellisation de l'expert). A chaque erreur commise par le système, nous corrigeons (mise à jour) les coefficients du réseau. C'est donc l'architecture du système de reconnaissance qui compte. Pour la phase de test, l'observation courante y est donnée en entrée du système et sa structure est projetée dans les différents sous-espaces du réseau jusqu'aux derniers sous-espaces représentants les différentes classes. Le sous-espace ayant reçu le plus d'informations détermine l'appartenance à la classe correspondante. Le système ne mémorise donc pas de descripteurs ni de données mais mémorise uniquement les coefficients de projection « appris » lors de la phase d'apprentissage. En entrée, il est également possible de prendre des descripteurs plutôt que les données brutes (*cf.* [HALKIAS et collab., 2013]). Voici ci-dessous les avantages et inconvénients de ce type d'approche.

Avantages :

- ✓ Ce type d'architecture est **intuitive** au sens où elle tend également à suivre un raisonnement humain. A chaque erreur commise, la bonne réponse est donnée au système afin qu'il « converge » vers une réponse satisfaisante, un peu comme le fait un être humain qui apprend une nouvelle aptitude à force de pratique et de correction.
- ✓ La phase de reconnaissance (phase de test) est **généralement très rapide** car elle ne met en jeu que des projections (produits matriciels) qui ont une **complexité très faible**.
- ✓ Au même titre qu'utiliser une distance, le réseau attribue une confiance associée à chaque classe possible. Ainsi, si chacun des coefficients de confiance associés aux classes possibles n'est pas très élevé alors il est probable que le réseau ne connaisse pas la classe de l'observation d'entrée. Néanmoins notre expérience des réseaux de neurones nous encourage à penser que cette situation n'arrive généralement pas et que, même dans l'erreur, les réseaux ont une grande confiance dans leurs décisions. Cela vient de l'apprentissage qui fixe généralement un taux d'erreur très faible et « force » le réseau à être « sûr de lui ».

Inconvénients :

- ✕ Il est impossible de prédire si le système va être capable de trouver des descripteurs implicites pertinents pour faire la reconnaissance. C'est-à-dire qu'il faut généralement **tester plusieurs architectures** afin de déterminer subjectivement le potentiel de reconnaissance d'un réseau.
- ✕ De la remarque précédente découle le fait que pour optimiser les résultats de reconnaissance cela va nécessiter de **tester beaucoup d'hyper-paramètres** (nombre de couches, nombre de neurones, régularisation des coefficients de projections,...) selon une heuristique basée sur une expertise, jusqu'à trouver une combinaison satisfaisante, si elle existe.
- ✕ Le réseau doit établir de lui-même les descripteurs pertinents lors de la phase d'apprentissage. Ce peut être un avantage car il n'y a pas à s'occuper des descripteurs et en même temps deux problèmes se présentent :
  - — Le **sur-apprentissage** : le réseau a beaucoup de neurones dans ses couches cachées et possède ainsi une grande capacité d'apprentissage (peu de réduction de la dimension de l'espace de départ). La solution consiste alors à utiliser beaucoup de données afin de prévenir le réseau d'apprendre « par cœur ».
  - — Le **sous-apprentissage** : la solution est d'accroître la capacités d'apprentissage (augmenter le nombre de neurones par couches intermédiaires) pour l'éviter et de la même façon que précédemment, devoir utiliser beaucoup de données.
- ✕ Les résultats de reconnaissance sont conditionnés par l'architecture du réseau. En conséquence, **la méthode n'est ni dynamique ni modulable**. C'est-à-dire qu'elle implique de remettre à jour tous les coefficients du réseau à partir du moment où la sortie (les classes considérées) est modifiée. De plus, le fait d'ajouter une nouvelle classe ne permet pas de garantir une continuité des résultats obtenus avec le nombre de classes précédentes.
- ✕ Dans la littérature bioacoustique, notamment dans [HALKIAS et collab., 2013], les réseaux de neurones utilisent généralement **une classe de bruit** pour gérer les fausses alarmes. Toutefois, ce n'est pas une bonne idée dans notre situation car cela reviendrait à devoir « montrer » à notre système l'ensemble des bruits réels existant en environnement marin, ce qui est impossible car **l'ensemble des signaux de la classe de bruit est infini.**

2.7.3 Discussion

Nous devons prendre en compte dans notre contexte, le fait que la méthode de reconnaissance est appliquée sur un bateau et assistée (ou non) par un opérateur en temps-réel. Elle doit donc être modulable afin de pouvoir sélectionner facilement les signaux à reconnaître ou non en fonction des missions organisées et de la répartition géographique des mysticètes correspondants. De plus, à notre connaissance, nous dénombrons environ une soixantaine de vocalises différentes réparties sur 17 espèces de mysticètes. Il n'est raisonnablement pas possible de traiter l'ensemble de ces vocalises en une seule fois (par manque de données par exemple). En conséquence, la méthode doit être simple d'utilisation. Ceci implique qu'elle doit avoir peu de paramètres et que ces paramètres doivent être faciles à choisir (méthode robuste). Pour finir, il faut rappeler que les campagnes sismiques impliquent un gros budget et qu'en cas de détection d'un mammifère marin dans le rayon des 500 m réglementaires autour du canon à air, la prospection doit parfois s'arrêter pour ne reprendre qu'une heure plus tard (30 min de *pre-watch* plus 30 min de temps de chauffe du canon à air). La gestion des fausses alarmes (classe de bruit) est donc essentielle. Pour toutes ces raisons, nous nous plaçons dans le cas des méthodes de reconnaissance utilisant une mesure de similarité qui permettent de travailler de façon incrémentale, c'est-à-dire que l'apprentissage

validé pour un nombre de classes fixé n'est pas à refaire en cas d'ajout de nouvelles classes (ou de retrait d'ancienne classe). De plus, une telle méthode peut également être utilisée à des fins statistiques, comme par exemple le comptage de vocalises dans des fichiers enregistrés.

2.8 La validation des méthodes de reconnaissance

2.8.1 La qualité d'une méthode proposée

La phase d'évaluation répond à une problématique essentielle : la qualité. Que ce soit pour évaluer la qualité d'un travail de recherche et/ou celle d'un travail industriel, l'objectif est de donner un **degré de confiance** associé au travail effectué. Ce degré de confiance nécessite de mettre en place un système d'évaluation qui assure que la méthode proposée répond bien aux besoins (ou spécifications) scientifiques et/ou industriels. De façon générale, nous pouvons décomposer la phase de test en trois entités : les données utilisées, la ou les méthodes proposées et les performances mesurées (*cf.* figure 2.17).

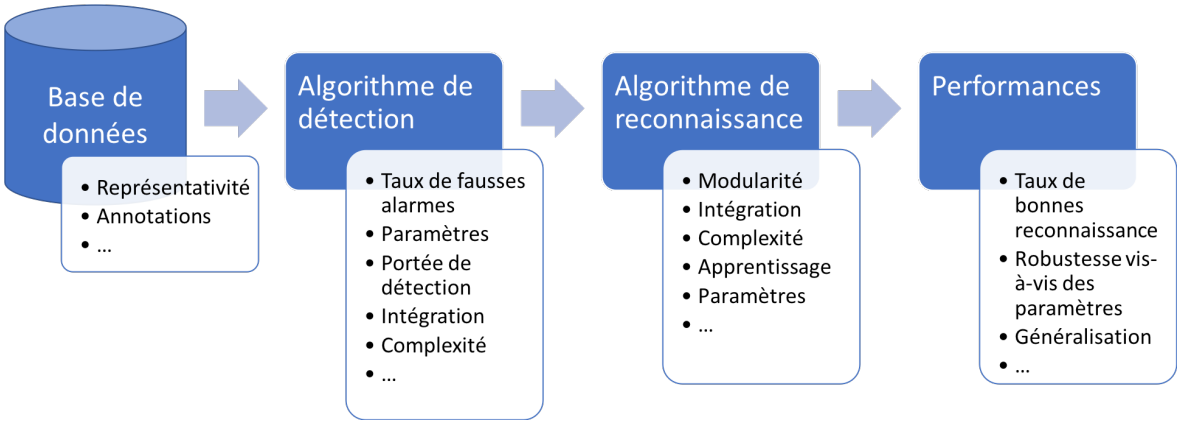


FIGURE 2.17 – Exemple général où chaque processus (Base de données, détection, reconnaissance et performances) est associé à un ensemble de caractéristiques ou critères de performances. Ces caractéristiques peuvent ensuite permettre de définir des spécifications comme par exemple « avoir un taux de fausse alarme de 1% ». Enfin la validation de ces spécifications permet de donner un degré de confiance général associé au système global.

Dans notre contexte, notre objectif est de remplacer (ou de seconder) un expert. Il s'agit dès lors de mettre en œuvre une méthode de reconnaissance qui, comme l'expert, soit capable de détecter et de reconnaître les signaux d'intérêt dans de nouvelles données, annoter ces nouvelles données et donner un indice de confiance par rapport aux identifications proposées. Le fait d'identifier de nouvelles données s'appelle la *généralisation* ou la capacité de la méthode à généraliser. Afin de mesurer la généralisation, l'idée est de mettre la méthode « en situation », c'est-à-dire de la confronter au contexte. Voici pourquoi il est nécessaire de travailler avec des données représentatives de ce contexte d'application. Deux types de données sont alors envisageables : les données simulées et/ou les données réelles. Dans les deux cas, ces données sont annotées (vérité terrain donnée par un ou plusieurs experts).

A notre connaissance, il n'y a pas, dans la littérature bioacoustique, de validation des performances de détection et/ou de reconnaissance effectuée uniquement sur des données simulées. Cela ne signifie pas pour autant qu'il n'existe pas la possibilité de créer un ensemble de données simulées représentatif. Cependant, comme il n'existe pas de modèle physique simple à utiliser, la communauté bioacoustique privilégie l'utilisation des données réelles.

2.8.2 Le besoin de références

L'utilisation de références (ou de *benchmark* en anglais) correspond aux problématiques de standardisation qui sont liées à la reproductibilité et donc à la productivité. Plus précisément,

pour un degré de confiance égal (qualité égale), il est pertinent de se demander quelles méthodes donnent les meilleurs résultats par rapport à notre contexte. Commençons par présenter le cas où le processus de validation est standardisé (*cf.* figure 2.18).

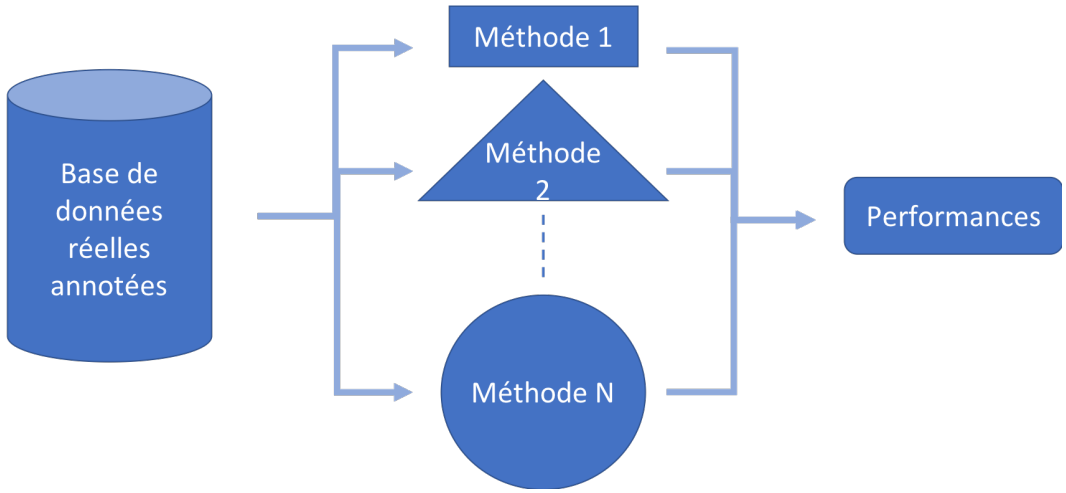


FIGURE 2.18 – Le cas où le processus de validation est standardisé.

Les problématiques de validation industrielles et de recherches étant très proches, nous proposons de présenter la situation dans le contexte de la recherche sans perte de généralité. Dans le cas d'un processus standardisé, cela implique que les données réelles annotées sont facilement accessibles et validées par la communauté bioacoustique. C'est-à-dire que ces données sont considérées comme représentatives du contexte considéré et n'ont plus besoin d'être présentées. De même, l'ensemble des performances est validé par la communauté et se suffit à lui-même. Ainsi, lorsque nous souhaitons publier notre méthode, il suffit simplement de citer la base de données utilisée et d'effectuer les tests proposés. De cette façon, chaque méthode est comparable et nous sommes capables de rapidement situer la qualité de notre travail par rapport aux autres méthodes. Nous sommes dès lors dans la situation la plus productive possible.

A présent, présentons le cas où aucun processus de validation n'a été défini et donc n'est pas validé par la communauté (*cf.* figure 2.19).

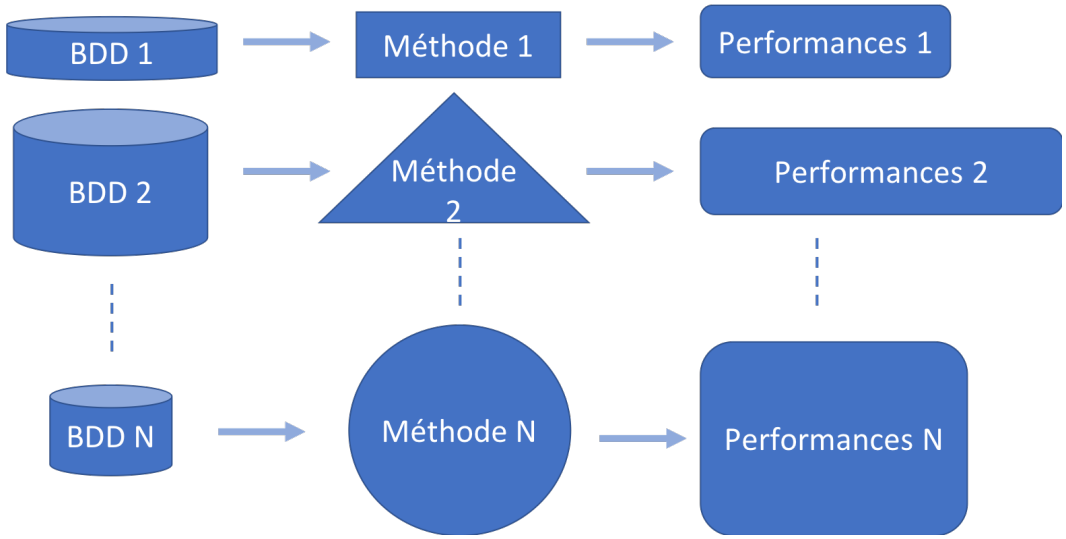


FIGURE 2.19 – Le cas où le processus de validation n'est pas standardisé. (BDD : Base De Données.)

Cette fois-ci, les bases de données utilisées sont différentes pour chaque méthode proposée. De même, elles ne sont pas facilement accessibles. Néanmoins, dans la situation présentée ici, ces N bases de données sont publiées c'est-à-dire qu'elles sont considérées comme des données

exploitables par la communauté. Après, au contraire du cas standardisé où la représentativité est établie pour l'ensemble des contextes et problématiques scientifiques, rien ne permet d'affirmer que la représentativité de ces N bases de données correspond également à notre contexte. En conséquence, il devient nécessaire de construire nous-mêmes notre base de données réelles. Cela implique d'avoir à disposition des moyens matériel et/ou financiers ainsi que du temps pour permettre l'acquisition de données d'intérêt.

Une fois cette base de données construites, elles ne sont pas encore validées par la communauté et nécessitent une expertise (annotations des signaux d'intérêts par un expert) qui exige égalment du temps et/ou de l'argent. Après l'obtention des annotations, il reste à convaincre la communauté et/ou le partenaire industriel que la base de données est suffisamment représentative du contexte considéré. Il vient alors un ensemble d'études statistiques de la base de données qui a pour objectif de détailler les propriétés internes (quantité et qualité des signaux, temps total d'enregistrement, ...) et externes(fréquence d'échantillonnage, type d'hydrophone, profondeur d'immersion, lieu d'enregistrement, ...) des données.

Après ce travail effectué, il devient possible d'envisager la construction du classifieur (phase d'apprentissage) et de prévoir un ensemble de tests (phase de validation) pour mesurer les performances de la méthode. L'objectif est toujours de garantir de bonnes capacité de généralisation en s'assurant que la méthode n'est pas « conditionnée » par les données. Dans ces conditions, la généralisation devrait, dans l'idéal, être mesurée sur une partie de la base de données dont on sait qu'elle est représentative du contexte et sur laquelle aucun test n'a été effectué. La base de données est scindée en trois ensembles, une base d'apprentissage, une base de test et un ensemble dit de *validation croisée*. Dans le cas où les données ne sont pas suffisamment nombreuses, il est possible que ce troisième ensemble ne soit pas faisable. Il convient alors de s'efforcer de travailler sur des performances moyennes en changeant par exemple la base d'apprentissage et la base de test de façon aléatoire afin d'être un peu moins dépendant de la base de données globale.

Imaginons que les performances sont obtenues sur un ensemble de validation croisée. Il est nécessaire à présent de se comparer aux méthodes pré-existantes afin de situer son travail parmi le travail de nos prédécesseurs. Cependant, la comparaison est difficile puisque les performances proposées par les N méthodes précédentes ont été mesurées dans des conditions différentes. C'est comme si nous voulions observer les variations du paramètre « méthode » alors que nous avons également fait varier les paramètres « base de données » et « performances ». Pour fixer ces « paramètres », il faut reprogrammer les méthodes présentées par nos prédécesseurs et mesurer, si possible, les mêmes performances que celles mesurées pour notre méthode. Dès lors, la comparaison prend autant de temps qu'il y a de méthodes à programmer, en sachant que, pour certaines méthodes, l'absence des détails techniques rend leur reproduction exacte impossible.

Dans la communauté bioacoustique, il n'existe pas encore de processus de validation complet des méthodes de détections et de reconnaissance qui soit référencé. Afin de palier ce manque, la communauté s'efforce de fournir de plus en plus de données réelles, notamment à chaque *workshop* de la conférence [DCLDE, 2015], ainsi que par l'intermédiaire de la base de données de mobysound présentée dans [MELLINGER et CLARK, 2006] accessible sur le site [MOBYSOUND.ORG]. Cependant, ces données ne sont malheureusement pas suffisamment représentatives de notre contexte pour être exploitables. Considérons par exemple la base de données Mobysound associée aux mysticètes. Neuf [1] des dix-sept mysticètes sont représentés dans cette base. Malheureusement, les enregistrements sont anciens (datant de 1988, 1992-1994, 1998-1999 et 2008) et nous savons que les vocalises évoluent d'année en année pour la plupart des espèces considérées. De plus, les enregistrements sont de mauvaise qualité par rapport à ce permettent les moyens actuels. Pour finir, même si la base contient de nombreuses vocalises, il arrive parfois que les annotations proposées ne soient pas parfaitement valides. C'est par exemple le cas des annotations des rorquals communs (dossier « fin-01 ») qui ont une durée de moins d'une seconde alors qu'en explorant les données nous notons que les vocalises impulsives « 20Hz-pulse » durent plus de 2 s. Cela

1. Il s'agit de la baleine bleue, la baleine du Groenland, la baleine de Brydes, la baleine grise, la baleine à bosse, la baleine de Minke, la baleine franche du Nord Pacifique, la baleine franche de l'hémisphère Sud et le rorqual commun.

implique de devoir ré-annoter les données pour les exploiter. Afin d'illustrer ces données, voici ci-dessous trois spectrogrammes générés à partir des fichiers de cette base de donnée :

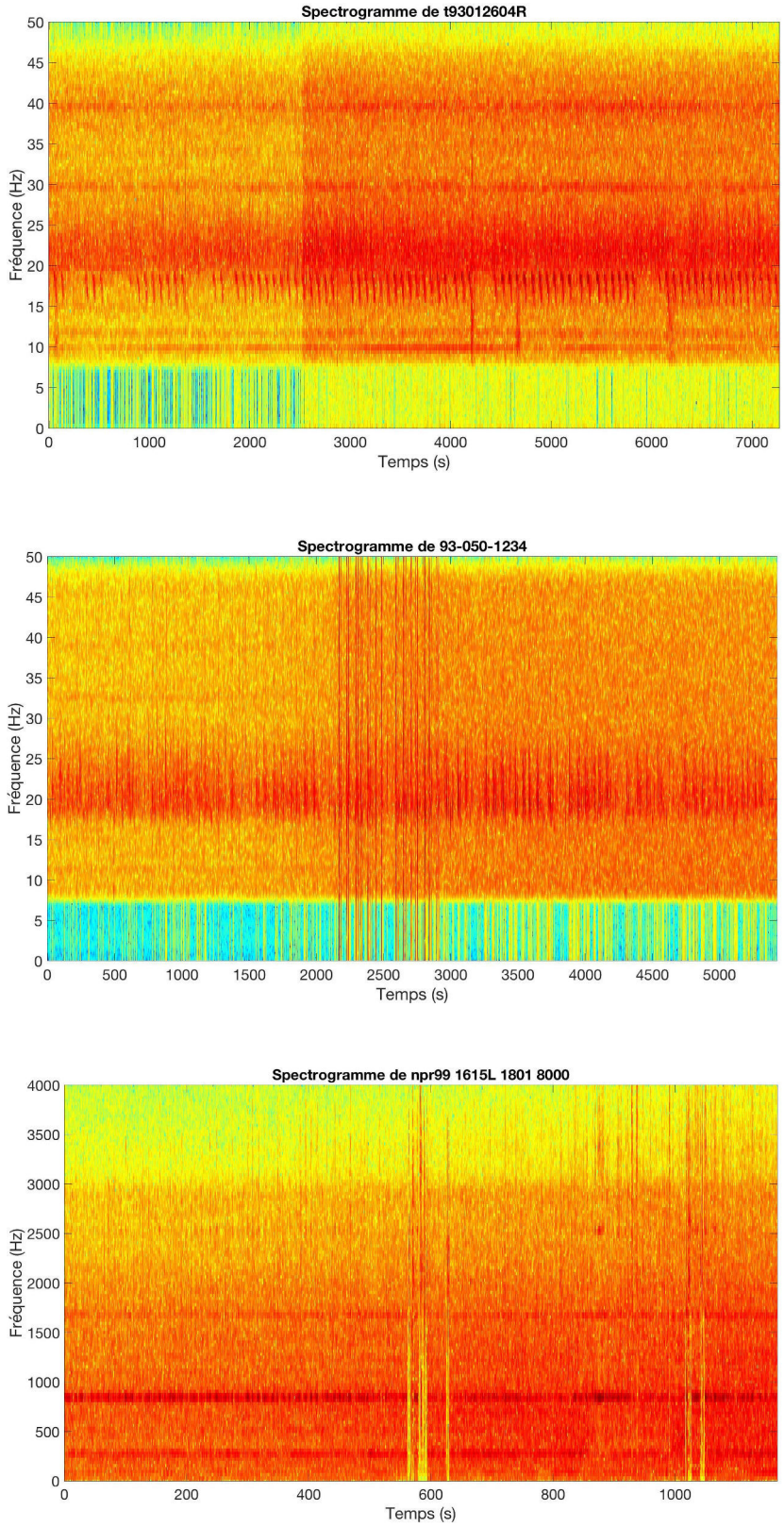


FIGURE 2.20 – Exemples de spectrogrammes représentatifs des données de baleines bleues (Premier spectrogramme), de rorqual commun (Deuxième spectrogramme) et de baleines franche du Nord de l'océan Pacifique (troisième spectrogramme) présentes dans [MOBYSOUND.ORG].

Comme nous pouvons le voir sur les spectrogrammes représentatifs des données de baleines

bleues de [MOBYSOUND.ORG] (*cf.* figure 2.20), l'ancienneté de l'enregistrement fait que la qualité des données n'est pas au niveau des moyens actuels. Sur le premier spectrogramme, il y a un brusque changement d'intensité qui ne semble pas naturel. De même, nous pouvons notamment observer sur le deuxième spectrogramme un passage entre 2 000 s à 3 000 s qui n'est, *a priori*, pas réaliste. Ainsi, même si ces données sont variées, après avoir analysé l'entièreté de la base, nous avons pris le parti de considérer qu'elles sont trop loin de notre contexte pour être exploitables.

2.8.3 Discussion sur la validation

La validation des méthodes de reconnaissance est toujours une problématique ouverte dans la littérature bioacoustique. Même si un processus de qualité peut être suffisant dans le cas de considérations industrielles, ce processus est insuffisant dans un contexte de recherche où la standardisation est urgente et indispensable. Sans référence, il devient impossible de véritablement comparer ses travaux en plus de ne pas être productif. C'est d'ailleurs pour ces raisons que nous avons été également contraint de gérer nous-mêmes notre base de données et nos critères de performances que nous présentons au chapitre 3. Une idée, peut-être un peu idéaliste, serait que la communauté bioacoustique et les entreprises définissent l'ensemble des besoins et/ou problématiques actuels, afin d'identifier et de mettre à jour par la suite des critères d'évaluation (performances fixées à mesurer) et des données correspondant à ces différents contextes. Il serait peut-être envisageable de créer un « comité de validation des données » qui aurait pour objectif de maintenir les données à jour en plus de valider et surtout de rendre facilement accessibles les nouvelles bases de données. C'est-à-dire qu'il faudrait poursuivre le travail proposé par le projet [MOBYSOUND.ORG] en allant jusqu'au processus complet de validation. De cette façon, tout nouveau membre de la communauté bioacoustique pourrait s'intégrer rapidement et pourrait situer ces résultats par rapport aux travaux proposés dans la littérature, ce qui n'est pas évident actuellement.

2.9 Conclusion

A travers ce chapitre, nous avons mis en avant les problématiques de la reconnaissance à travers la littérature bioacoustique. Après avoir discuté de la difficulté à définir la notion de classes, nous avons vu les avantages et les inconvénients des deux approches principales pour faire de la reconnaissance, à savoir une approche basée sur de la réduction de dimension et une autre basée sur une mesure de similarité. Nous retenons que la démarche employée dépend de la problématique inhérente au contexte et des moyens disponibles pour répondre à cette problématique. Ainsi, même si l'objectif de reconnaître des classes est généralement flou, car défini empiriquement par le contexte d'application, il est possible de fixer des spécifications basées sur des paramètres mesurables, comme par exemple le respect d'un taux de fausses alarmes pour un détecteur. Dans notre contexte, nous rappelons que nous choisissons une architecture d'apprentissage basée sur une mesure de similarité, afin de garantir les résultats obtenus à chaque étape de la mise en œuvre de la méthode considérée. Afin, d'être le plus représentatif possible des données, nous allons voir au chapitre suivant que l'utilisation de dictionnaires semble être une bonne alternative à la limitation des représentations des signaux proposées plus haut.

Chapitre 3

SINR-SRC

« On fait la science avec des faits, comme on fait une maison avec des pierres : mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison. »

— Henri Poincaré

Sommaire

Ce chapitre développe notre cheminement méthodologique qui a donné lieu à l'utilisation des dictionnaires pour représenter les données et réaliser la reconnaissance. Notre méthode repose sur des représentations parcimonieuses qui supposent que les cris ou vocalises de mysticètes se trouvent dans un sous-espace linéaire décrit par une représentation basée sur un dictionnaire. Le classifieur tient compte du bruit (et plus généralement des classes inconnues) en refusant d'assigner le signal observé à une classe donnée s'il n'est pas inclus dans le sous-espace linéaire couvert par les dictionnaires de vocalises de mysticètes. Le rejet du bruit (ou des classes inconnues) est réalisé sans apprentissage de descripteurs. De plus, la méthode proposée est modulaire au sens où les classes de vocalises peuvent être ajoutées ou retirées du système de reconnaissance sans nécessiter de « re-apprentissage ». Le classifieur est facile à concevoir puisqu'il repose sur quelques paramètres. Des résultats expérimentaux sur cinq types de vocalises de mysticètes sont présentés. Ils comprennent des cris en Z ou Z-call de baleine bleue Antarctique, deux types de cris de baleine bleue pygmée de Madagascar, des 20Hz-pulse de rorqual commun et des D-call de baleine bleue du Pacifique Nord. Sur cet ensemble de données, contenant 2 385 cris et 15 000 échantillons de bruit, un rappel moyen de 92,1 % est obtenu et 97,3 % des données de bruit (persistantes et transitoires) sont correctement rejetées par le classifieur. En perspectives, nous développons ensuite la possibilité (preuve de concept) de faire de l'apprentissage incrémentale semi-supervisée et mettons en place un niveau de confiance associée à chaque vocalise. Pour finir nous illustrons également la possibilité de prendre en compte les effets de la dispersion modale dans l'apprentissage de dictionnaire.

3.1 Méthodologie

Comme nous l'avons vu au chapitre 2, nous sommes dans un contexte de classification supervisée. Rappelons que l'apprentissage supervisé donne les moyens à un système de reconnaissance d'identifier automatiquement une nouvelle observation, après avoir appris des exemples labellisées (ou étiquetés) par des experts. A partir d'un ensemble labellisé ou *ensemble d'apprentissage* constitué de N paires $\{(\boldsymbol{s}_i, \ell_i)\}_{1\leq i\leq \mathrm{N}}$ représentatives de C *classes, i.e.*, C types de vocalises dans notre cas, où $\boldsymbol{s}_i$ est le i^{e} vecteur de caractéristiques de l'ensemble d'apprentissage et ℓ_i est la classe correspondante ou label de $\boldsymbol{s}_i$, e.g., $\ell_5 = 3$ signifie que le 5^{e} élément de l'ensemble d'apprentissage appartient à la 3^{e} classe. Cet ensemble d'apprentissage est utilisé pour déterminer une application $f(\cdot|\{(\boldsymbol{s}_i, \ell_i)\}_{1\leq i\leq \mathrm{N}})$ qui infère un label à partir d'un vecteur de caractéristiques donné.

L'application f est liée aux deux approches vues au chapitre précédent. Elle est apprise sur un ensemble d'apprentissage, soit en minimisant une *fonction de coût* représentant les erreurs de prédictions (c'est-à-dire les différences entre le label de la classe prédit par le système de reconnaissance et le label réel), soit elle découle d'une *mesure de similarité* qui compare les nouvelles données (données de test) aux données apprises par le système de reconnaissance.

Comme nous le verrons plus loin, notre méthode repose sur la seconde approche. Ce choix est principalement motivé par la volonté de construire une méthode robuste et modulaire où la mesure de similarité ne dépend pas de l'ensemble d'apprentissage ni du nombre de classes. Il est également souhaitable d'éviter d'utiliser trop d'hyper-paramètres ("pas si faciles à régler") pour faciliter le déploiement de la méthode.

Dans la suite, $\{\boldsymbol{s}_k : k > \mathrm{N}\}$ représente les vecteurs de caractéristiques (ou descripteurs) en entrée du système de reconnaissance. Étant donné un vecteur de caractéristiques $\boldsymbol{s}_k$ avec $k > \mathrm{N}$, $\widehat{\ell}_k = f(\boldsymbol{s}_k|\{(\boldsymbol{s}_i, \ell_i)\}_{1\leq i\leq \mathrm{N}})$ est le label de sortie dans $\{1, 2, \ldots, \mathrm{C}\}$ assigné à $\boldsymbol{s}_k$.

Dans la méthode proposée ci-dessous, les vecteurs de caractéristiques sont des séries temporelles digitales de cris. Nous supposons que la détection des régions d'intérêt dans les séries temporelles a déjà été réalisée automatiquement ou manuellement.

3.1.1 D'une mesure de similarité aux représentations parcimonieuses

Il existe une grande variété de mesures de similarité, par exemple la distance euclidienne, la valeur absolue, la vraisemblance, la corrélation, etc. Soit $|\langle \boldsymbol{s}_k, \boldsymbol{s}_i \rangle|$ le produit scalaire normalisé positif ou *corrélation* entre un signal $\boldsymbol{s}_k$ et un signal $\boldsymbol{s}_i$. Pour les approches telles que les banques de filtres adaptés ou les corrélations de spectrogrammes, l'application f choisit la classe qui maximise la corrélation entre un signal de test $\boldsymbol{s}_k$, $k > \mathrm{N}$, et tous les signaux de l'ensemble d'apprentissage, c'est-à-dire

$$\hat{\ell}_k = \ell_{i^*}, \text{ où } i^* = \underset{i \in \{0,1\ldots,\mathrm{N}-1\}}{\operatorname{argmax}} |\langle \boldsymbol{s}_k, \boldsymbol{s}_i \rangle|. \tag{3.1}$$

Une extension bien connue d'une telle approche est l'algorithme des K plus proches voisins (K *Nearest Neighbors [KNN]* en anglais) [FRIEDMAN et collab., 2001], où $\boldsymbol{s}_k$ est assigné à la classe la plus présente parmi K plus proches voisins (par exemple les K signaux dans l'ensemble de données d'apprentissage ayant la plus forte corrélation avec $\boldsymbol{s}_k$). En général, choisir K supérieur à 1 est bénéfique car il réduit le bruit global [STEVENS et collab., 1967].

Au-delà des KNN, la reconnaissance peut être basée sur une mesure de similarité entre le signal de test $\boldsymbol{s}_k$ à reconnaître et une *combinaison linéaire* des K signaux les plus proches de $\boldsymbol{s}_k$. Tous les signaux d'apprentissage deviennent alors des *atomes* élémentaires qui peuvent être combinés pour créer de nouveaux signaux. De cette façon, le nouvel espace de représentation permet de couvrir un espace plus grand que l'ensemble de données d'apprentissage original et, à ce titre, doit mieux capturer la structure « intrinsèque » des signaux d'intérêt. D'une part, K doit être suffisamment petit pour empêcher le sur-apprentissage (*overfitting* en anglais), surtout en présence de bruit. D'autre part, à partir d'un signal de test, la mesure de similarité doit aider à sélectionner une combinaison linéaire d'atomes de la même classe que le signal pour garantir une comparaison significative entre celui-ci et chaque modèle moyen de chaque classe. De là vient un compromis nécessaire dans le choix de K entre le risque de sur-apprentissage et la nécessité de bien « approcher » le signal de test.

Formellement, on suppose que tout signal de test $\boldsymbol{s}_k$ de dimension n de la classe c se situe approximativement dans l'ensemble des combinaisons linéaires des signaux d'apprentissage associé à cette classe, c'est-à-dire

$$\boldsymbol{s}_k \approx \mathbf{A}_c \boldsymbol{w}_c, \text{ avec } \|\boldsymbol{w}_c\|_0 \leq \mathrm{K} \ll \mathrm{N}_c, \tag{3.2}$$

où $\mathbf{A}_c \in \mathbb{R}^{n \times \mathrm{N}_c}$ est une matrice contenant tous les N_c signaux d'apprentissage de taille n appartenant à la classe c, $\boldsymbol{w}_c \in \mathbb{R}^{\mathrm{N}_c}$ est un vecteur de poids utilisé dans la combinaison linéaire et $\|\boldsymbol{w}_c\|_0$ dénote la pseudo-norme ℓ_0 qui retourne le nombre de coefficient non nuls dans $\boldsymbol{w}_c$. Lorsque $\boldsymbol{s}_k$ peut être représenté par un petit nombre de coefficients non nuls dans la base $\mathbf{A}_c$, le modèle (3.2) est appelé **« représentation parcimonieuse »** dans la littérature du traitement du signal [ELAD, 2010]. L'inégalité $\|\boldsymbol{w}_c\|_0 \leq \mathrm{K}$ est appelée la **contrainte de parcimonie**. Cette contrainte K est directement liée à la complexité de chaque vocalise à classifier. Les signaux combinant variabilité et grande complexité (tels que les signaux erratiques) doivent être construits à partir d'un grand nombre d'atomes tandis que les signaux de faible complexité doivent être composés de quelques atomes.

Par exemple les D-call des baleines bleues [THOMPSON et collab., 1996] sont des chirps (signaux modulés en fréquence) qui peuvent être approximés par une combinaison linéaire de quelques atomes. Néanmoins, ces cris présentent un grande variabilité (fréquence initiale, durée, largeur de la bande de fréquences, vitesse de modulation fréquentielle). Par conséquent, la norme ℓ_0 du symbole $\boldsymbol{w}_c$ est petite pour chaque cri, mais les atomes actifs, correspondant à des entrées non nulles de $\boldsymbol{w}_c$, peuvent être différents d'un cri à l'autre de sorte que N_c doit être grand. Notons que le modèle (3.2) est une approximation, car les cris peuvent être affectés par les conditions locales de propagation et par le bruit. Cependant, les très bons résultats obtenus en section 3.2 indique que ce modèle est suffisamment précis pour faire de la reconnaissance.

3.1.2 La reconnaissance basée sur les représentations parcimonieuses

Basé sur un modèle linéaire similaire à (3.2), Wright et collab. ont proposé dans [WRIGHT et collab., 2009] l'algorithme SRC (pour *Sparse Representation-based Classification* en anglais). Cet algorithme a obtenu des résultats impressionnants dans un large éventail d'applications (reconnaissance de chants d'oiseaux [TAN et collab., 2013], reconnaissance de signaux EEG [SHIN et collab., 2015] et reconnaissance de visages [ORTIZ et collab., 2013; WRIGHT et collab., 2009]). Initialement appliquée à la reconnaissance faciale, nous proposons d'adapter cette approche à notre contexte. À cette fin, cette sous-section rappelle la procédure de l'algorithme SRC, tandis que les deux suivantes proposent des fonctionnalités supplémentaires pour améliorer les performances de SRC dans notre application.

SRC suppose que les signaux de test peuvent être représentés par une combinaison linéaire de signaux d'apprentissage. Dans notre contexte, ces signaux sont des séries temporelles numérisées et représentent les vecteurs de caractéristiques d'entrée du classifieur. L'algorithme SRC est une procédure en deux étapes : (i) il recherche la combinaison linéaire de signaux d'entraînement qui se rapproche le mieux, au sens de la parcimonie, du signal de test et (ii) choisit la classe qui contribue le plus à cette approximation. Plus précisément, le véritable label du signal de test $\boldsymbol{s}_k$ étant inconnu, $\boldsymbol{s}_k$ est d'abord représenté comme une combinaison linéaire de tous les signaux d'entraînement stockés dans une matrice $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_\mathrm{C}] \in \mathbb{R}^{n \times \sum_{c=1}^{\mathrm{C}} \mathrm{N}_c}$, où C est le nombre de classes de vocalises, *i.e.*,

$$\boldsymbol{s}_k \approx \mathbf{A}\boldsymbol{w}, \text{ avec } \|\boldsymbol{w}\|_0 \leq \mathrm{K}. \tag{3.3}$$

Idéalement, les entrées de $\boldsymbol{w} \in \mathbb{R}^{\sum_{c=1}^{\mathrm{C}} \mathrm{N}_c}$ sont toutes des zéros exceptées, au plus, les K entrées liées aux signaux d'apprentissage de la même classe que le signal de test. Par exemple, si $\boldsymbol{s}_k$ appartient à la classe c, c'est-à-dire, $\ell_k = c$, alors $\boldsymbol{w}$ devrait idéalement satisfaire $\boldsymbol{w} = [0, \cdots, 0, \boldsymbol{w}_c^{\mathrm{T}}, 0, \cdots, 0]^{\mathrm{T}}$ où $\boldsymbol{w}_c \in \mathbb{R}^{\mathrm{N}_c}$ et $\|\boldsymbol{w}_c\|_0 \leq \mathrm{K}$. Par conséquent, la classe réelle du signal de test peut être obtenue en estimant le vecteur $\boldsymbol{w}$ et en trouvant les indices des entrées non nulles de $\boldsymbol{w}$. Cependant, dans la pratique, en raison du bruit et de la non-orthogonalité des signaux d'apprentissage des différentes classes entre eux, des entrées non nulles de $\boldsymbol{w}$ peuvent apparaître à des indices qui ne sont pas associés à la vraie classe du signal de test. Par conséquent, le label de la classe du signal de test n'est pas déterminé en trouvant les indices des entrées non nulles de $\boldsymbol{w}$ mais en trouvant les entrées de $\boldsymbol{w}$ qui donnent la meilleure approximation de $\boldsymbol{s}_k$ dans (3.3).

Plus précisément, la procédure en deux étapes de l'algorithme SRC est la suivante :

1. Estimer $\boldsymbol{w}$ par encodage parcimonieux de $\boldsymbol{s}_k$ sur la base $\mathbf{A}$. C'est-à-dire, en résolvant les problèmes suivants.

$$\boldsymbol{w}^* = \underset{w}{\operatorname{argmin}} \|\boldsymbol{s}_k - \mathbf{A}\boldsymbol{w}\|_2^2, \text{ avec } \|\boldsymbol{w}\|_0 \leq \mathrm{K}. \tag{3.4}$$

   L'encodage parcimonieux peut être effectué avec des algorithmes de poursuite [ELAD, 2010] ou de minimisation de la norme ℓ_1 [MAIRAL et collab., 2010]. Dans la section 3.2, cette étape est mise en œuvre avec l'algorithme de « poursuite de correspondance orthogonale » ou *orthogonal matching pursuit* (OMP) [PATI et collab., 1993].

2. Associer $\boldsymbol{s}_k$ à la classe $\hat{\ell}_k$ qui satisfait

$$\hat{\ell}_k = \underset{1 \leq c \leq \mathrm{C}}{\operatorname{argmin}} \|\boldsymbol{s}_k - \mathbf{A}\delta_c(\boldsymbol{w}^*)\|_2^2, \tag{3.5}$$

   où $\delta_c(\boldsymbol{w}^*)$ est une fonction caractéristique qui sélectionne les coefficients de $\boldsymbol{w}^*$ associés à la c^{e} classe. Pour tout $\boldsymbol{w} \in \mathbb{R}^{\sum_{c=1}^{\mathrm{C}} \mathrm{N}_c}$, $\delta_c(\boldsymbol{w}) \in \mathbb{R}^{\sum_{c=1}^{\mathrm{C}} \mathrm{N}_c}$ est un vecteur dont les entrées non nulles sont les entrées dans $\boldsymbol{w}$ qui sont liées à la c^{e} classe. Par exemple, si $\boldsymbol{w} = [\boldsymbol{w}_1^{\mathrm{T}}, \boldsymbol{w}_2^{\mathrm{T}}, \cdots, \boldsymbol{w}_\mathrm{C}^{\mathrm{T}}]^{\mathrm{T}}$ où chaque $\boldsymbol{w}_i$ appartient à la classe i, alors $\delta_c(\boldsymbol{w}) = [0, \cdots, 0, \boldsymbol{w}_c^{\mathrm{T}}, 0, \cdots, 0]^{\mathrm{T}}$. La solution à (3.5) est trouvée par une recherche exhaustive à travers toutes les classes.

3.1.3 Option de compression

Idéalement, l'ensemble de données d'apprentissage $\mathbf{A}$ doit couvrir l'espace qui inclut tous les cris de mysticètes que nous souhaitons reconnaître. En particulier, pour chaque classe, le symbole $\mathbf{A}_c$ doit incorporer suffisamment de variabilité pour modéliser tous les cris possibles de la même classe. Il est donc souhaitable d'injecter dans $\mathbf{A}$ la quantité maximale d'informations dont nous disposons sur ces vocalises. Cependant, la complexité liée au calcul de (3.4) augmente avec la taille de $\mathbf{A}$ sans nécessairement améliorer les performances si $\mathbf{A}$ contient des signaux redondants. Pour limiter la redondance de $\mathbf{A}$ et ainsi obtenir un compromis entre la variabilité et la charge de calcul, nous suggérons de construire un dictionnaire de dimensions inférieures $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_\mathrm{C}]$ par rapport à l'ensemble de données d'apprentissage, où chaque sous-matrice $\mathbf{D}_c$ a $\mathrm{N}'_c \leq \mathrm{N}_c$ colonnes, c'est-à-dire, $\mathbf{D}_c \in \mathbb{R}^{n \times \mathrm{N}'_c}$. Chaque $\mathbf{D}_c$ est défini comme le sous-dictionnaire qui conduit à la meilleure représentation possible pour chaque signal d'apprentissage de la classe c avec la contrainte de parcimonie (3.4). Plus précisément, le nouveau sous-dictionnaire $\mathbf{D}_c$ pour la classe c est obtenu en résolvant le problème de minimisation :

$$\begin{aligned} &\min_{\mathbf{D}_c, \mathbf{W}} \|\mathbf{A}_c - \mathbf{D}_c \mathbf{W}\|_\mathrm{F}^2 \\ &\text{sous la contrainte } \|\boldsymbol{w}_i\|_0 \leq \mathrm{K}, \forall\, 1 \leq i \leq \mathrm{N}_c, \end{aligned} \tag{3.6}$$

où $\mathbf{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{\mathrm{N}_c}]$ et $\boldsymbol{w}_i \in \mathbb{R}^{\mathrm{N}'_c}$. Le problème de minimisation (3.6) est communément appelé « apprentissage de dictionnaire » et n'est effectué qu'une seule fois hors ligne. Les solutions numériques de (3.6) peuvent être obtenues avec la méthode de direction optimisée (*method of optimized direction* [MOD]) [Engan et collab., 1999], K-SVD [Aharon et collab., 2006] ou en apprentissage en ligne [Mairal et collab., 2010]. Une fois le dictionnaire de dimensions inférieures appris, $\mathbf{A}$ et $\mathbf{A}_c$ sont remplacés par $\mathbf{D}$ et $\mathbf{D}_c$ dans (3.4) et (3.5), respectivement et $\delta_c(\cdot)$ est adapté à la taille de la matrice $\mathbf{D}$. En plus de supprimer l'information redondante dans le processus d'apprentissage, l'apprentissage du dictionnaire extrait les caractéristiques structurelles pertinentes de $\mathbf{A}$ et doit ainsi limiter la dépendance aux signaux d'apprentissage bruités et/ou les problèmes de surapprentissage.

3.1.4 Option de rejet

Un défi majeur dans la classification automatique des sons sous-marins est la gestion du « bruit » (ou plus généralement des classes inconnues). Dans notre contexte, le bruit est défini comme n'importe quel signal de test (entrée du système) qui n'appartient pas à l'une des C classes de vocalises de mysticètes (sortie du système). Ce bruit peut être :

— Un bruit transitoire ou une interférence qui désigne tout signal transitoire sans intérêt pour le classifieur, par exemple, les cris d'autres baleines, le bruit des navires, les canons à air, les tremblements de terre, les craquements de glace, etc.

— Un bruit de fond qui est un mélange de nombreuses sources sonores ambiantes non identifiables qui n'incluent aucun signal transitoire.

L'option de rejet offre la possibilité de refuser d'assigner le signal examiné à une classe. Dans [Wright et collab., 2009, Sec. 2.4], une option de rejet est proposée pour l'algorithme SRC. Cette option repose sur l'hypothèse qu'un signal de test valide a une représentation parcimonieuse dont les entrées non nulles se concentrent principalement sur une classe, alors qu'un signal à rejeter a généralement des coefficients largement répartis entre plusieurs classes. Ainsi, bien qu'une telle hypothèse puisse être valide dans des applications telles que la reconnaissance faciale [Wright et collab., 2009], elle n'est pas applicable dans notre contexte. La raison principale est que les bruits acoustiques sous-marins transitoires peuvent avoir une quantité non négligeable de leur énergie dans un sous-espace dans lequel réside une classe spécifique de vocalises. Par exemple, les coefficients parcimonieux de bruits impulsifs sont susceptibles de se concentrer sur les classes liées aux cris impulsifs (comme les 20Hz-pulse de rorqual commun présentés dans la section 3.2.2),

alors que les coefficients de bruits tonaux seront liés aux vocalises tonales ayant des fréquences similaires. Pour traiter le bruit, nous proposons d'appliquer une procédure de post-traitement qui décide si le signal de test se trouve réellement dans le sous-espace « couvert » par les colonnes du sous-dictionnaire correspondant à la classe choisie par SRC. Plus précisément, le résultat de SRC est validé si le rapport Signal sur Interférence plus bruit (*Signal to Interference plus Noise Ratio* [SINR]) estimé

$$\mathrm{SINR}(\boldsymbol{s}_k, \hat{\ell}_k) = \frac{||\mathbf{D}\delta_{\hat{\ell}_k}(\boldsymbol{w}^*)||_2^2}{||\boldsymbol{s}_k - \mathbf{D}\delta_{\hat{\ell}_k}(\boldsymbol{w}^*)||_2^2} \tag{3.7}$$

est supérieur à un certain seuil. Basé sur le modèle (3.2), $\mathbf{D}\delta_{\hat{\ell}_k}(\boldsymbol{w}^*)$ est une estimation du signal d'intérêt et $\boldsymbol{s}_k - \mathbf{D}\delta_{\hat{\ell}_k}(\boldsymbol{w}^*)$ est une estimation de l'interférence plus le bruit de fond. Ce critère mesure la qualité de reconstruction du signal de test $\boldsymbol{s}_k$ lorsqu'il est approché par une combinaison linéaire des éléments de $\mathbf{D}_{\hat{\ell}_k}$. Il est inspiré des détecteurs à taux de fausses alarmes constant (CFAR pour Constant False Alarm Rate) de signal connu dans le bruit avec une puissance inconnue, qui montrent des propriétés optimales en ce qui concerne la performance de détection [Scharf, 1991; Socheleau et collab., 2015; Socheleau et Samaran, 2017]. La méthodologie utilisée pour définir le seuil SINR est présentée à la section 3.2.4. Un aspect clé de notre approche est que le classifieur n'a pas besoin d'apprendre les caractéristiques des bruits transitoires pour les rejeter. Cela diffère des méthodes telles que [Halkias et collab., 2013] où les caractéristiques du bruit sont apprises par les réseaux de neurones ou de [Urazghildiiev et collab., 2009] où, pour chaque classe de bruit, « *un modèle paramétrique du bruit est introduit. Les modèles sont basés sur les propriétés spectrales de bruits impulsifs typiques observés dans les données* » [Urazghildiiev et collab., 2009, pp. 360]. Cela implique de trouver des exemples exhaustifs de bruits sous-marins, ce qui semble difficile étant donné la complexité de l'environnement sous-marin. Les caractéristiques des sons sous-marins détectés dépendent fortement de l'environnement anthropique, biologique, géologique ou océanographique ainsi que de la façon dont les capteurs sont placés dans la colonne d'eau. Ainsi, le bruit appris ou modélisé dans un contexte peut difficilement être transposé dans un autre contexte.

3.1.5 Procédure global

Le processus de classification résultant des considérations précédentes est dénommé SINR-SRC. Il est résumé comme suit et illustré dans le cas de deux classes dans la figure 3.1.

1. Sélection hors ligne des signaux d'apprentissage représentatifs de leur classe de vocalise.
2. Application de l'option de compression (3.6) si la complexité de calcul doit être réduite.
3. Avec un signal de test donné $\boldsymbol{s}_k$, effectuez un encodage parcimonieux de $\boldsymbol{s}_k$ sur le dictionnaire $\mathbf{D}$ en calculant :

$$\boldsymbol{w}^* = \underset{w}{\operatorname{argmin}} \|\boldsymbol{s}_k - \mathbf{D}\boldsymbol{w}\|_2^2, \text{ avec } \|\boldsymbol{w}\|_0 \leq \mathrm{K}.$$

4. Application de SRC en calculant la classe qui contribue le plus au signal de test $\boldsymbol{s}_k$:

$$\hat{\ell}_k = \underset{1\leq c\leq \mathrm{C}}{\operatorname{argmin}} \|\boldsymbol{s}_k - \mathbf{D}_c\delta_c(\boldsymbol{w}^*)\|_2^2.$$

5. Gestion des classes de bruit : si $\mathrm{SINR}(\boldsymbol{s}_k, \hat{\ell}_k)$est supérieur à un certain seuil, le résultat fourni par SRC est validé, sinon $\boldsymbol{s}_k$ est considéré comme du bruit.

Cette procédure de SINR-SRC peut être illustrée par le schéma de la figure 3.1. En plus des bonnes performances de classification obtenues par SINR-SRC (voir section 3.2), notez aussi que la méthode est modulaire, ce qui peut être très utile dans un contexte opérationnel. Par exemple, si une nouvelle classe de vocalises de mysticètes doit être ajoutée à un classifieur SINR-SRC existant, il n'est pas nécessaire de « ré-entraîner » tout le classifieur comme dans les approches basées sur

de la réduction de dimension, rappelons notamment le cas des réseaux neuronaux (*cf.* Chapitre 2). Seul le nouveau sous-dictionnaire associé à la nouvelle classe doit être appris. De plus, pour réduire les classifications erronées de la surveillance acoustique passive en ligne, des informations préalables telles que la position géographique du capteur pourraient être prises en compte en supprimant les sous-dictionnaires dans **D** correspondant aux espèces dont les habitats sont connus comme étant loin du capteur.

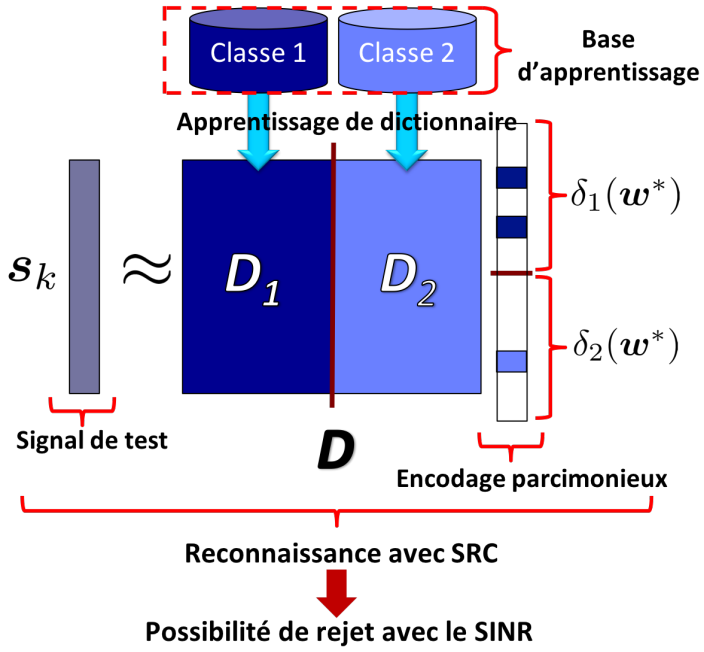
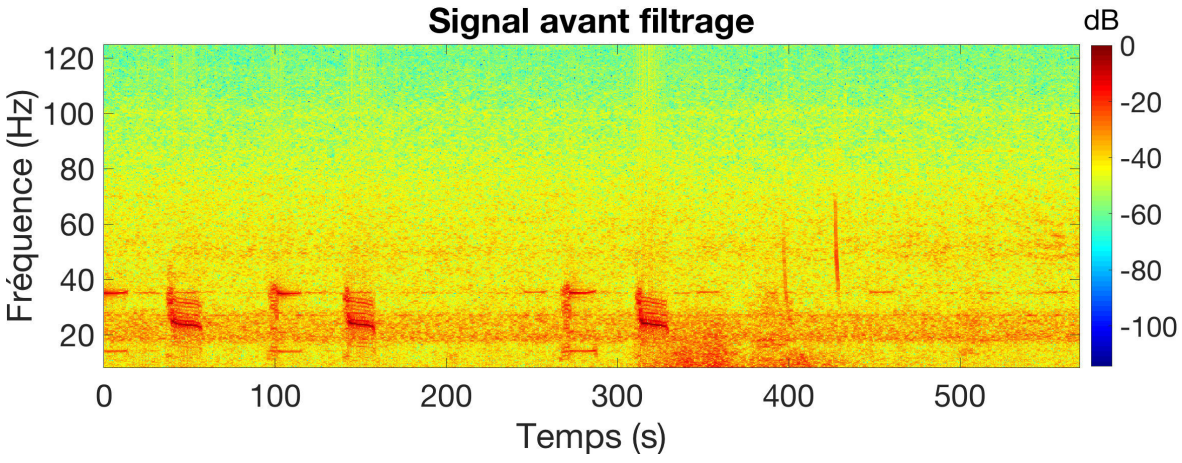


FIGURE 3.1 – Vue d'ensemble de la méthode de reconnaissance pour 2 classes

3.2 Résultats expérimentaux

3.2.1 Pré-traitement des données d'entrées

En amont de la reconnaissance, le bruit est « stationnarisé » en temps et en fréquence en appliquant un filtre de blanchiment du bruit sur les données d'entrées. L'intérêt est de « normaliser » les données d'entrées afin d'être le plus invariant possible aux différents types de bruit de fond. Nous illustrons ci-dessous l'effet du blanchiment du bruit (*cf.* avant le filtrage et après le filtrage (*cf.* figure 3.2) :



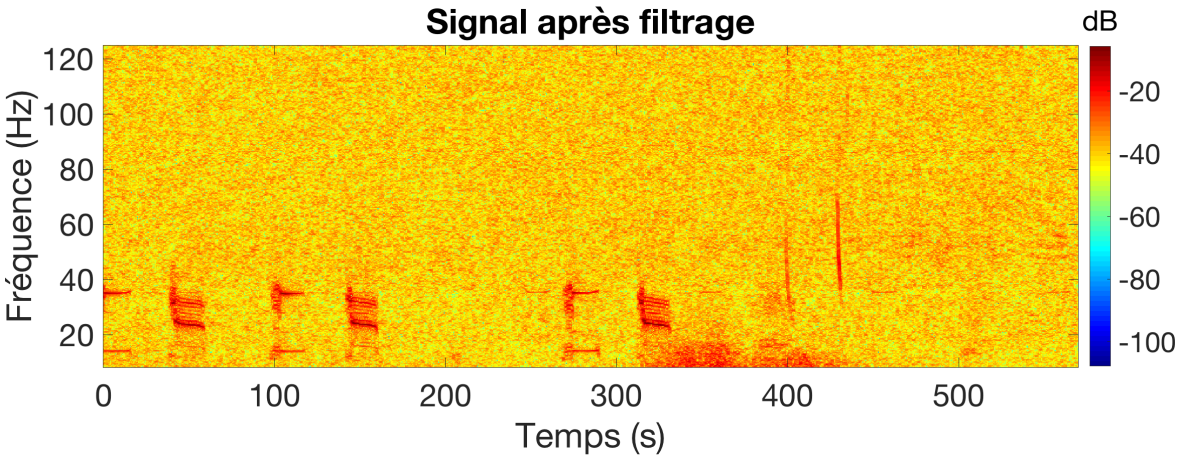


FIGURE 3.2 – Illustrations des effets avant (figure du haut) et après le blanchiment du bruit (figure du bas)

En se focalisant sur le bruit de fond, nous pouvons noter qu'il est devenu stationnaire en temps et en fréquence.

3.2.2 Jeu de données

Nous évaluons SINR-SRC sur les 5 types de vocalises suivants : les Z-call (ou cri en Z) de la baleine bleue Antarctique [SAMARAN et collab., 2013; STAFFORD et collab., 2004], les 2 classes dénommées Mad1 et Mad2 de baleine bleue pygmée de Madagascar [SAMARAN et collab., 2013], les 20Hz-pulse (ou train pulsé de 20Hz) de rorqual commun [SIROVIC et collab., 2009] et les D-call (ou cri en D) de la baleine bleue de l'Atlantique Nord [DCLDE, 2015; THOMPSON et collab., 1996]. Nous avons choisi ces vocalises pour les raisons suivantes :

— Elles se recoupent en fréquence et certaines ont des durées proches ce qui les rend difficiles à discriminer avec ces deux descripteurs temps-fréquence élémentaires (*cf.* figures 3.3 et 3.4).

— Elles offrent de la variété en terme de type de signaux : sons pulsés, sons tonaux ou modulation en fréquence (*cf.* figure 3.5).

— Elles mettent en valeur différents niveaux de variabilité : des vocalises stéréotypées (par exemple les Z-call) à des vocalises qui varient en durée, en bande et en modulation de fréquence (par exemple les D-call).

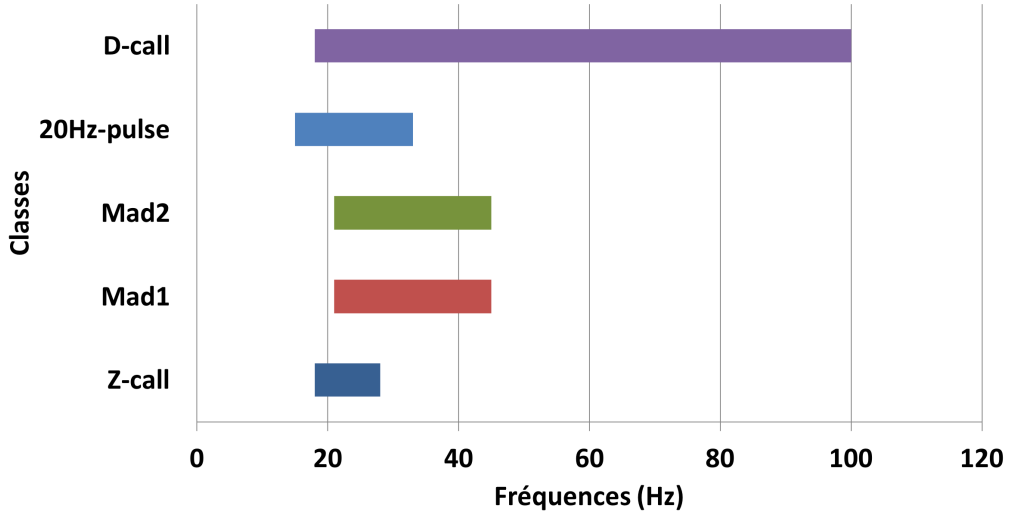


FIGURE 3.3 – Plage de fréquences de chaque type de vocalise

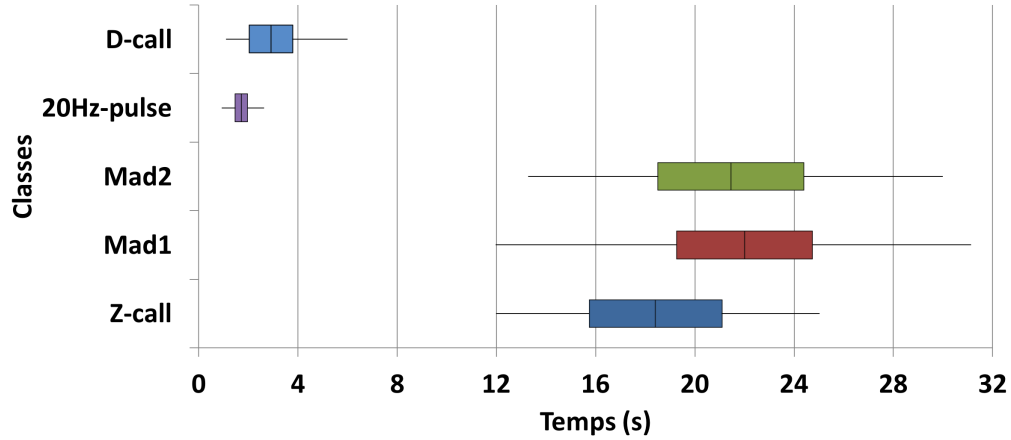


FIGURE 3.4 – Boxplot des durées de chaque type de vocalise

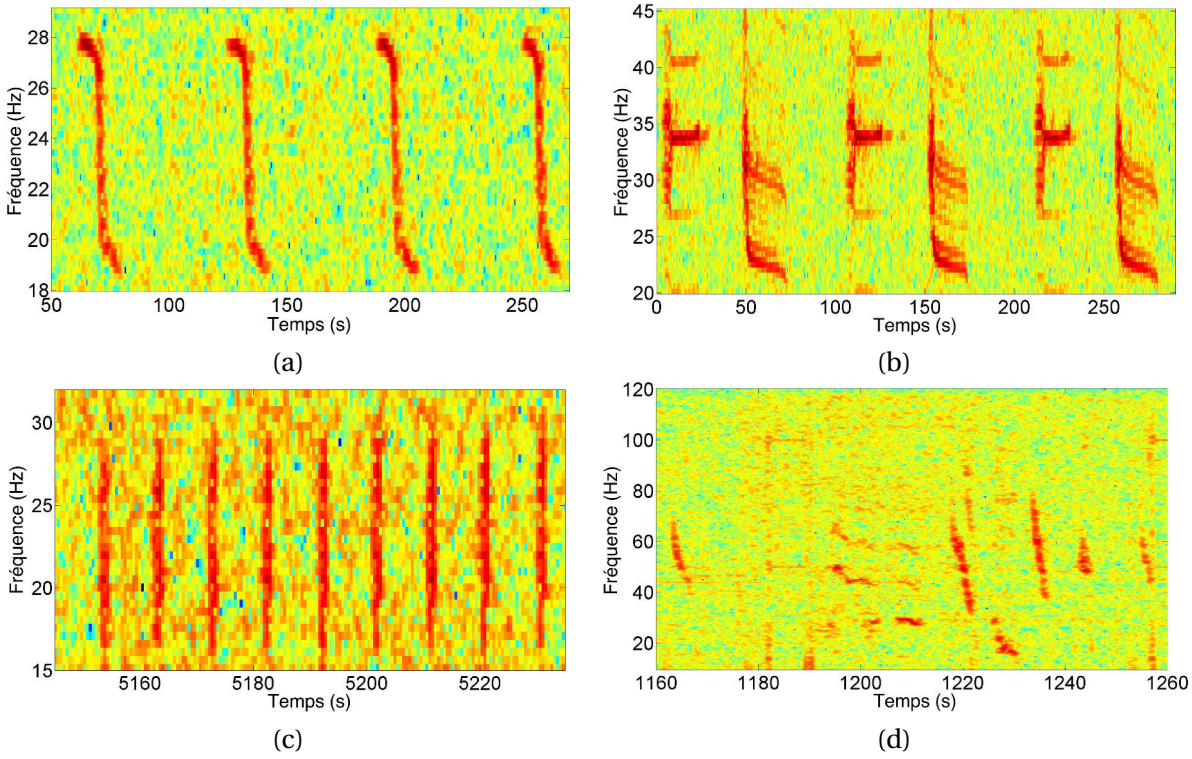


FIGURE 3.5 – Exemples de spectrogrammes de notre jeu de données. (a) 4 Z-call produits par la baleine bleue de l'Antarctique, (b) une alternance des 2 types de vocalises produites par la baleine bleue pygmée de Madagascar, (c) Un train de 20Hz-pulse produit par le rorqual commun, (d) 5 D-call produits par la baleine bleue du Pacifique Nord.

Ces cinq types de vocalises ont été extraites manuellement de trois bases de données :

La base de donnée DEFLOHYDRO : Trois hydrophones autonomes ont été déployés près des territoires français dans le Sud de l'Océan Indien d'octobre 2006 à janvier et avril 2008. L'objectif de ce projet était de surveiller les signaux acoustiques basses fréquences et notamment ceux produits par les grandes baleines [DZIAK et collab., 2008]. Ces trois instruments ont été espacés et positionnés dans le bassin de Madagascar, à 320 milles nautiques (NM) au Sud de l'Île de la Réunion, 470 NM du Nord-Est (NEAMS) et 350 NM du Sud-Ouest (SWAMS) de l'île d'Amsterdam. Les lignes d'amarrage ont été ancrées sur le fond marin entre 3 410 et 5 220 m de profondeur et les hydrophones ont été déployés près de l'axe du canal sonore (SOFAR) entre 1 000 m et 1 300 m. Les instruments enregistrent des sons en continu à une fréquence d'échantillonnage de 250 Hz (gamme de fréquences 0,1-110 Hz) [SAMARAN et collab., 2013]. 254 Z-call et 1 000 20Hz-pulse ont

été extraits manuellement de cet ensemble de données.

L'ensemble de données OHASISBIO : Dans le prolongement de l'expérimentation des données de DEFLOHYDRO, un réseau d'hydrophones a été initialement déployé en décembre 2009 sur cinq sites dans le sud de l'Océan Indien. Cette expérience a été conçue pour surveiller les sons basse fréquence, produits par des événements sismiques et volcaniques, et par les grands mysticètes [LEROY et collab., 2016; TSANG-HIN-SUN et collab., 2016]. 551 vocalises de baleines bleues pygmées de Madagascar ont été extraites des données enregistrées par l'hydrophone de la Réunion dans le bassin de Madagascar (coordonnées géographiques : +26° 05' S, +058 °08' E) en mai 2015. 264 sont des cris de type 1 et 287 sont de type 2 (*cf.* figure 3.5).

La base de données de DCLDE 2015 : Ces données ont été obtenues avec du matériel d'enregistrement acoustique à haute fréquence déployé dans la baie de Californie du Sud. 380 D-call ont été extraits des données enregistrées sur le site CINMS B (latitude : 34 - 17,0 N, longitude : 120 - 01,7 W) à l'été 2012 [DCLDE, 2015].

La base de données totale représente 2 385 vocalises de mysticètes. Chaque vocalise a été annotée manuellement en temps et en fréquence : les heures de début et de fin sont identifiées ainsi que la fréquence la plus basse et la plus haute de chaque vocalise. Toutes les vocalises sont filtrées par bande passante en fonction de leur annotation et rééchantillonnées à 250 Hz. Pour appliquer SRC, toutes les vocalises doivent avoir le même nombre d'échantillons, ce qui est facilement réalisé par du remplissage de zéro (*zéro-padding* en anglais). Comme montré à la figure 3.6, la base de donnée contient des signaux avec une grande variété de rapports signal à bruit (*Signal to noise ratio*, SNR en anglais). Le SNR est ici défini comme le rapport de la puissance du signal sur la puissance du bruit, *mesurée dans la bande de fréquence de chaque vocalise individuelle.*

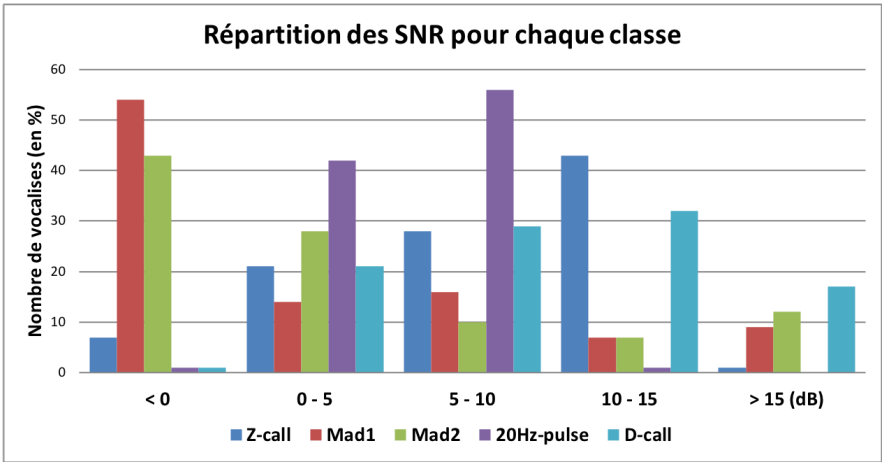


FIGURE 3.6 – Répartition des SNR (en décibels [dB]) de toutes les vocalises de la base de données

Notons que quatre types de vocalises (Z-call, 20Hz-Pulse, Mad1, Mad2) ont été enregistrés dans l'océan Indien et un type (D-call) de vocalise dans la baie de Californie du Sud. Les capteurs des réseaux OHASISBIO ou DEFLOHYDRO peuvent détecter les quatre premiers types de cris dans les mêmes enregistrements [LEROY et collab., 2017] mais les D-call des baleines bleues du Pacifique Nord sont observés séparément. Dans la pratique, ce type de vocalises peut donc être différencié des autres vocalises en fonction des habitats supposés. Afin de mettre en difficulté notre méthode, l'information sur l'emplacement n'a pas été prise en compte. Une approche similaire a été envisagée dans [HALKIAS et collab., 2013]. De plus, les baleines bleues de l'Océan Indien produisent également des D-call [SHANNON et collab., 2005]. Bien que légèrement différents des D-call des baleines bleues du Pacifique Nord, ces D-call sont aussi des chirps, c'est-à-dire des signaux modulés en fréquence avec une fréquence initiale, une vitesse de modulation, une durée et une bande passante en fréquence variables. Cela suggère que notre méthode pourrait également s'appliquer à ce type de vocalises.

3.2.3 Base de données de bruits

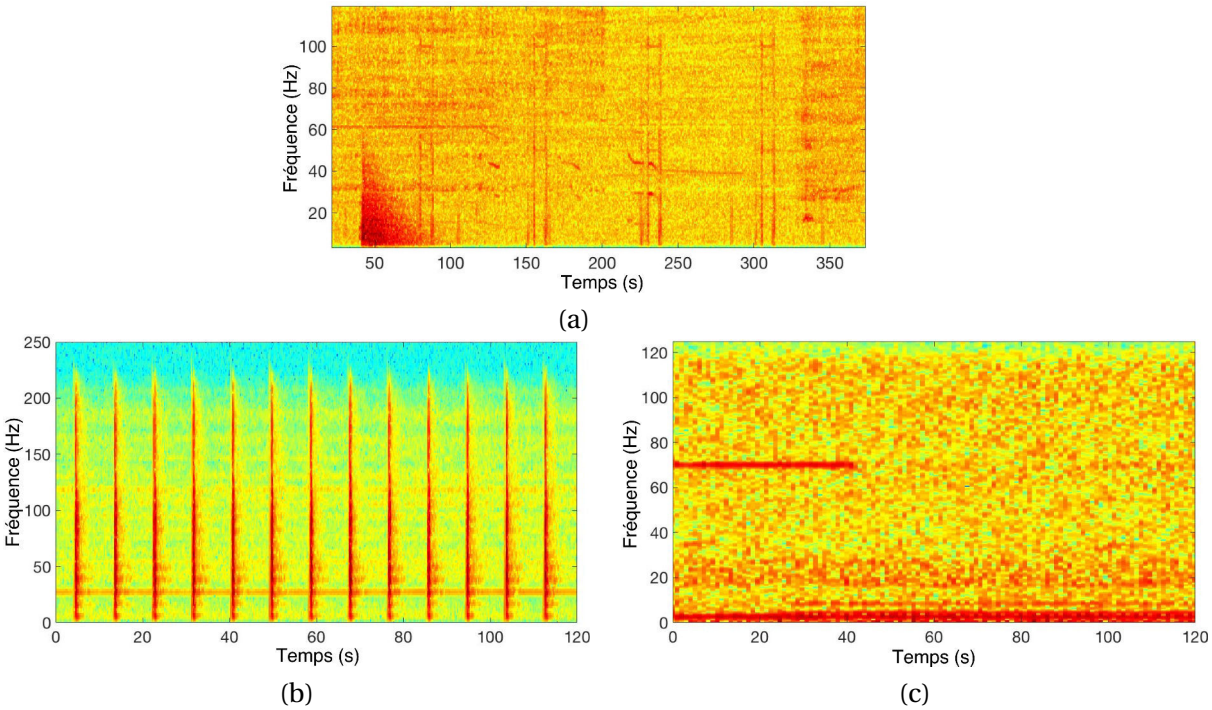


FIGURE 3.7 – Exemples de spectrogrammes de la base de données de bruits. (a) Extrait de DCLDE 2015, (b) bruit de campagne sismique provenant de SERCEL et (c) bruit océanique extrait de DEFLOHYDRO.

Pour tester la robustesse de SINR-SRC par rapport au bruit, une base de données de bruit a également été créée. 5 000 échantillons de bruits ont été extraits aléatoirement de l'ensemble de données DEFLOHYDRO, plus 5 000 de l'ensemble de données DCLDE 2015 et enfin 5 000 autres d'un ensemble de données fourni par SERCEL et enregistré lors d'études sismiques. Les 5 000 premiers échantillons de bruit correspondent principalement à ce que l'on appelle le « bruit de fond » dans la section 3.1.4 et les autres sont pour la plupart des signaux transitoires sans intérêt pour le classifieur, c'est-à-dire, des « interférences » (*cf.* figure 3.7). En pratique, les caractéristiques (durée, largeur de bande fréquentielle, puissance, etc.) des exemples de bruit injectés dans le classifieur dépendent du comportement réel du détecteur utilisé pour identifier la région d'intérêt avant la classification. Comme nous souhaitons tester les performances de notre méthode indépendamment du détecteur, les exemples de bruit ont été extraits au hasard des ensembles de données. De plus, pour mettre en difficulté notre méthode, les exemples de bruit ont été filtrés de manière à ce que leurs largeurs de bande en fréquence et durées soient identiques aux largeurs de bande en fréquence et durées des vocalises de mysticètes à identifier. Cela correspond au pire scénario pour le système de reconnaissance, car les exemples de bruits filtrés auront une plus grande quantité d'énergie dans les sous-espaces dans lesquels résident les cris, ce qui entraînera nécessairement une augmentation du SINR (*cf.* équation (3.7)).

3.2.4 Performances

Nous commençons par analyser et comparer les performances de SINR-SRC avec une méthode de référence de l'état de l'art sans utiliser l'option de rejet. Puis nous présentons les résultats lorsque l'option de rejet est activée. Pour finir nous discutons des effets, sur les performances, de la taille du dictionnaire et de la contrainte de parcimonie en fin de cette section. Les performances du classifieur sont mesurées à l'aide d'une validation croisée. Comme le montre le tableau 3.1, pour chaque classe (à l'exception du bruit), 100 cris sont sélectionnés au hasard pour l'apprentissage et les autres cris de cette classe sont utilisés pour les tests. La moyenne de tous les tests

présentés est calculée sur 100 sélections aléatoires de l'ensemble d'apprentissage afin de s'assurer que les résultats et les conclusions dépendent le moins possible d'un choix spécifique des données d'apprentissage. Pour chaque classe, la mesure de rappel, utilisée ci-dessous, est définie comme le rapport entre les cris correctement classés et le nombre total de cris dans cette classe. Cette mesure est parfois appelée sensibilité ou taux positif réel. Un rappel de 100 % pour la classe des Z-call signifie que tous les cris en Z ont été correctement classés.

Classes	Signaux d'apprentissage	Signaux de test	Total
Z-call	100	154	254
Mad1	100	164	264
Mad2	100	187	287
20Hz-pulse	100	900	1000
D-call	100	280	380
Bruits	-	15000	15000

TABLEAU 3.1 – Nombre de signaux d'apprentissage et de test pour chaque classe et pour chaque itération de la validation croisée.

Résultats sans l'option de rejet

	Z-call	Mad1	Mad2	20Hz-pulse	D-call
Z-call	**100**	0.00	0.00	0.00	0.00
	0.10	0.00	0.10	0.00	0.10
Mad1	0.00	**97.7**	1.90	0.00	0.40
	0.00	1.10	1.10	0.00	0.50
Mad2	0.00	0.30	**99.60**	0.00	0.10
	0.00	0.30	0.30	0.10	0.20
20Hz-pulse	0.00	0.00	0.00	**100**	0.00
	0.00	0.00	0.00	0.00	0.00
D-call	0.00	0.00	0.00	0.00	**100**
	0.00	0.10	0.10	0.00	0.10

TABLEAU 3.2 – Matrice de confusion moyenne de l'algorithme SRC (en %) sans l'option de rejet. Pour chaque classe, la ligne supérieure contient la moyenne et la ligne inférieure l'écart-type obtenu pour 100 itérations de validation croisée.

Le tableau 3.2 montre la matrice de confusion moyenne de l'algorithme SRC sans rejet et sans bruit injecté dans le classifieur. Chaque colonne de la matrice correspond à la prédiction de la méthode tandis que les lignes correspondent à la vérité terrain de l'expert. Le résultat dans chaque case est exprimé comme le ratio du nombre de vocalises prédites sur le nombre total de vocalises considérées. Par exemple, le résultat correspondant à l'intersection de la 3^{e} ligne et de la 5^{e} colonne signifie que 4 ‰ des vocalises de types Mad1 ont été prédite comme des D-call. L'écart-type des résultats de classification est également affiché dans le tableau 3.2. Pour ce test, aucune réduction de la dimension du dictionnaire n'est appliquée, c'est-à-dire que $\mathbf{D} = \mathbf{A}$ et la contrainte de parcimonie K est fixée à 3 (l'impact de ces paramètres sur la performance de classification est discuté dans la section 3.2.4). Un rappel moyen global de 99 % est obtenu. Le classifieur SRC ne fait pas seulement très peu d'erreurs, mais il est également robuste aux changements d'ensembles de données d'apprentissage.

A titre de comparaison, les résultats de classification obtenus avec une méthode générale inspirée de [BAUMGARTNER et MUSSOLINE, 2011] sont présentés dans le tableau 3.3. Cette méthode

est basée sur le calcul du spectrogramme et l'extraction de quatre attributs temps-fréquence pondérés en amplitude pour chaque vocalise : la fréquence moyenne, la variation de fréquence, la variation temporelle et la pente de l'énergie de la vocalises dans l'espace temps-fréquence. Ces attributs sont ensuite utilisés comme les entrées d'un classifieur basé sur une fonction discriminante quadratique (analogue au maximum de vraisemblance *a posteriori* ou méthode MAP). La méthode inspirée de [BAUMGARTNER et MUSSOLINE, 2011] fournit des performances acceptables sur l'ensemble de données de test. Néanmoins, il persiste plus de confusion entre les classes et la variance des résultats est plus élevée notamment pour la reconnaissance des Z-call. Ceci est dû à l'ensemble des caractéristiques choisi qui n'est pas suffisamment discriminant. Les attributs temps-fréquence sont trop grossiers et provoquent plus de confusion que pour SINR-SRC.

	Z-call	**Mad1**	**Mad2**	**20Hz-Pulse**	**D-call**
Z-call	**79.89**	0.00	19.66	0.45	0.00
	15.96	0.00	16.06	0.44	0.00
Mad1	0.25	**96.77**	2.70	0.00	0.29
	0.66	1.44	1.22	0.00	0.33
Mad2	3.42	0.69	**95.89**	0.00	0.00
	3.09	0.37	3.14	0.00	0.00
20Hz-Pulse	0.01	0.00	0.00	**93.00**	6.99
	0.02	0.00	0.02	5.13	5.14
D-call	3.73	0.00	0.00	0.00	**96.27**
	1.17	0.00	0.00	0.00	1.17

TABLEAU 3.3 – Matrice de confusion (en %) de la méthode présentée dans [BAUMGARTNER et MUSSOLINE, 2011]. Pour chaque classe, la ligne supérieure contient la moyenne et la ligne inférieure l'écart-type obtenu pour 100 itérations de validation croisée.

Résultat avec option de rejet

Nous illustrons maintenant les performances de SINR-SRC lorsque l'option de rejet est activée. Nous rappelons que, contrairement aux méthodes alternatives telles que [HALKIAS et collab., 2013; URAZGHILDIIEV et collab., 2009], le rejet du bruit est obtenu sans apprentissage ni modélisation d'un modèle de descripteurs du bruit, c'est-à-dire qu'aucun dictionnaire n'est construit à partir des données de bruit ni dédié à la gestion du bruit. Une entrée est rejetée par le classifieur si le SINR estimé, obtenu en calculant (3.7), est inférieur à un certain seuil. Cette approche est très efficace pour distinguer les données de bruit des vocalises d'intérêt [SOCHELEAU et SAMARAN, 2017].

Par exemple, nous présentons ci-après une méthode fondée sur l'estimation d'une probabilité de fausses alarmes, comme cela se fait couramment dans le cadre de Neyman-Pearson pour la vérification d'hypothèses binaires. En supposant que la fonction de densité de probabilité (pdf) de la métrique SINR est connue lorsque des exemples de bruit sont injectés dans le classifieur, un seuil de rejet garantissant une probabilité de fausses alarmes spécifiée par l'utilisateur peut alors être trouvé. Cependant, comme l'espace de tous les bruits transitoires sous-marins possibles est très grand, il n'est guère possible de connaître précisément cette pdf dans la pratique. Par conséquent, nous recourons à une approche empirique et injectons dans le classifieur des exemples de bruit aléatoire synthétique pour obtenir une pdf à partir de laquelle nous pouvons fixer un seuil. Ce bruit est synthétique afin d'être aussi indépendant que possible d'un ensemble de données spécifiques. Dans notre expérimentation, le bruit est généré à partir d'échantillons gaussiens filtrés en temps et en fréquence, avec des largeurs de bande fréquentielle et durées identiques aux largeurs de bande fréquentielle et durées des cris de mysticètes à identifier.

Comme expliqué à la section 3.2.3, cela correspond au pire scénario pour notre méthode car un tel bruit produira un SINR plus grand qu'un bruit avec n'importe quelle autre largeur de bande

fréquentielle et durée. Dans la pratique, il est peu probable que les détecteurs utilisés déclenchent le système de reconnaissance avec un faux signal d'alarme dont la bande passante fréquentielle et la durée correspondent exactement à celles d'un cris de mysticète réel. L'examen des scénarios les plus pessimistes est justifié par la volonté de mesurer des performances de reconnaissances réalistes, quel que soit le détecteur.

Les seuils de rejet sont estimés sur chaque distribution de SINR obtenue après injection d'échantillons gaussiens *dans chaque dictionnaire*. La figure 3.8 montre un exemple de seuil de rejet choisi en fixant une probabilité de fausse alarme à 1‰ sur la distribution SINR obtenue avec des échantillons gaussiens filtrés injectés dans le dictionnaire Z-call. Notez que des distributions autres que gaussiennes pourraient être pertinentes pour modéliser des bruits. Cependant, la figure 3.8 indique que la distribution SINR (en rouge) des bruits réels (pas nécessairement gaussiens) obtenus après SRC est proche de la distribution obtenue avec des échantillons d'entrée gaussiens.

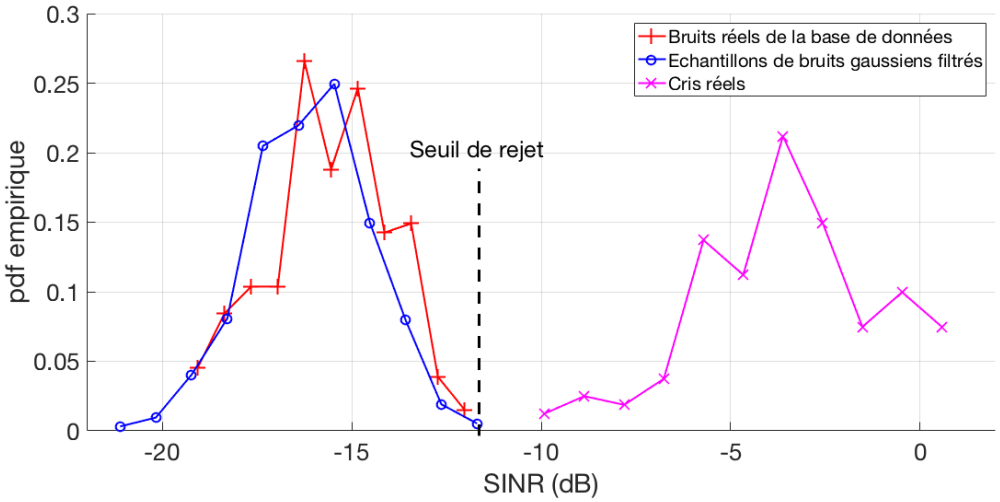


FIGURE 3.8 – Distribution du SINR, tel que calculé dans (3.7), pour les échantillons gaussiens (en bleu), le bruit réel (en rouge) et les cris réels de l'ensemble de données de test (en magenta), tous identifiés comme Z-call selon l'algorithme SRC sans l'option de rejet. Dans cette situation, pour une probabilité de fausse alarme de 1 ‰, le seuil de rejet est fixé à -12 dB.

Le tableau 3.4 montre la matrice de confusion moyenne de l'algorithme SINR-SRC avec rejet. Comme prévu, l'activation de l'option de rejet produit une légère baisse du rappel moyen. Cette baisse est principalement significative pour les D-call en raison de leur forte variabilité de durée, de plage de fréquence et de répartition d'énergie, ce qui fait que certaines vocalises dans l'ensemble de données de test sont considérées comme des interférences et sont donc rejetées. Néanmoins, 97,1% des entrées de bruit sont correctement rejetées. Cela montre clairement que SINR-SRC est capable de gérer efficacement les données d'entrée qui sont inconnues du classifieur. Cette propriété est hautement souhaitable dans l'environnement sous-marin à basse fréquence où les sources sonores, telles que les bruits et les interférences, peuvent être très actives. Pour une analyse plus approfondie de la performance de rejet de la métrique SINR (3.7), le lecteur est invité à se référer à [SOCHELEAU et SAMARAN, 2017, Sec. 4], où les courbes des caractéristiques de fonctionnement des récepteurs sont affichées pour les D-call et les cris Mad2 du jeu de données DCLDE 2015 et OHASISBIO, respectivement. A des fins de comparaison, les résultats de reconnaissance de SRC avec l'option de rejet *désactivée* sont affichés dans le tableau 3.5 lorsque des entrées de bruit sont injectées dans le classifieur. On peut voir que les entrées de bruit sont réparties entre les 5 classes avec une probabilité légèrement plus élevée pour les classes cris incorporant des structures impulsives avec une forte pente en fréquence. Cela s'explique par le grand nombre de signaux transitoires dans la bibliothèque de bruit.

	Z-call	Mad1	Mad2	20Hz-pulse	D-call	Rejected
Z-call	**99.0**	0	0	0	0	1.0
	0.7	0	0	0	0	0.7
Mad1	0	**93.4**	0.3	0	0	6.3
	0	1.6	0.3	0	0	1.6
Mad2	0	0.4	**96.5**	0	0	3.1
	0	0.3	1.0	0	0	1.0
20Hz-Pulse	0	0	0	**92.6**	0	7.4
	0	0	0	2.0	0	2.0
D-call	0	0	0	0	**79.0**	21.0
	0	0.1	0	0	3.8	3.8
Noise	0.3	0.2	1.2	0	1.0	**97.3**
	0.3	0.2	1.1	0	1.4	2.8

TABLEAU 3.4 – Matrice de confusion de l'algorithme SRC avec l'option de rejet activée. Pour chaque classe, la ligne supérieure contient la moyenne et la ligne inférieure l'écart-type obtenu pour 100 itérations de validation croisée.

	Z-call	Mad1	Mad2	20Hz-pulse	D-call
Bruit	11.8	5.0	35.1	21.7	26.5
	19.6	7.3	27.9	29.5	30.9

TABLEAU 3.5 – Résultats de la classification du SRC avec des entrées de bruit seulement. L'option de rejet est désactivée. La ligne supérieure contient la moyenne et la ligne inférieure l'écart-type obtenu pour 100 itérations de validation croisée.

Jusqu'à présent, aucune réduction de la dimension du dictionnaire n'a été prise en compte, c'est-à-dire que $\mathbf{D} = \mathbf{A}$. Comme mentionné dans la section 3.1.2, limiter la redondance en résolvant (3.6) pendant la phase d'apprentissage peut être utile pour réduire la complexité calculatoire.

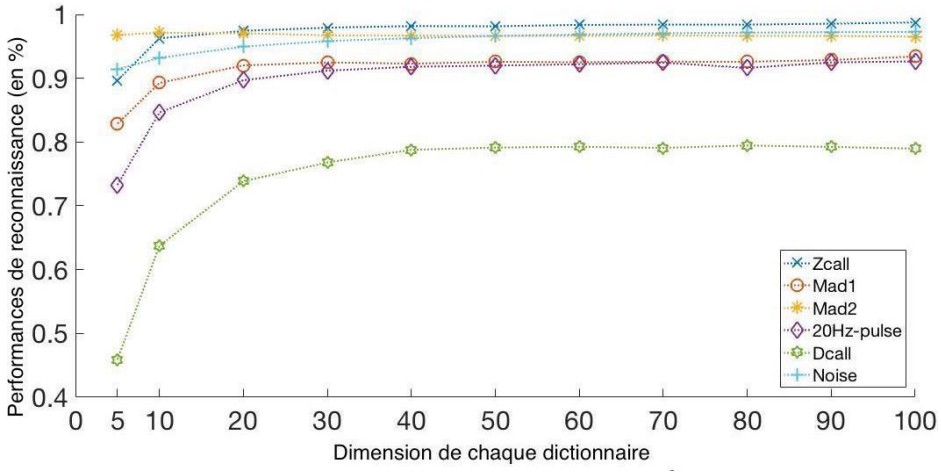


FIGURE 3.9 – Rappel moyen en fonction de la taille du dictionnaire N'_c.

La figure 3.9 montre l'impact de la taille du dictionnaire N'_c sur la performance de reconnaissance pour chaque classe de cris. Pour ce test, l'équation (3.6) a été résolue en utilisant le dictionnaire en ligne MAIRAL et collab. [2010] (le code MATLAB™ est disponible à `http://spams-devel.gforge.inria.fr/`). La taille du dictionnaire influe sur le rappel et il est intéressant de noter que son impact dépend de la classe. Pour les cris stéréotypés tels que les Z-call, la taille du dictionnaire peut être petite puisque la dimension de l'espace du signal est liée à la variabilité du cri considéré, qui est faible dans ce cas. Cependant, pour des signaux variables tels que les D-call, qui

ont aussi des caractéristiques de chevauchement avec les 20Hz-pulse, le rappel de reconnaissance augmente (en moyenne) avec la taille du dictionnaire. Dans cette expérience, choisir $N'_c = 40$ pour chaque classe est suffisant pour atteindre la performance optimale.

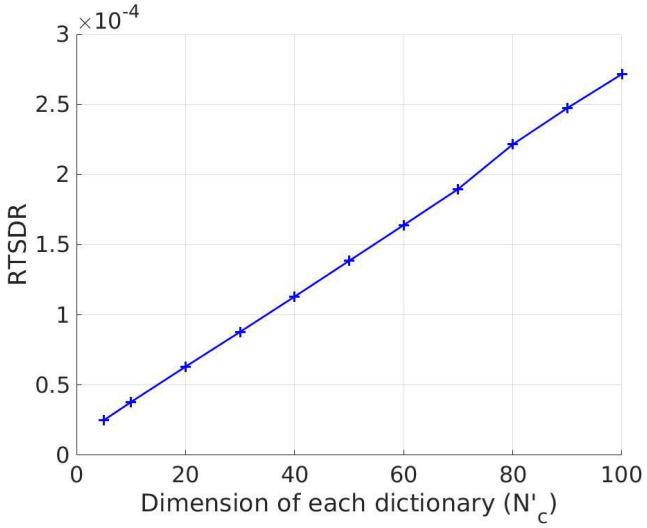


FIGURE 3.10 – RTSDR en fonction de la taille du dictionnaire N'_c.

L'impact de la taille du dictionnaire sur la complexité du calcul est visible dans la figure 3.10 où le rapport du temps d'exécution sur la durée du signal (RTSDR pour *run-time-to-signal-duration ratio* en anglais) de SINR-SRC est montré en fonction de la taille du dictionnaire N'_c. Ce rapport est calculé comme étant la durée du temps de traitement divisée par la durée totale de l'ensemble des données de tests (58 h). SINR-SRC est implémenté en MATLAB™ (sans calcul parallèle) et fonctionne sur une station de travail avec le processeur Intel Core i7 à 2,9 GHz, 8 Go de mémoire RAM et un disque dur interne DDR3. La plupart du temps de calcul est passé à résoudre (3.4) en utilisant OMP, ce qui fait que le RTSDR augmente avec N'_c. Dans cette expérience, le temps de traitement augmente linéairement avec N'_c. Ainsi, selon la figure 3.9, le temps de traitement peut être divisé par 2,5 en choisissant $N'_c = 40$ au lieu de $N'_c = 100$ sans perte de performance. Pour $N'_c = 40$, SINR-SRC a pris moins de 24 secondes pour traiter les 58 heures de signaux de tests, ce qui répond aux exigences de la plupart des applications PAM. Notez que ce temps devrait augmenter avec le nombre de classes considérées par le système de reconnaissance.

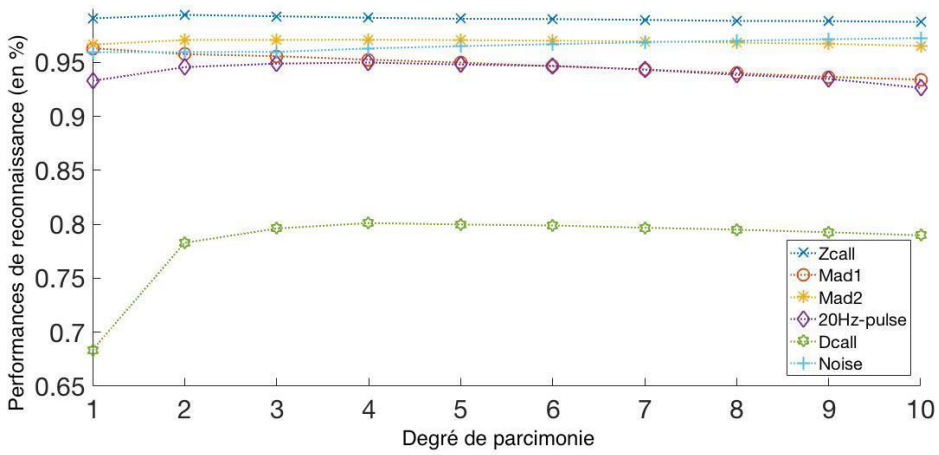


FIGURE 3.11 – Rappel moyen en fonction de la contrainte de parcimonie K. $N'_c = 100$ et l'option de rejet est activée.

Comme le montre la figure 3.11, la contrainte de parcimonie K peut également affecter le rappel de classification. De même que pour la taille du dictionnaire, la valeur optimale pour K dépend

de la variabilité et de la complexité des signaux de test et dépend donc de la classe. Cependant, aucun réglage précis n'est requis. SINR-SRC fonctionne mieux pour toutes les classes lorsque K est supérieur à 1, K = 1 correspondant à un banc de filtres adaptés (ou banc de corrélateur). Pour une contrainte de parcimonie supérieure à 3 et inférieur à 10, ce test montre que SINR-SRC est robuste au choix de K. Puisque K contribue à la complexité de notre algorithme, il peut être pertinent de le limiter à 3 ou 4 pour les classes de vocalises testées dans cette expérience. De plus, le choix d'une valeur élevée pour K (supérieure à 10 par exemple) peut être préjudiciable pour les performances de reconnaissance car la métrique SINR aura tendance à rejeter moins d'échantillons de bruit [SOCHELEAU et SAMARAN, 2017, Sec. 4.1.2].

Performances en présence de propagation modale

Nous allons à présent observer l'évolution des performances en présence de propagation modale (preuve de concept) et mettre ainsi en évidence la robustesse de la méthode SINR-SRC face aux conditions environnementales. Pour ce faire, nous allons simuler la réponse impulsionnelle du milieu désiré et nous allons appliquer ce filtre uniquement à la base de test (pas de prise en compte par le dictionnaire) pour évaluer la dégradation des performances par rapport à l'augmentation de la distance à la source sonore. Pour notre test, nous avons choisi d'utiliser la modélisation d'un guide de Pekeris avec le simulateur ORCA [WESTWOOD et collab., 1996] en utilisant les paramètres suivants :

— célérité des ondes dans la colonne d'eau : 1500 m.s^{-1}
— hauteur de la colonne d'eau : 50 m
— profondeur du récepteur : 15 m
— profondeur de la source : 20 m
— distance de la source : 1, 2, 3, 4, 5 puis 10 km
— 1 socle de taille semi-infini avec une célérité de propagation des ondes de 1800 m.s^{-1} et une densité de 1.8 tonnes.m^{-3}

Voici les spectrogrammes des réponses impulsionnelles échantillonnées à 250 Hz avec une distance de la source à 1, 5 puis 10 km (observation de la dispersion modale) appliquée aux vocalises (*cf. figures 3.12 à 3.14*) :

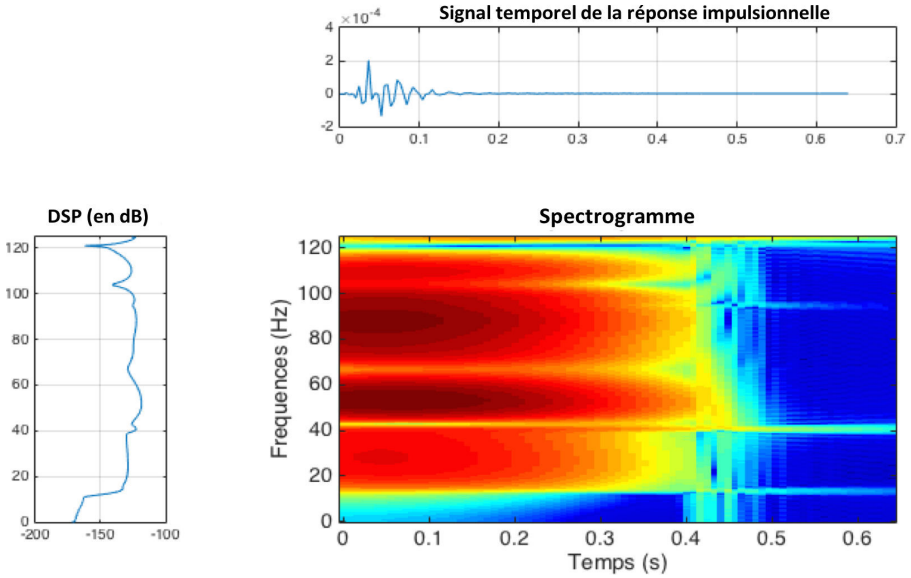


FIGURE 3.12 – Spectrogramme de la réponse impulsionnelle du milieu pour une distance à la source de 1 km. (DSP : Densité Spectrale de Puissance)

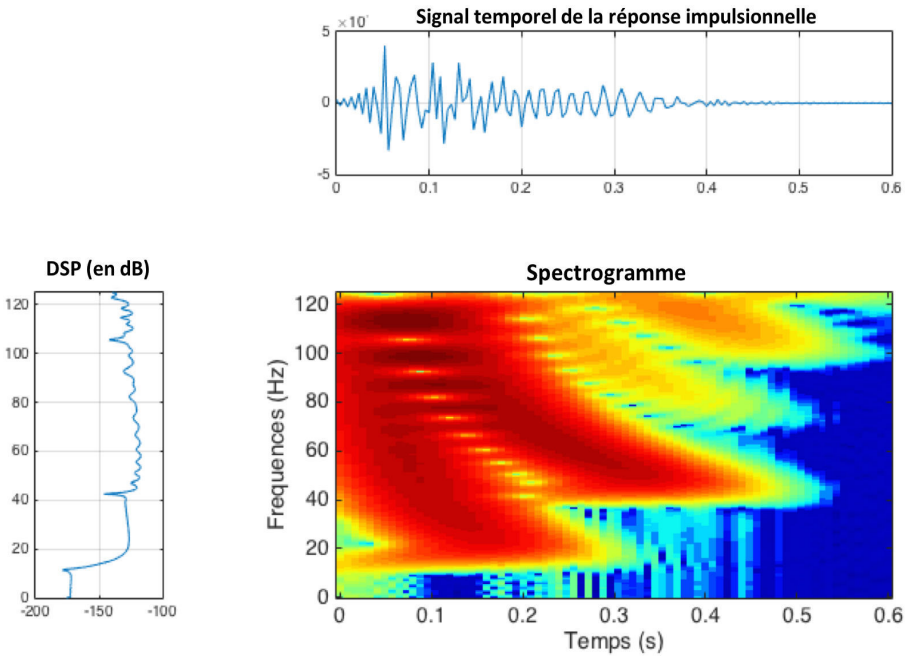


FIGURE 3.13 – Spectrogramme de la réponse impulsionnelle du milieu pour une distance à la source de 5 km. (DSP : Densité Spectrale de Puissance)

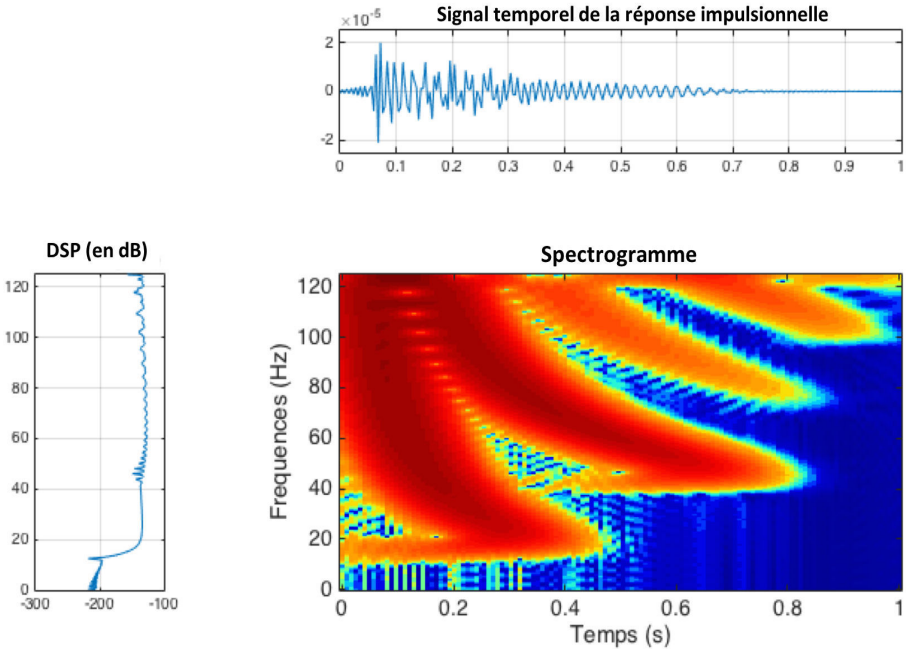


FIGURE 3.14 – Spectrogramme de la réponse impulsionnelle du milieu pour une distance à la source de 10 km. (DSP : Densité Spectrale de Puissance)

Comme nous pouvons l'observer, plus la distance à la source est grande et plus l'effet de la propagation modale est important. Plus précisément nous sommes en présence de dispersion intermodale (la séparation des modes en petit fond augmente avec la distance ainsi que lorsque la fréquence diminue) et intramodale (deux fréquences ont des vitesses de groupe différentes et cet effet tend à étaler les signaux en temps) de façon simultanée. Dans notre cas, nous avons choisi de ne pas aller plus loin que 10 km pour observer les conséquences sur les performances tout en respectant amplement le rayon des 500 mètres réglementaires relatif à la protection des mammifères marins en présence de tirs sismiques. Voici la courbe des performances de reconnaissance en fonction de la distance à la source (*cf.* 3.15) :

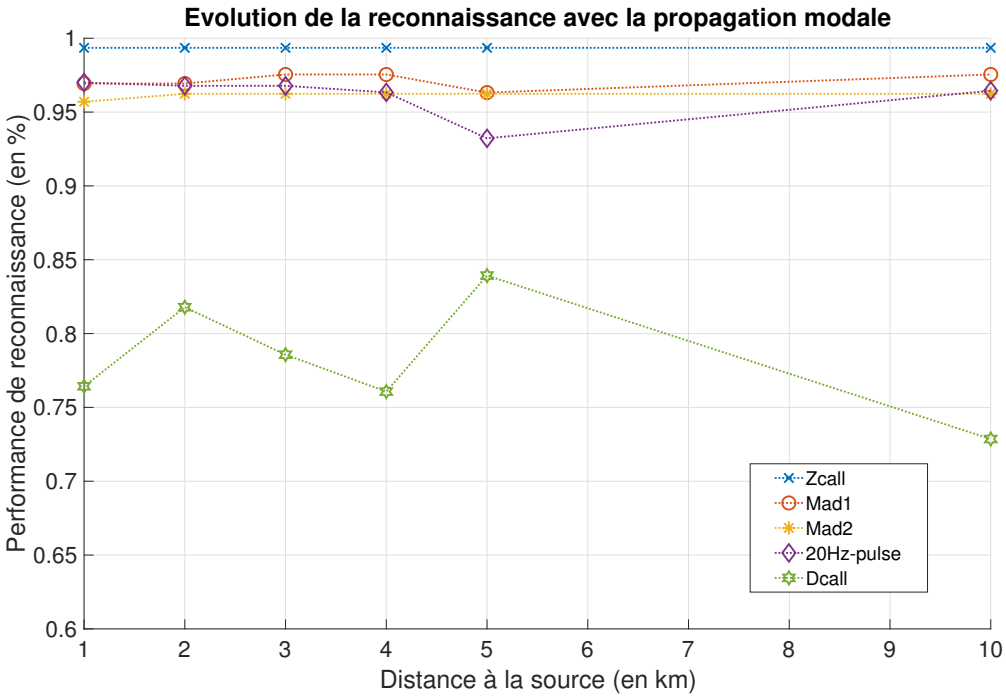


FIGURE 3.15 – Performances de reconnaissance en fonction de la distance à la source (K = 3 et $N'_c = 100$).

Ces résultats sont quasi-constants avec l'éloignement de la source et subissent peu de dégradation par rapport à la dispersion modale générée. La méthode est donc suffisamment robuste par rapport à la dispersion modale simulée dans un modèle de Pekeris. Pour finir, au cas où les résultats ne seraient pas suffisants, nous pouvons noter qu'il est toujours possible d'étendre la connaissance du système de reconnaissance (le dictionnaire) en apprenant des vocalises qui ont subi de la dispersion modale et/ou de prendre en compte la transformation considérée en l'appliquant aux atomes du dictionnaire.

3.3 Auto-apprentissage incrémental semi-supervisée

Cette section présente un test qui permet de mettre en avant une propriété forte de la méthode SINR-SRC qui est de pouvoir réaliser de l'auto-apprentissage incrémental semi-supervisé. Dans notre cas, nous avons choisi des vocalises de baleine bleue de la base MOBYSOUND.org [MOBYSOUND.ORG] qui sont, comme nous l'avons souligné au chapitre 2, des données de mauvaise qualité. Nous avons donc conservé les données « exploitables » qui constituent trois types de vocalises que nous pouvons observer sur le spectrogramme représenté figure 3.16 (avec les annotations dans des rectangles de couleur).

Nous avons réalisé un premier test qui consiste à considérer ces vocalises comme une classe inconnue et donc non apprise par le dictionnaire. Nous avons obtenu **100 %** des vocalises qui ont été identifiées comme « classe inconnue » (sur 486 vocalises). Nous avons ensuite appris une centaine de vocalises pour constituer un dictionnaire et nous avons évalué ses performances sur l'ensemble des vocalises restantes devenue notre nouvelle base de test (386 vocalises). Nous avons obtenu **100 % de bonne reconnaissance** sans seuil de gestion de classe inconnue (option de rejet désactivée) et **97 % de bonne reconnaissance** avec seuil (option de rejet activée). En conclusion de ce test de faisabilité, nous mettons en avant le fait que notre méthode a un très fort potentiel applicatif et pratique non seulement par sa construction, mais aussi par rapport à la littérature bioacoustique actuelle qui ne propose aucune méthodes aussi flexibles à notre connaissance.

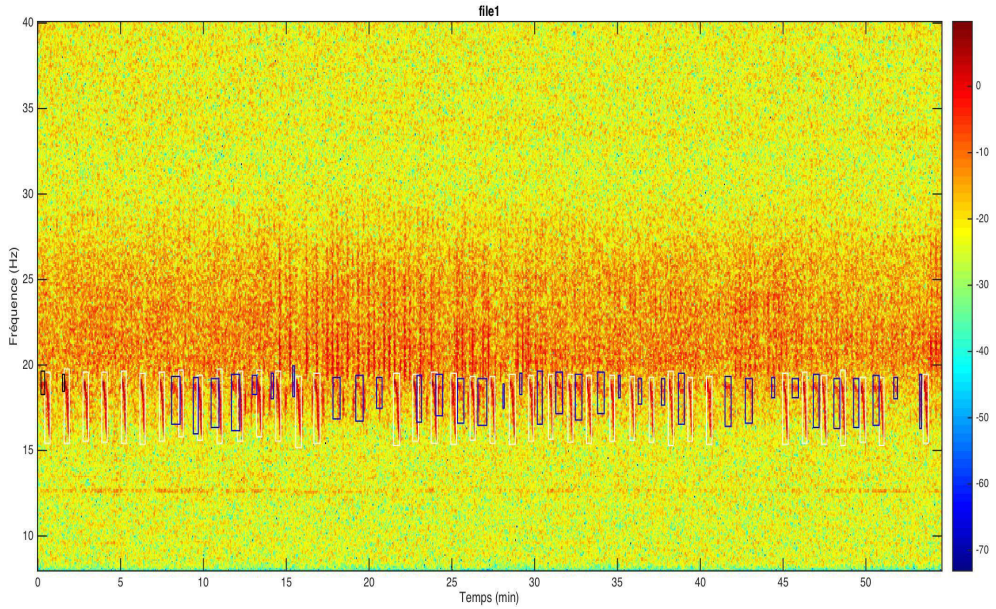


FIGURE 3.16 – Vocalises de baleines bleues de Mobysound

3.4 Niveau de confiance

Précédemment, nous avons évalué les performances de notre méthode et la robustesse par rapport aux choix des paramètres. A présent, nous proposons de mettre en œuvre un indice de confiance associé à chaque reconnaissance effectuée par notre méthode. Nous présentons ci-dessous deux métriques : une première pour le rejet du bruit et une seconde associée à la reconnaissance des vocalises. D'un point de vue théorique, lorsque nous considérons les SINR des bruits (gaussiens et/ou réels) et ceux de la base de test, alors nous nous trouvons dans un problème de détection multi-hypothèses avec les hypothèses suivantes :

— $\mathscr{H}_0$: « Le signal observé est du bruit »

— $\mathscr{H}_{1\leq i\leq \mathrm{C}}$: « Le signal observé est une vocalise de la classe i »

Dans notre cas, nous favorisons la gestion de la fausse alarme et nous plaçons donc le seuil avec pour exigence que la probabilité de retourner le label « bruit » quand l'observation est véritablement du bruit doit être très élevée, soit $\mathbb{P}(x < x_0|\mathscr{H}_0) = 0.999$ avec x le SINR de notre observation y et x_0 la valeur du seuil.

Nous respectons cette contrainte par notre choix du seuil, en utilisant notre hypothèse sur la distribution des SINR des exemples de bruits synthétiques, c'est-à-dire qu'un ensemble de « bruits gaussiens » de même durée que des vraies vocalises et filtrées en fréquence dans les mêmes plages que des vocalises peut être représentatif de l'ensemble des bruits réels à travers chaque dictionnaire.

Dans le même esprit, nous proposons d'utiliser cette distribution pour définir un niveau de confiance associé à la reconnaissance d'une observation d'un bruit, c'est notre première métrique. Plus précisément, nous proposons un **pourcentage de confiance** directement relié à la fonction de répartition empirique de la distribution des SINR des bruits gaussiens (ici, cette distribution est considérée comme une « densité » de probabilité de la variable aléatoire notée $\mathrm{SINR_B}$). Formellement ce niveau de confiance associé à l'estimation du label $\hat{\ell}_y$ de l'observation y se définit tel que :

$$\text{Confiance}(\hat{\ell}_y) = \begin{cases} \mathbb{P}(\mathrm{SINR_B} > x) & \text{si } x \leq x_0 \\ 0 & \text{sinon} \end{cases}$$

avec x le SINR de l'observation y, x_0 la valeur du seuil et $\mathbb{P}$ la mesure de probabilité associée. Pour bien comprendre cette expression, il suffit de noter que le niveau de confiance tend vers 0

quand x s'approche de x_0 et tend vers 100% lorsque x devient suffisamment petit.

En image (*cf.* figure 3.17), voici les étapes de construction du niveau de confiance basé sur la fonction de répartition empirique (noté F_x) associée à la distribution des SINR des bruits gaussiens :

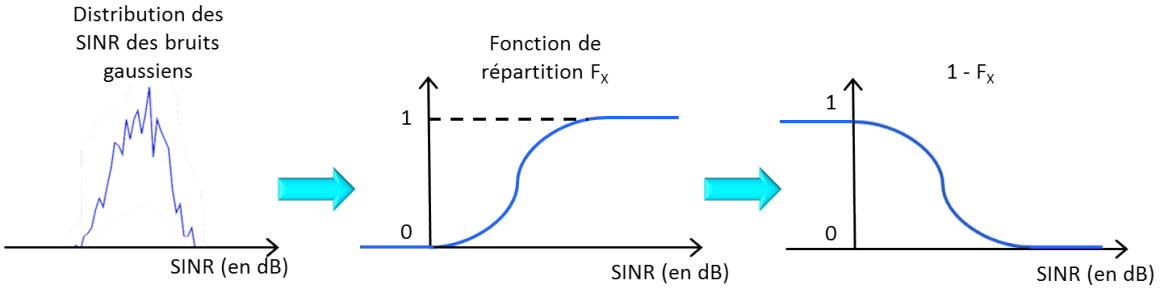


FIGURE 3.17 – Représentation schématique du passage de la distribution des SINR des bruits gaussiens à $1 - F_X$

De cette façon, il nous est possible d'avoir le schéma suivant (*cf.* figure 3.18), qui représente le niveau de confiance associé au rejet du bruit pour une distribution de SINR de bruits gaussiens :

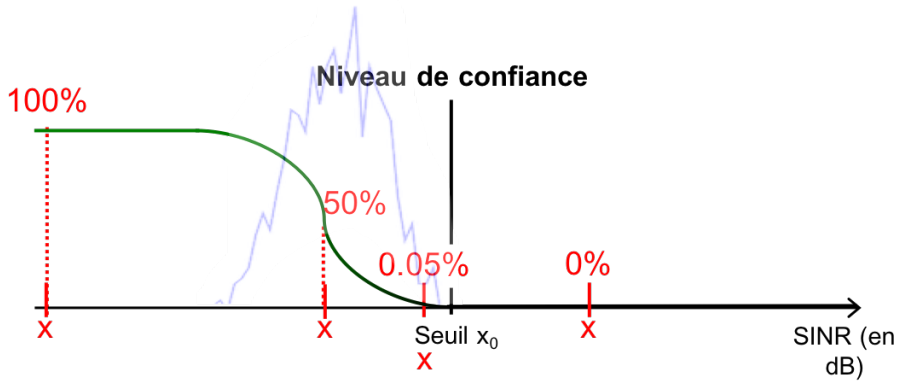


FIGURE 3.18 – Représentation schématique du niveau de confiance associé à la reconnaissance du bruit

Nous pouvons observer quelques x rouges qui représentent chacun une observation considérée ici comme du bruit. Le niveau de confiance se lit par projection sur notre fonction définie ci-dessus. Nous retrouvons bien le fait que lorsque le SINR d'une observation est à droite du seuil alors elle a 0 % de chance d'être considérée comme du bruit.

Certes ce niveau de confiance est tout à fait « naturel » du point de vue du rejet des bruits, mais celui-ci ne permet pas de donner un niveau de confiance du point de vue de la reconnaissance des vocalises de la classe considérée. En effet, il faudrait connaître la distribution des SINR de la base de test pour transposer le raisonnement fait précédemment, ce qui n'est *a priori* pas possible. Néanmoins, il est envisageable d'utiliser une partie de la base d'apprentissage et d'en faire une base auxiliaire (première base de test) ce qui nous permettrait d'estimer la vraie distribution de la base de test. Cela représente notre seconde métrique. A la fin, en notant SINR_{Aux} la distribution des SINR de la base auxiliaire nous avons la fonction suivante :

$$\text{Confiance}(\hat{\ell}_y) = \begin{cases} \mathbb{P}(\text{SINR}_{\text{B}} > x) & \text{si } x \leq x_0 \\ \mathbb{P}(\text{SINR}_{\text{Aux}} < x) & \text{sinon} \end{cases}$$

avec x le SINR de l'observation y et x_0 la valeur du seuil.

Ainsi nous obtenons finalement l'ensemble de confiance suivant :

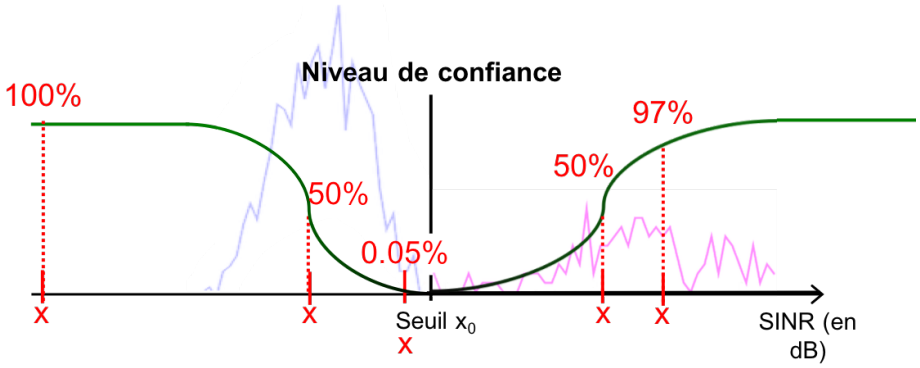


FIGURE 3.19 – Représentation schématique le niveau de confiance associé à la reconnaissance du bruit (ou des classes inconnues) et des vocalises)

Pour chaque x nous pouvons lire la confiance de part et d'autre du seuil, c'est-à-dire que nous avons le niveau de confiance quand l'observation est considérée comme étant du bruit (à gauche du seuil) et quand elle est identifiée comme étant une vocalise de la classe considérée (à droite du seuil).

3.4.1 Généralisation de la méthode avec mise en pratique du niveau de confiance

Comme nous l'avons vu au chapitre 2, les performances de généralisation d'une méthode de reconnaissance représentent les performances attendues sur de nouvelles données. La mesure de la généralisation se fait alors sur des données de tests que la méthode n'a pas apprises. Néanmoins, la démarche généralement employée peut biaiser les résultats. Sans perte de généralité, prenons le cas où nous avons une base de test et une base d'apprentissage. Imaginons que le premier test sur la base de test ne soit pas très concluant, alors nous allons amélioré et/ou changé la méthode jusqu'à obtenir de suffisamment « bons résultats » sur cette même base de test. En procédant ainsi, à chaque test effectué sur la base de test, nous avons « appris » et/ou « exploré » un peu plus cette base de test. En conséquence, manuellement ou non, nous finissons par converger vers une méthode qui a été conditionnée par cette base de test. Il devient alors difficile de garantir la capacité de généralisation de la méthode. Voici pourquoi, il convient d'avoir une base de données représentatives du contexte, mais dont aucun test n'a été effectué afin de mesurer véritablement la méthode quand elle se trouve en situation réelle.

Dans notre cas, nous avons utilisé une validation des performances par validation croisée (choix aléatoire de base d'apprentissage et de base de test) pour valider nos résultats et être le plus indépendant possible aux données d'apprentissage. Néanmoins, il reste une possibilité d'être dépendant à la base de données utilisée et dans l'idéal, il convient de tester la méthode sur des données complètement nouvelles par rapport aux données apprises, mais qui contiennent les mêmes classes de signaux à identifier. Afin de considérer cette problématique, nous proposons par la suite une démonstration de faisabilité. Nous testons ici la généralisation de la méthode, ainsi que la mise en pratique du niveau de confiance, sur une base de test différente de celles déjà vues auparavant par la méthode. Il s'agit de vocalises de rorqual commun données par SERCEL. Cela signifie que les données ne sont pas enregistrée avec un hydrophone fixe, mais par un bateau en mouvement et surtout les données proviennent d'une zone de campagne sismique où des tirs d'airgun sont présents au contraire des données de la base DEFLOHYDRO.

Après avoir appris un dictionnaire sur les rorquals communs de DEFLOHYDRO, nous avons eu 100 % de bonnes reconnaissances des vocalises de rorquals communs de SERCEL sans seuil de gestion de bruit et 85 % avec le seuil ce qui prouve, dans cette situation, que la méthode a *a priori* une bonne capacité de généralisation. Cet exemple est l'occasion de mettre en pratique le niveau de confiance (défini précédemment). Nous présentons ici les distributions des bruits gaussiens avec celui d'une base de test auxiliaire (200 vocalises de test de « 20Hz-pulse » de DEFLOHYDRO)

avec les deux ensembles de confiance associés :

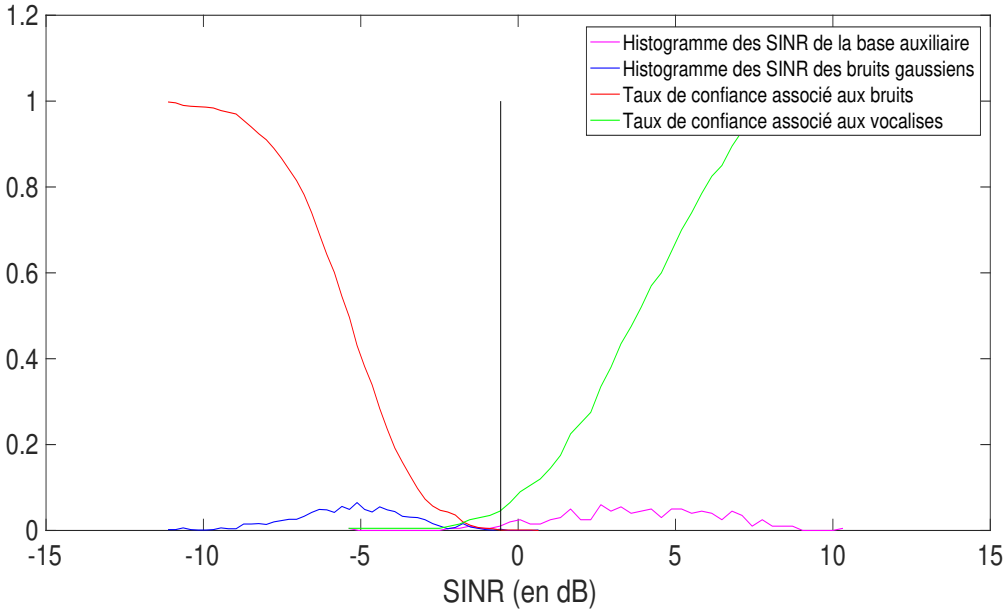


FIGURE 3.20 – SINR des bruits gaussiens et de la base de test auxiliaire avec les niveaux de confiance associés

Ainsi, en « fusionnant » les informations des deux niveaux de confiance, il est alors possible d'appliquer cet ensemble sur la base de test des « 20Hz-Pulse » de SERCEL :

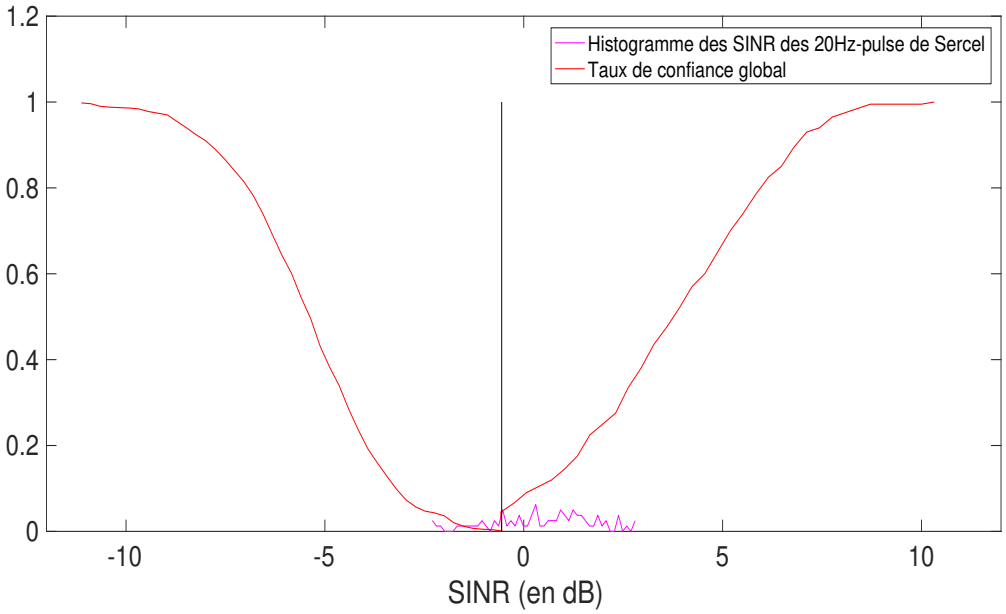


FIGURE 3.21 – SINR des bruits gaussiens et de la base de test auxiliaire avec les niveaux de confiance associés

Nous pouvons observer la distribution des SINR de la base de test des vocalises de SERCEL qui sont reconnues ici avec un niveau de confiance moyen de 13%±10% sur l'ensemble des 100 vocalises. Plus précisément, nous avons en moyenne 2%±2% de confiance pour les observations reconnues comme étant du bruit (15 vocalises) et en moyenne 16%±1% pour les observations reconnues comme des vocalises (85 vocalises). Les valeurs de confiance obtenues sont faibles ce qui est cohérent car même si la méthode reconnaît les vocalises de la base de données de SERCEL, ces données restent différentes de celles apprises par la méthode (durée des vocalises de SERCEL plus courte et dans un environnement marin différent de celle apprise sur la base DEFLOHYDRO).

Pour confirmer notre interprétation, nous effectuons un test supplémentaire qui va prendre en compte l'apprentissage des nouvelles vocalises pour visualiser son effet sur la distribution des SINR de la base de test. Nous divisons la base de données sercel en 20 vocalises d'apprentissage et 80 de tests. Après avoir effectuer l'apprentissage du dictionnaire, la méthode reconnaît 75 vocalises dans la classe « 20Hz-pulse SERCEL » (93.75%) et 5 comme étant « inconnues » (6.25%), voici les distributions des SINR obtenus :

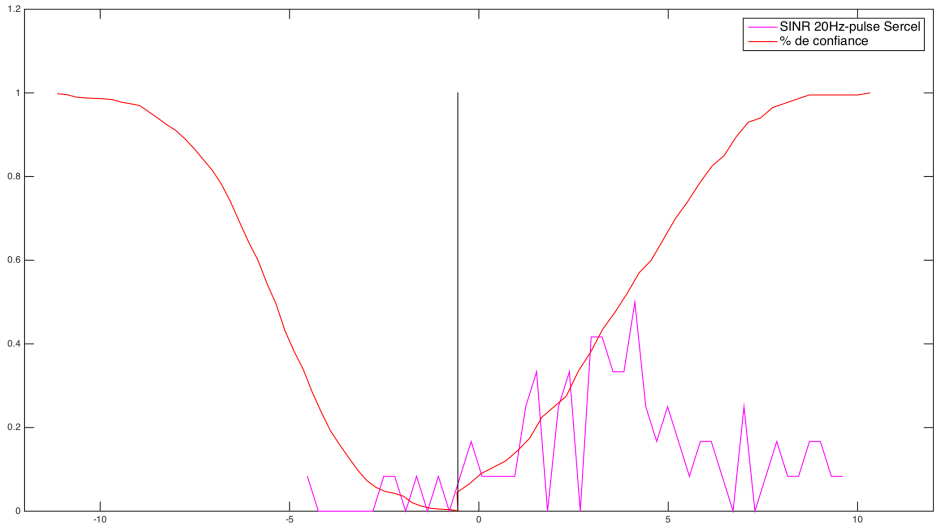


FIGURE 3.22 – SINR des bruits gaussiens et de la base de test auxiliaire avec les niveaux de confiance associés

Le niveau de confiance moyen est de 51%±30% sur l'ensemble des 80 vocalises de test. Plus précisément, nous avons en moyenne 9%±13% de confiance pour les observations reconnues comme étant du bruit (5 vocalises) et en moyenne 53%±29% pour les observations reconnues comme des vocalises (75 vocalises).

Ces résultats confirment que la distribution des SINR « s'étale » avec uniquement quelques vocalises apprises par le système (ici seulement 20). Même si le dictionnaire construit n'est pas suffisamment représentatif de la classe considérée, nous observons une augmentation significative du niveau de confiance qui vient confirmer qu'effectivement la méthode a plus « d'expérience » vis-à-vis des données SERCEL.

3.5 Conclusion

Les représentations parcimonieuses se sont avérées efficaces pour classer les vocalises basses fréquences des mysticètes. De telles représentations modélisent les vocalises comme des combinaisons linéaires d'atomes dans un dictionnaire (base sur-complète) dans lequel de nombreux coefficients sont nuls. Dans ce cadre, le classifieur cherche à approcher les signaux de test donnés en entrée avec (peu de) combinaisons linéaires des vocalises apprises précédemment et associe l'étiquette de la classe qui donne la meilleure approximation. La méthode proposée traite directement les signaux digitaux en représentation temporelle et ne subit donc aucune perte d'information due à une éventuelle projection dans un autre espace (comme cela peut être fait lors de l'extraction de caractéristiques à partir de spectrogrammes ou de cepstres). Elle ne repose que sur quelques paramètres, à savoir : la taille du dictionnaire et la contrainte de parcimonie. Ces paramètres reflètent le degré de variabilité et de complexité d'une classe de vocalises donnée. Comme montré dans les expérimentations numériques, ces paramètres sont faciles à régler (la méthode

est robuste) et ne nécessitent pas un réglage fin.

Les représentations parcimonieuses permettent également de construire des métriques de confiance simples pour rejeter les données de bruit. La statistique SINR (3.7) a été utilisée à la sortie du classifieur et a rejeté 97.1 % des données de bruit réelles. Avec cette approche, le bruit est géré sans que l'algorithme n'apprenne les caractéristiques des données de bruit réelles. La méthode globale a été testée sur cinq types de vocalises de mysticètes avec des caractéristiques temps-fréquence qui se chevauchent et différents degrés de variabilité. Les résultats numériques ont montré que, sur l'ensemble de données de test, 92,1 % sont correctement reconnues en moyenne. Comme prévu, les vocalises stéréotypées, telles que les Z-call de baleine bleue Antarctique, sont plus faciles à reconnaître que les vocalises plus variables comme les D-call de la baleine bleue du Pacifique Nord, qui peuvent être incorrectement rejetées par la statistique SINR.

Les classes peuvent facilement être supprimées ou ajoutées à la méthode proposée. Cela peut être utile pour la surveillance acoustique passive opérationnelle où des informations préalables telles que l'emplacement du capteur et/ou la période de l'année Sont connues. Ces informations peuvent être prises en compte pour se concentrer sur des espèces spécifiques.

Dans un récent travail [SOCHELEAU et SAMARAN, 2017], les représentations parcimonieuses ont montré de bonnes performances pour détecter les vocalises de mysticètes. Une extension possible de ce travail serait donc de fusionner les deux approches pour détecter et classer conjointement les sons des mysticètes. Puisque les cris sont affectés par les conditions de propagation locales et par le bruit, il conviendrait également d'étudier l'avantage potentiel de construire des dictionnaires à partir d'un modèle(s) paramétrique(s) de cris plutôt qu'à partir des cris eux-mêmes. En outre, la statistique SINR pourrait être utilisée comme une mesure de confiance (liée à la position du seuil) et également comme un détecteur de nouveauté. De cette façon, l'algorithme SINR-SRC offrirait non seulement la capacité à rejeter le bruit, mais pourrait également être utilisé pour développer un algorithme automatique d'apprentissage incrémental semi-supervisé qui construit de nouveaux dictionnaires en ligne. Plus concrètement, après détection par l'algorithme SINR-SRC d'un signal structuré inconnu, un analyste humain pourrait l'étiqueter et décider de l'ajouter à un nouveau dictionnaire pour la raconnaissance automatique des occurrences futures de cette nouvelle classe de signaux. Dans la partie suivante, nous allons développer l'idée d'un détecteur-classifieur basé sur la méthode SINR-SRC .

Chapitre 4

Une extension de SINR-SRC : le détecteur multiclasses

« *Il faut toujours penser par soi-même. Ne rien apprendre par cœur, mais tout redécouvrir et, en tout cas, ne rien accepter qui ne soit prouvé. Ne rien négliger de ce qui est concevable ou imaginable.* »

— Albert Einstein

Sommaire

Dans ce chapitre, nous proposons de développer une extension de l'algorithme SINR-SRC afin qu'il puisse réaliser la détection et la reconnaissance conjointement. Les données d'entrée ne sont plus traitées à la sortie d'un détecteur, mais au contraire, le signal brut est balayé et traité. Notre méthode, dénommée le *détecteur multiclasses*, est capable de détecter et reconnaître les signaux d'intérêt avec un coefficient de confiance (indice de confiance et/ou SNR) pour chaque détection. Après avoir présenté la mise en œuvre de la méthode, nous discuterons des résultats de détection et de reconnaissance. Enfin, nous discuterons des éventuelles perspectives d'application de notre méthode comme l'obtention de résultats statistiques (nombre de vocalises, qualité des signaux, etc.), l'annotation automatique de fichier et l'apprentissage incrémental de nouveauté de façon semi-supervisée, pour des applications temps-réels ou non.

4.1 Mise en œuvre d'un détecteur multiclasses

4.1.1 Procédure générale

Comme nous l'avons vu au chapitre 3 à la section 3.4, l'utilisation de l'option de rejet (seuillage du SINR) nous permet de nous trouver dans un problème de détection multi-hypothèses :

— $\mathscr{H}_0$: « Le signal observé est du bruit »

— $\mathscr{H}_{1\leq i\leq \mathrm{C}}$: « Le du signal observé est une vocalise de la classe i »

De cette façon, plutôt que de considérer le système de reconnaissance en sortie d'un détecteur, nous proposons d'appliquer SINR-SRC directement comme « détecteur-classifieur ». Afin de prendre en compte les différentes propriétés temps-fréquence (connaissance *a priori*) des différentes classes de signaux apprises par le dictionnaire, nous considérons autant de « fenêtres d'observation » qu'il existe de classes. Une fenêtre d'observation correspond à la sortie du signal d'entrée après filtrage temps-fréquence, ce filtrage permet d'extraire le « rectangle » temps-fréquence correspondant à l'espace d'observation dans lequel existe potentiellement les vocalises de la classe considérée. La méthode générale traite chaque classe en parallèle à partir des données d'entrées (*cf.* figure 4.1). Pour chaque classe $1 \leq i \leq \mathrm{C}$, le processus est résumé comme suit et est schématisé dans la figure 4.2.

1. Une « fenêtre d'observation » en temps-fréquence est définie pour la classe $1 \leq i \leq \mathrm{C}$ considérée. Elle est définie par deux applications. La première application est une fenêtre d'extraction temporelle de durée T^i secondes. T^i est égal à la durée maximale des signaux du dictionnaire de la classe i considérée. La seconde application est un filtre passe-bande $h^i_{[f_{min},f_{max}]}$ (avec une fréquence de coupure à -6 dB) construit tel que f_{min} et f_{max}, les bornes fréquentielles minimum et maximum, correspondent à la plage d'observation fréquentielle des vocalises de la classe i considérée.
2. Pour un signal de test donné $\boldsymbol{s}$ de durée totale $\mathrm{T}_s \geq \mathrm{T}^i$. Le filtre $\boldsymbol{h}^i_{[f_{min},f_{max}]}$ est appliqué à $\boldsymbol{s}$ et la fonction d'extraction temporelle est utilisée pour balayer le vecteur $\boldsymbol{s}$ avec un pas de lecture Δ tel que, après extraction temporelle, nous avons :

$$\boldsymbol{s}^i_k = \boldsymbol{s}[k.\Delta, k.\Delta + \mathrm{L}^i - 1] * \boldsymbol{h}^i_{[f_{min},f_{max}]},$$

   où L^i est le nombre d'échantillons correspondant à la durée de T^i secondes et k est un entier allant de 0 à $\lfloor \mathrm{T}_s/\mathrm{T}^i \rfloor$ par pas de 1.
3. Pour chaque portion $\boldsymbol{s}^i_k$ du signal $\boldsymbol{s}$, l'algorithme SRC est appliqué et donne en sortie une estimation de la classe $1 \leq j \leq \mathrm{C}$ associée à la classe la plus proche de l'observation $\boldsymbol{s}^i_k$
4. Si la classe j estimée est différente de la classe i considérée alors l'observation est rejetée. Sinon, le SINR est calculée et géré comme dans SINR-SRC, à savoir : si le SINR est supérieur à un certain seuil alors le résultat fourni par SRC est validé ; sinon $\boldsymbol{s}^i_k$ est rejeté.

5. Si l'observation est rejetée, alors aucune détection n'est comptée. Sinon, la détection n'est pas faite tout de suite, mais mise en attente. Les valeurs du SINR sont alors mémorisées tant que chaque nouvelle observation $\boldsymbol{s}_k$ est validée comme étant de la classe i. A la première observation $\boldsymbol{s}_k$ qui n'est pas validée, la détection est réalisée de la façon suivante. Si les valeurs mémorisées du SINR forment un « support trop grand » (l'ensemble des valeurs dépasse la durée de la fenêtre d'observation de la classe i) alors la détection est rejetée. Sinon, la détection est validée. Elle débute à compter du maximum des valeurs du SINR mémorisées et dure T^i secondes. Les valeurs du SINR comprises entre le début et la fin de la détection ne sont alors pas prise en compte.

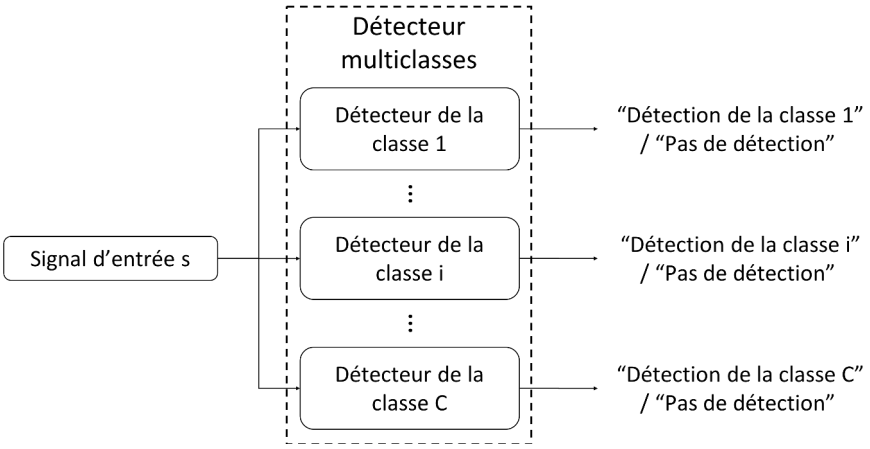


FIGURE 4.1 – Représentation générale du détecteur multiclasses. Notons l'architecture de traitement parallèle associé à chaque classe.

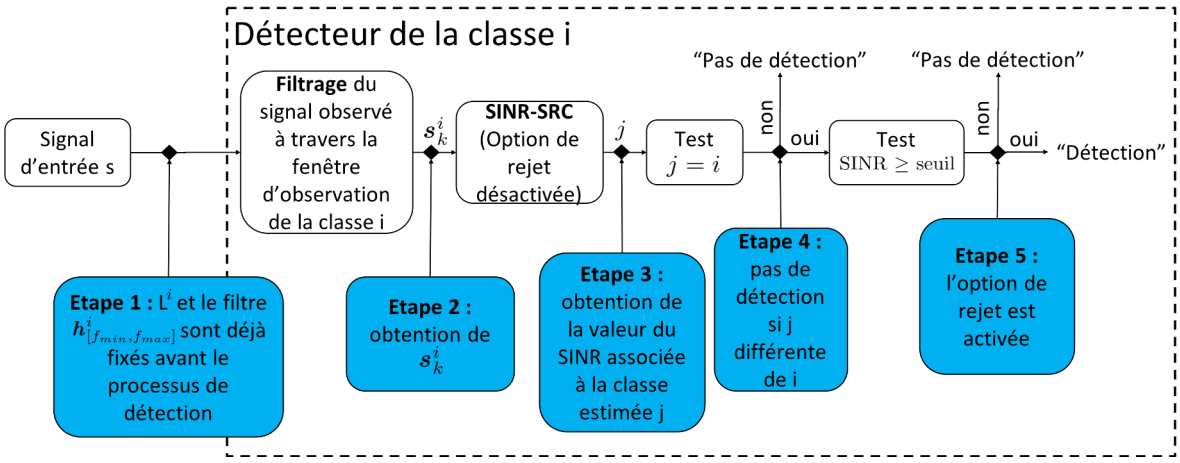


FIGURE 4.2 – Représentation des étapes du détecteur multiclasses pour une classe

4.1.2 Illustration de la mise en œuvre du détecteur multiclasses sur un exemple

Notre méthode s'applique **sur le signal temporel** et de façon séquentielle. Nous proposons de visualiser le processus de façon globale jusqu'à la détection. Pour ce faire, nous considérons un signal contenant deux classes, ici des vocalises correspondantes à des Z-call et à des 20Hz-pulse extraites de la base de données DEFLOHYDRO (*cf.* figure 4.3). L'intérêt de ce signal est qu'il contient une zone bruitée (autour de 1500-1550 s) et permet de bien visualiser les deux types de vocalises qui sont annotées ici sur le spectrogramme (*cf.* figure 4.3). Les annotations correspondent aux rectangles temps-fréquence et sont la vérité terrain donnée par l'expert. A présent, nous balayons le signal suivant cinq fenêtres d'observations en appliquant SINR-SRC (avec l'option de rejet désactivée) et un dictionnaire contenant les cinq mêmes classes (Z-call, Mad1, Mad2, 20Hz-pulse et

D-call) qu'au chapitre précédent. Cette partie correspond à l'étape 3 (*cf.* figure 4.2) de notre méthode. Nous représentons ci-dessous les cinq ensemble de SINR obtenu suivant les cinq fenêtres d'observations (*cf.* figures 4.5 à 4.9). Dans la suite, nous allons observer différents SINR. Les SINR proviennent de fenêtre d'observation différentes. Pour bien se représenter quelles sont les fenêtres d'observation utilisées, nous affichons, sous forme de rectangles temps-fréquence colorés, une représentation schématique (à l'échelle) de notre étape 1 (*cf.* figure 4.2) des fenêtres d'observations utilisées sur un spectrogramme (*cf.* figure 4.4). Les couleurs utilisées sont représentatives des classes considérées tout au long de ce chapitre.

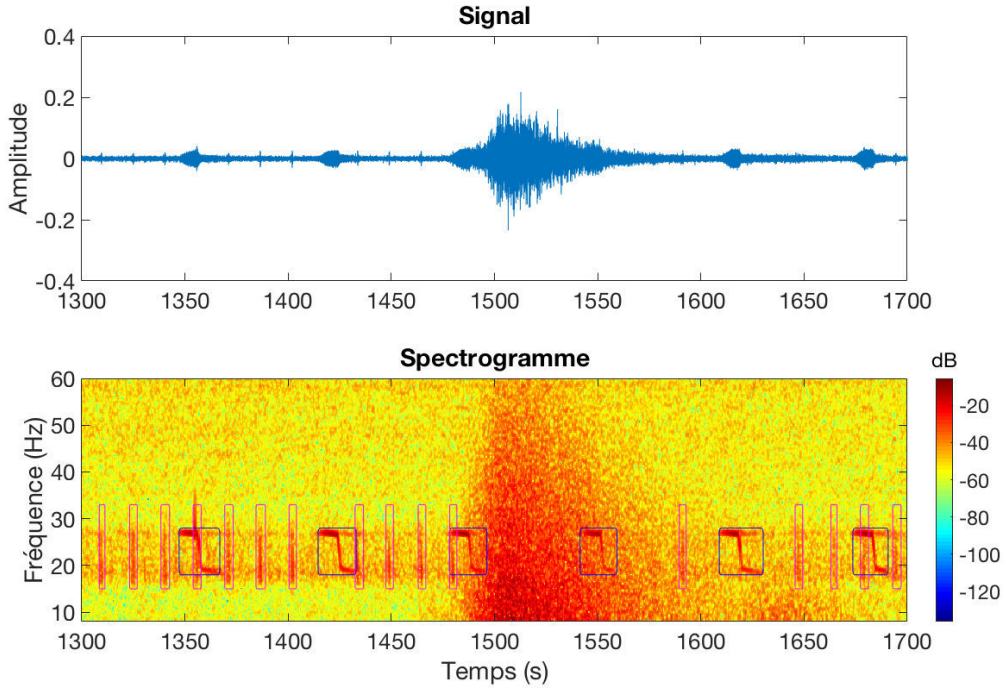


FIGURE 4.3 – Représentation temporelle (en haut) et spectrogramme avec annotations (en bas) d'un signal contenant deux classes (les Z-call annotés en bleu et les 20Hz-pulse annotés en magenta)

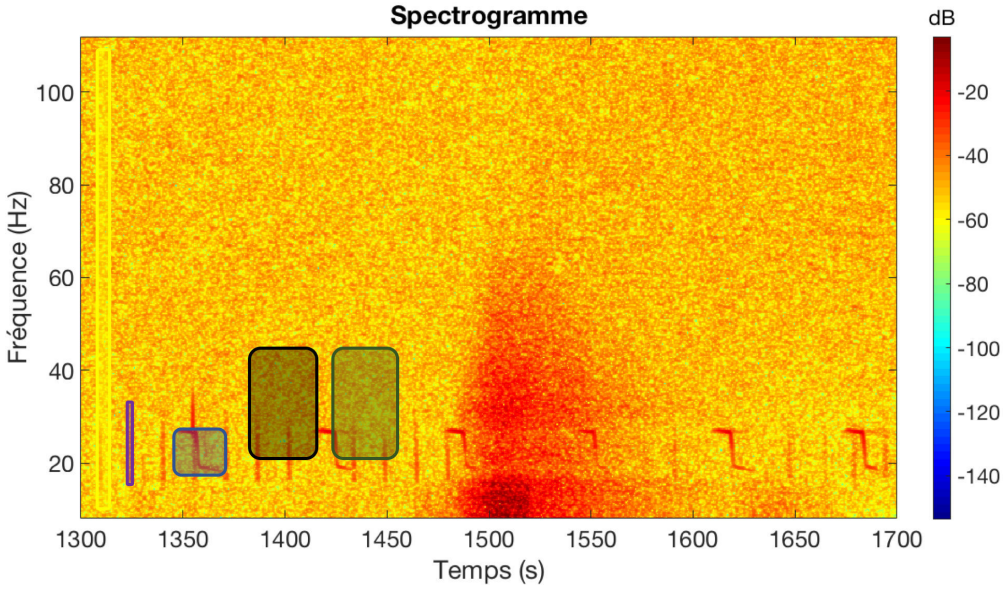


FIGURE 4.4 – **Etape 1 :** Représentation (à l'échelle) des fenêtres d'observations pour chaque classe sur le spectrogramme de Fourier, de gauche à droite : D-call (10-110 Hz, 6 s, en jaune), 20Hz-pulse (15-33 Hz, 2.64 s, en magenta), Z-call (18-28 Hz, 25.03 s, en bleue), Mad1 (21-45 Hz, 32.41 s, en noir) et Mad2 (21-45 Hz, 31.24 s, en vert)

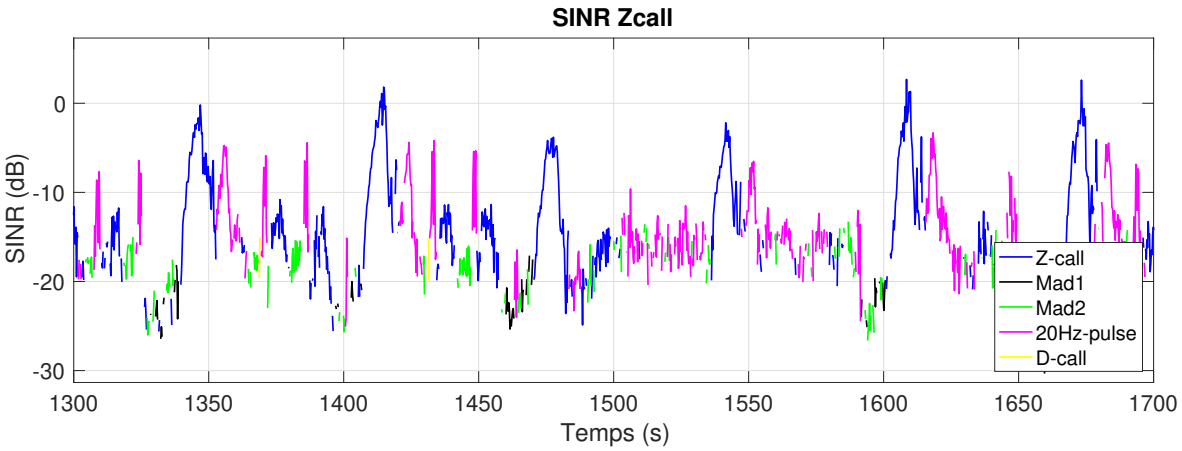


FIGURE 4.5 – **Etape 3 :** Ensemble des SINR obtenus après l'application de SINR-SRC (avec l'option de rejet désactivée) à travers la fenêtre d'observation de la classe Z-call. Les paramètres de la fenêtre d'observation sont 18-28 Hz en fréquence pour une durée de 25.03 s avec un delta de 240 ms. La légende indique la classe reconnue (notre code couleur) associée à la valeur du SINR considérée.

Nous observons les données de sortie de l'étape 3 (*cf.* figure 4.2). Chaque valeur du SINR est associée à une reconnaissance d'une des cinq classes du dictionnaire. La reconnaissance est représentée avec les mêmes couleurs que les fenêtres d'observation de la figure 4.4. Le « SINR Z-call » (*cf.* figure 4.5) nous indique que l'algorithme SINR-SRC avec l'option de rejet désactivée, à travers la fenêtre de la classe Z-call, identifie bien des Z-call (en bleu) en formant des « pics » supérieur à -5 dB. Nous pouvons également noter que certains 20Hz-pulse (en magenta) sont identifiables par des pics dont les valeurs dépassent -10 dB. Dans notre méthode, les 20Hz-pulse sont gérés par la fenêtre d'observation associée à cette classe et non par la fenêtre des Z-call qui n'est pas adapté à cette classe. Néanmoins, il est intéressant de noter l'effet de la fenêtre d'observation des Z-call sur le SINR associé aux 20Hz-pulse. Par construction, la fenêtre des Z-call est environ 9,5 fois plus large temporellement que celle des 20Hz-pulse. Cela signifie que, par rapport à la fenêtre d'observation des 20Hz-pulse, elle prend en compte le « voisinage temps-fréquence » des 20Hz-pulse. Ce voisinage peut potentiellement contenir des signaux indésirables et rendre difficile la reconstruction de l'observation. Ce phénomène s'observe sur les valeurs du SINR. Au fur et à mesure que la fenêtre d'observation balaye le signal et prend en compte le fort bruit de fond (autour de 1500 s) alors les valeurs des SINR associés aux 20Hz-pulse diminuent. Enfin, notons que le système génèrent des fausses reconnaissances avec les classes Mad1, Mad2 et D-call mais avec des valeurs de SINR plus faibles que celles de la classe Z-call considérée.

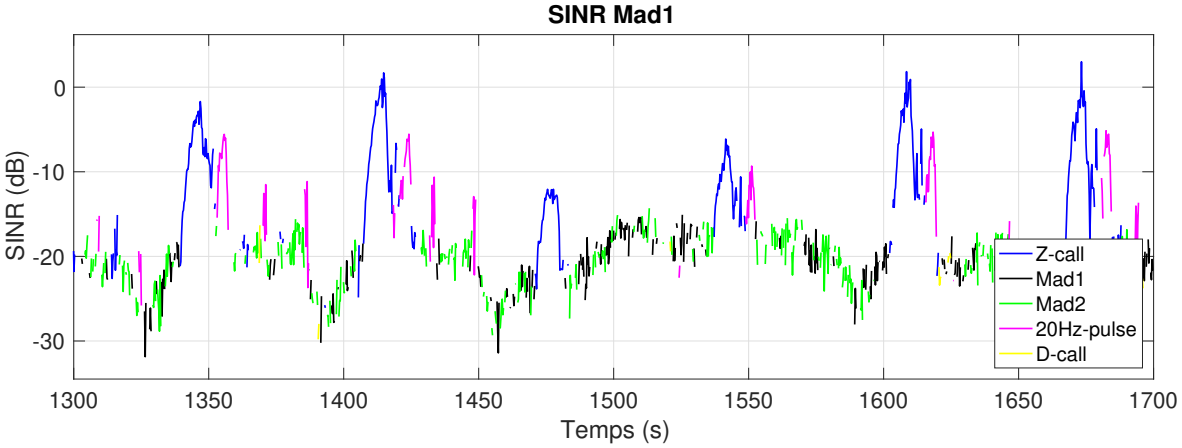


FIGURE 4.6 – **Etape 3 :** Ensemble des SINR obtenus après l'application de SINR-SRC (avec l'option de rejet désactivée) à travers la fenêtre d'observation de la classe Mad1. Les paramètres de la fenêtre d'observation sont 21-45 Hz en fréquence pour une durée de 32.42 s avec un delta de 240 ms.

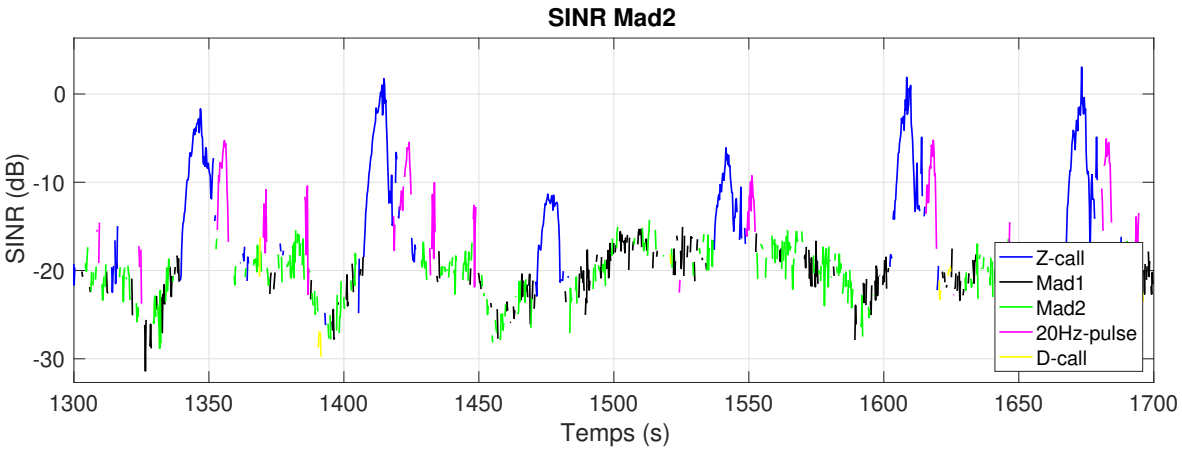


FIGURE 4.7 – **Etape 3 :** Ensemble des SINR obtenus après l'application de SINR-SRC (avec l'option de rejet désactivée) à travers la fenêtre d'observation de la classe Mad2. Les paramètres de la fenêtre d'observation sont 21-45 Hz en fréquence pour une durée de 31.24 s avec un delta de 240 ms.

Les fenêtres d'observation de Mad1 et Mad2 sont très proches (à 1 s près en durée). Les SINR obtenus sont très proches également (*cf.* figures 4.6 et 4.7). Malgré la plage de fréquence de 21 à 45 Hz qui ne « voit » plus le signal de 18 à 21 Hz comme la fenêtre précédente, nous pouvons noter que les vocalises de la classe Z-call sont toujours reconnues par la méthode mais avec des valeurs de SINR plus faibles que pour le cas du SINR Z-call. Comme pour les 20Hz-pulse à travers la fenêtre des Z-call, les fenêtres d'observations Mad1 et Mad2 ne sont pas adaptées aux Z-call. Par exemple, nous observons que les valeurs du SINR reconnues comme Z-call diminuent lorsque le bruit de fond est très présent dans la fenêtre d'observation Mad1 et Mad2 (3^{e} Z-call autour de 1500 s). A nouveau, notons que le système génèrent des fausses reconnaissances avec les classes Mad1, Mad2 et D-call mais sans valeurs significatives du SINR.

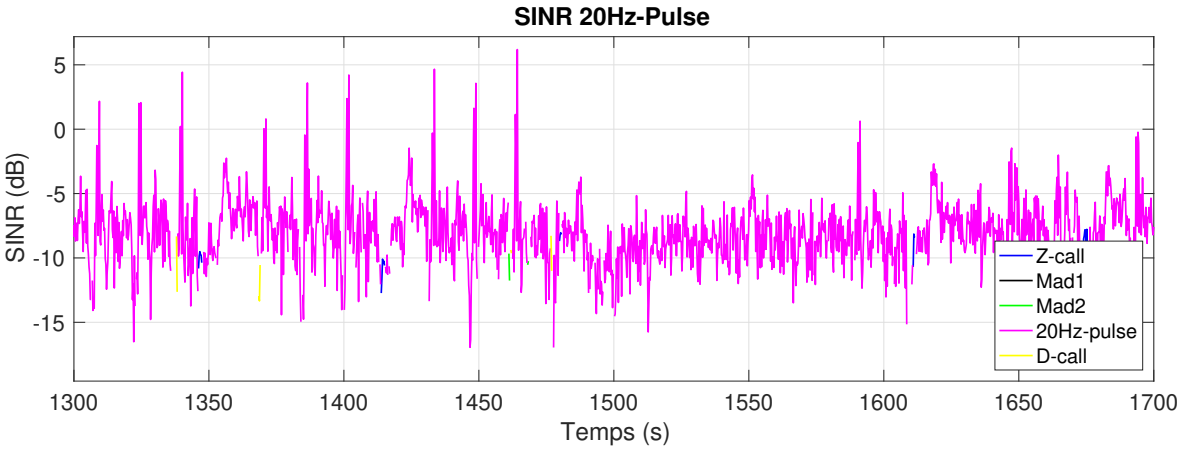


FIGURE 4.8 – **Etape 3 :** Ensemble des SINR obtenus après l'application de SINR-SRC (avec l'option de rejet désactivée) à travers la fenêtre d'observation de la classe 20Hz-pulse. Les paramètres de la fenêtre d'observation sont 15-33 Hz en fréquence pour une durée de 2.64 s avec un delta de 240 ms.

Concernant les valeurs des SINR associées à la fenêtre d'observation des 20Hz-pulse (*cf.* figure 4.8), elles sont beaucoup plus élevées (en moyenne autour de -7 dB) que les valeurs associées aux fenêtre précédentes. De plus, les identifications ne sont presque exclusivement que des 20Hz-pulse. La raison est que le filtrage temporel est tel que la fenêtre d'observation des 20Hz-pulse empêche de reconnaître les autres classes qui ont des propriétés temporelles très différentes. Notons que les pics à valeurs de SINR élevées (supérieur à -2 dB) coïncident avec les annotations des vocalises de 20Hz-pulse. Néanmoins, les seize 20H-pulse ne sont pas tous identifiés par des pics bien définis. L'explication est que les SNR des 20Hz-pulse considérés sonr plus faibles, ce qui empêche alors SINR-SRC de reconstruire le signal observé avec suffisamment peu d'erreur.

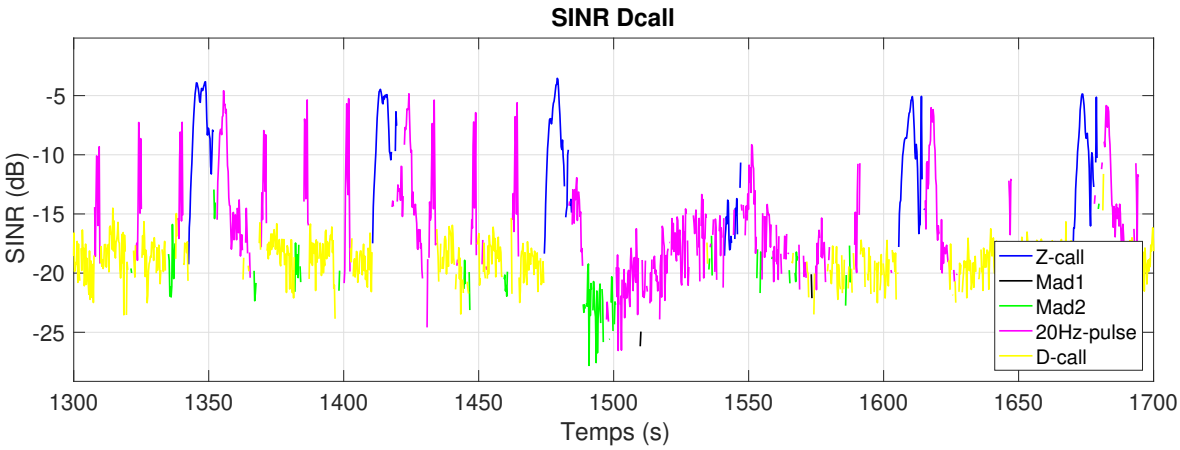


FIGURE 4.9 – **Etape 3 :** Ensemble des SINR obtenus après l'application de SINR-SRC (avec l'option de rejet désactivée) à travers la fenêtre d'observation de la classe D-call. Les paramètres de la fenêtre d'observation sont 10-110 Hz en fréquence pour une durée de 6 s avec un delta de 240 ms.

A travers la fenêtre d'observation des D-call, nous pouvons constater que les valeurs du SINR (*cf.* figure 4.9) sont du même ordre de grandeur que celle des classes Mad1 et Mad2. Ce sont d'ailleurs les valeurs de SINR, qui en moyenne, sont les plus basses avec les fenêtres d'observation les plus grandes. Intuitivement, nous pouvons penser que plus la fenêtre d'observation est grande, plus la probabilité d'augmenter l'erreur de reconstruction augmente, car plus difficile devient la reconstruction. Néanmoins, nous observons que les maxima locaux correspondant aux pics de SINR ne sont pas franchement moins élevés. Nous pouvons émettre l'hypothèse que la valeur du SINR dépend d'un ratio entre l'« occupation » de la vocalise considérée et la taille de la fenêtre.

4.1.3 Résultats de détection sur notre exemple

A présent, nous proposons d'illustrer la suite de la démarche utilisée en ne gardant uniquement que les valeurs des SINR associées aux deux classes présentes dans ce signal. Comme nous l'avons décrit plus haut, si la classe identifiée est différente du label de la fenêtre d'observation utilisée, alors les valeurs des SINR sont rejetées (étape 4 *cf.* figure 4.2). Voici les identifications après cette étape (*cf.* figures 4.10 et 4.11).

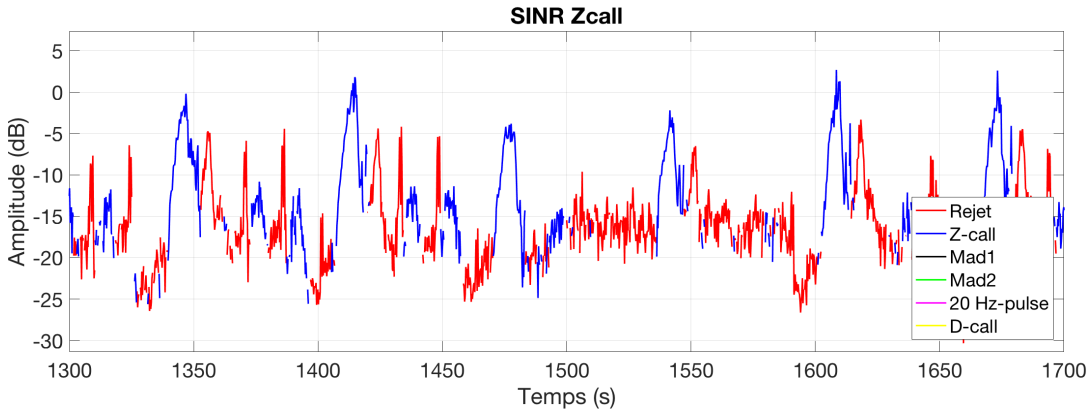


FIGURE 4.10 – **Etape 4 :** SINR associés à la classe Z-call après rejet des classes différentes du label de la fenêtre d'observation.

A présent il est possible d'appliquer la gestion du rejet par le seuillage des valeurs du SINR (étape 5 *cf.* figure 4.2). Dans le cas de cet exemple, nous proposons d'utiliser un seuil à -10dB pour gérer la classe des Z-call et un seuil à -3dB pour gérer les 20Hz-pulse. Nous avons précisé plus haut que le début de la détection correspond au maximum des valeurs du SINR dans le même

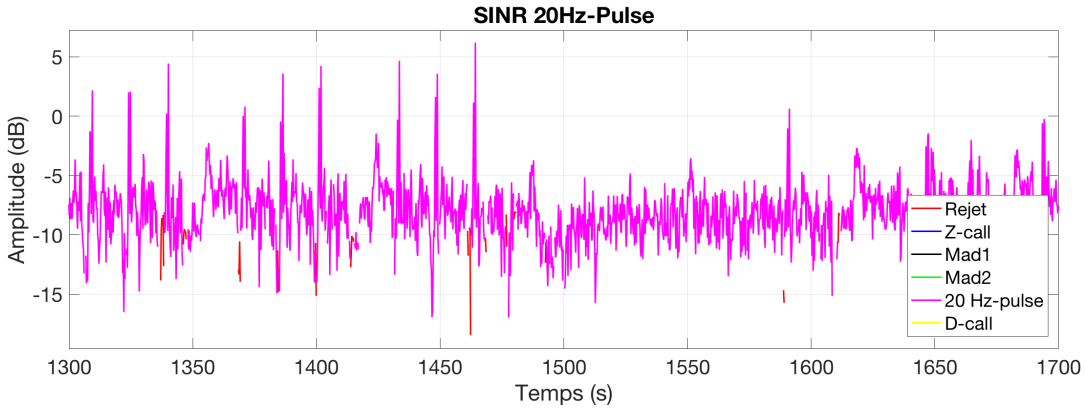


FIGURE 4.11 – **Etape 4 :** SINR associés à la classe 20Hz-pulse après rejet des classes différentes du label de la fenêtre d'observation.

« support continu ». De même nous faisons l'hypothèse qu'il n'existe qu'une unique vocalise de la chaque classe à l'intérieur de la fenêtre d'observation. D'un autre côté, si le support observé est « trop grand », la détection est rejetée. C'est le cas par exemple, si le seuil de détection est réglé trop bas. Nous affichons ci-dessous les valeurs avec les reconnaissances des SINR et les résultats de la détection retenue (*cf.* figure 4.12).

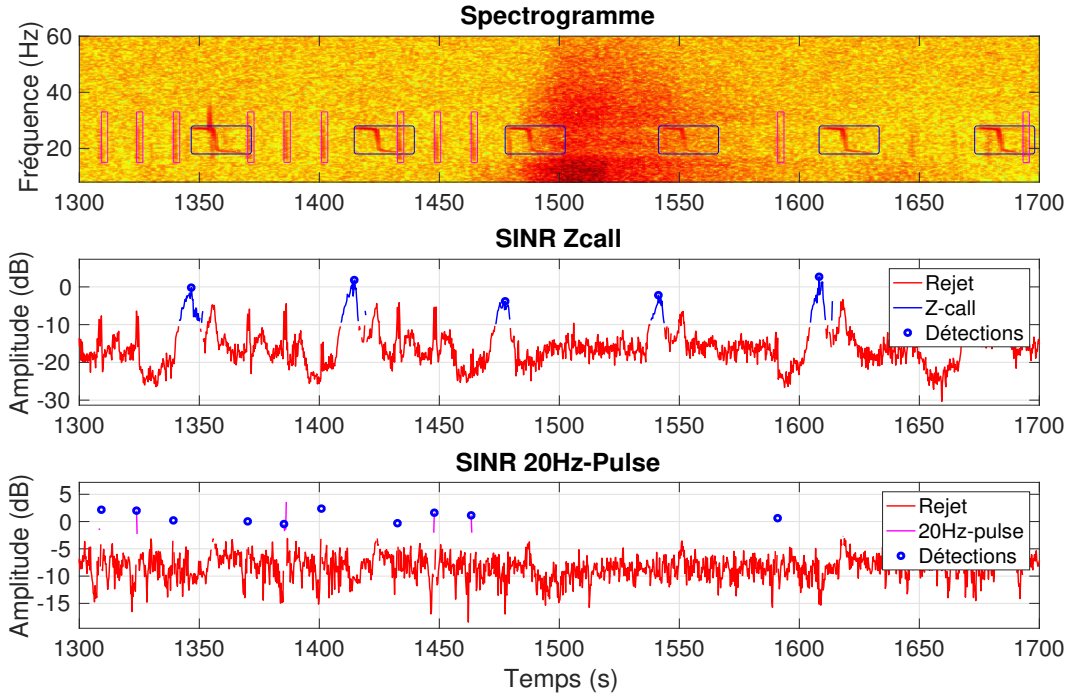


FIGURE 4.12 – **Etape 5 :** Visualisation de la détection (rectangles bleus pour les Z-call et magenta pour les 20Hz-pulse) sur le spectrogramme et valeurs des SINR associés (cercle bleus).

Nous rappelons que les valeurs des SINR compris entre le début et la fin d'une détection en sont pas pris en compte. Cela explique la 5^{e} détection des 20Hz-pulse qui ne semble pas correspondre au maximum du support des SINR considérés. En réalité, il s'agit de deux supports très proches et la détection a eu lieu sur le premier support. Le second support étant compris entre le début et la fin de la détection, il n'est pas pris en compte. Pour finir, comme nous pouvons le voir, le fait de prendre en compte des fenêtres d'observation indépendantes permet de faire de la multi-détections au sens où nous pouvons détecter plusieurs classes différentes au même instant.

4.2 Résultats expérimentaux

4.2.1 Le jeu de données utilisé

Comme au chapitre précédent, nous utilisons les quatre mêmes bases de données, à savoir les bases DEFLOHYDRO, OHASISBIO, DCLDE et les données de SERCEL. Néanmoins nous travaillons cette fois-ci, non pas sur des vocalises ou des bruits extraits de ces données, mais sur l'entièreté de ces signaux par balayage. La base de données représente 61h50min d'enregistrement. Par rapport à la base de données utilisée pour SINR-SRC, excepté la classe des D-call, la base de test a subi une évolution car les annotations des fichiers utilisés ne prenaient pas en compte l'entièreté des signaux considérés. Plus précisément, il n'existait pas de « zone temporelle continue bien annotée » permettant d'évaluer correctement les performances du point de vue de la détection. De ce constat, les données de tests des classes Mad1 et Mad2 sont modifiées. Quant aux données de test des classes Z-call et 20Hz-pulse, elles ont surtout subi une augmentation de taille. Voici plus en détails la composition de la base de données en termes de quantité de vocalises, de durée et de répartition des SNR (*cf.* figure 4.13) :

— Classe Z-call : 317 vocalises ou environ 2h
— Classe Mad1 : 321 vocalises ou environ 3h
— Classe Mad2 : 375 vocalises ou environ 3h
— Classe 20Hz-pulse : 3 242 vocalises ou environ 2h23min
— Classe D-call : 380 vocalises ou environ 38min
— bruit d'étude sismique provenant de SERCEL : 26h42min
— bruit océanique provenant de DEFLOHYDRO : 6h19min

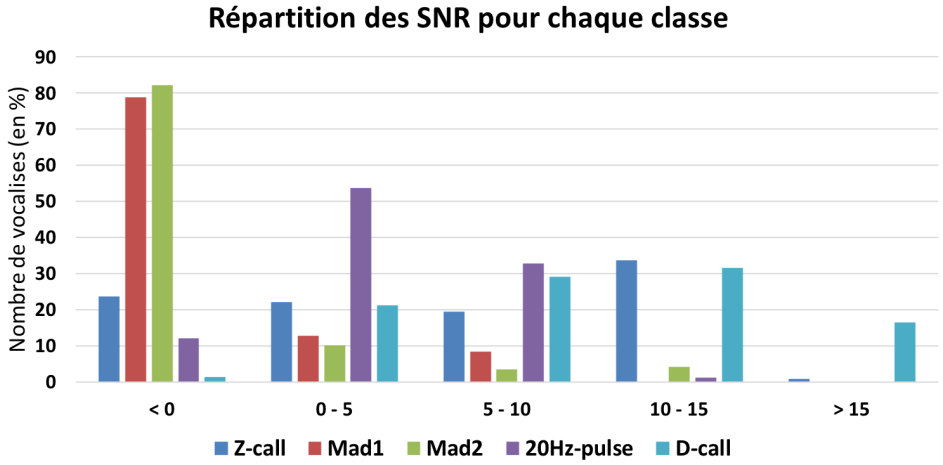


FIGURE 4.13 – Répartition des SNR des vocalises utilisées pour la base de test.

Le reste de la base de données correspond à des interférences et différents bruits de fond entre les vocalises. En terme de répartition, la quantité de vocalises représentent environ 11h soit un peu moins de 18 % des signaux traités. Ce constat permet de réaliser que les données contiennent principalement du bruit et des interférences. Néanmoins, il est important de souligner le fait que notre base de données contient beaucoup d'activité en termes de variété d'environnement, d'interférences et de nature de bruits. Aussi, l'évaluation de la fausse-alarme est à prendre avec précaution. En effet, nous mesurons un taux de fausses-alarmes par heure et cette métrique est liée au type de données et surtout à la quantité des données. Cela signifie que, dans notre cas, le taux de fausse alarme risque d'être un peu surévalué par rapport à une véritable mise en situation en plein mer où la plupart des données ne contiennent généralement pas autant d'activité. Néanmoins, ces résultats permettent de prendre conscience de la faisabilité d'une telle méthode et mettent en lumière ses points forts et ses limitations.

4.2.2 Courbes ROC

Les courbes ROC permettent d'apprécier les performances d'un détecteur avec sa sensibilité (taux de détection) en fonction de sa robustesse (taux de fausse-alarme). Dans notre cas, l'architecture de la méthode est telle qu'une bonne détection est nécessairement une bonne reconnaissance et une mauvaise détection (fausse alarme) est nécessairement une mauvaise reconnaissance. Ainsi, les performances de reconnaissance et de détection se lisent directement sur le même graphique. Afin d'interpréter au mieux les résultats, nous proposons tout d'abord d'afficher pour chaque classe les courbes ROC dans le cas où le détecteur ne passe que sur les fichiers qui contiennent les classes considérées. La courbe ROC de la classe Z-call est associée au résultat du détecteur multiclasses uniquement sur les fichiers contenant des Z-call et ainsi de suite pour chaque classe. Puis, nous proposons d'agrandir la base de données de test en faisant passer le détecteur multiclasses sur toutes les classes de vocalises. De cette façon, nous testons la robustesse de la méthode à une plus grande activité. Notamment, nous observons comme la méthode réagit dans le cas où la fenêtre d'observation utilisée ne « voit » aucune vocalises qui lui est associée. Ensuite, nous ajoutons du bruit océanique et enfin du bruit sismique. Dans cette manière de procéder, excepté pour les bruits océaniques qui contiennent assez peu d'activité, nous éprouvons progressivement notre méthode à travers des situations de plus en plus complexes.

Dans le cadre de notre partenariat industriel, nous cherchons à vérifier que nous sommes capables de détecter des vocalises qui sont émises suffisamment près du canon à air, c'est-à-dire avec un SNR assez élevé. En accord avec SERCEL, nous considérons que le SNR est satisfaisant lorsqu'il est supérieur à 5 décibels. Pour l'ensemble des résultats, les paramètres ont été choisis empiriquement afin de rester dans le cadre d'une démonstration de faisabilité. Le degré de parcimonie est K = 3 et le nombre de vocalises apprises N'_c pour chaque classe est de 30 cris pour les Z-call, 50 cris pour Mad1 et Mad2, 95 cris pour les 20Hz-pulse et 100 cris pour les D-call. Chaque courbe ROC est calculée en sélectionnant une plage de SNR.

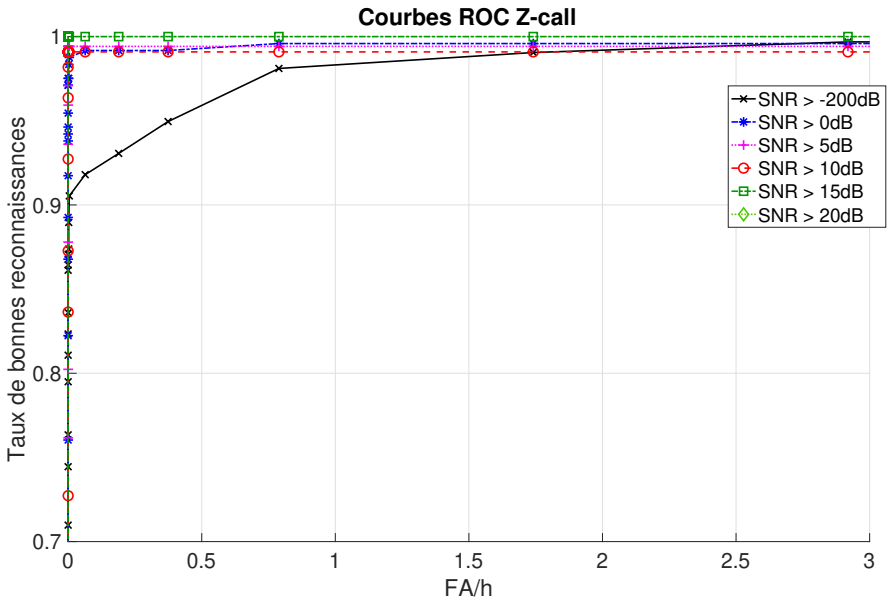


FIGURE 4.14 – Courbes ROC associées à la classe Z-call avec les performances par SNR où le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des Z-call.

La figure 4.14 permet d'apprécier les résultats de notre détecteur multiclasses dans le cas où les fichiers correspondent à l'environnement où les Z-call sont présents. La courbe noire (>-200dB) représente les résultats généraux car elle prend en compte toutes les vocalises. Dans cette partie, les SNR sont évalués de la même façon que pour SINR-SRC et il est intuitif d'observer que plus le SNR est élevé et meilleur sont les performances. Pour le cas des Z-call, à partir d'un SNR dépassant les 0dB, nous observons de bons résultats sur les fichiers contenant des Z-call.

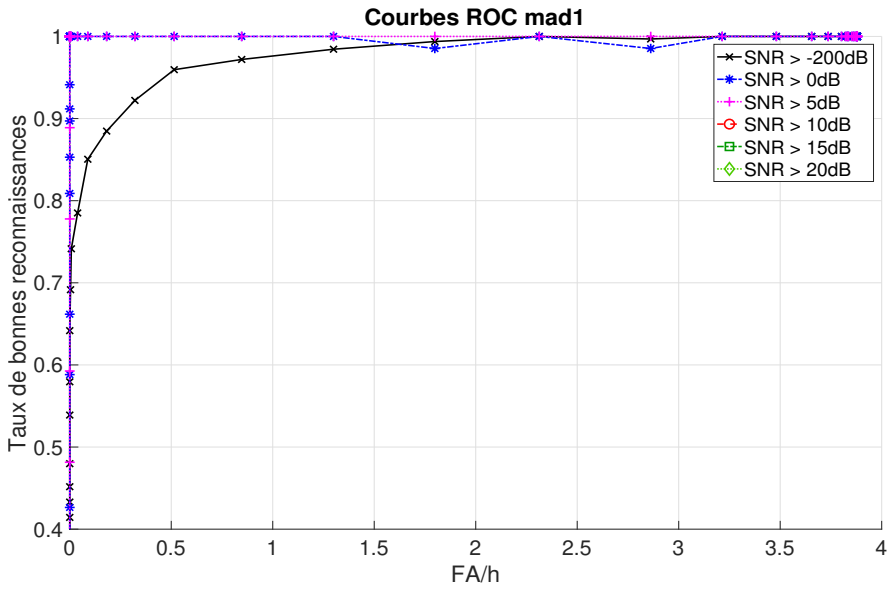


FIGURE 4.15 – Courbes ROC associées à la classe Mad1 avec les performances par SNR où le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des Mad1.

Pour les détections associée à la classe mad1 (figure 4.15), on remarque que la courbe, associée aux vocalises >0dB (courbe bleue), n'est pas parfaitement monotone. Ce phénomène est dû à notre post-traitement. D'une part, nous rejetons la détection si le « support continu » des valeurs du SINR est devenu trop grand. Il est alors possible que le nombre de détections diminue avec la diminution du seuil. Le rejet une détection si le support des valeurs du SINR est trop grand correspond au cas où le réglage du seuil n'a pas été effectué correctement. Il est alors probablement trop bas et par conséquent la détection n'a plus aucun sens. D'autre part, nous ne considérons qu'une détection par fenêtre d'observation. Cela signifie que le début de la détection (maximum local des valeurs du SINR) peut être décalé par rapport au vrai début. Le choix de ne garder qu'une unique vocalise est cohérent afin d'éviter de comptabiliser plusieurs fois un même événement. Comme nous pouvons le voir sur les résultats, pour un seuil bien choisi, les performances sont très bonnes, notamment pour les vocalises de la classe Mad1 dont le SNR dépasse 0 dB et correspond au contexte recherché.

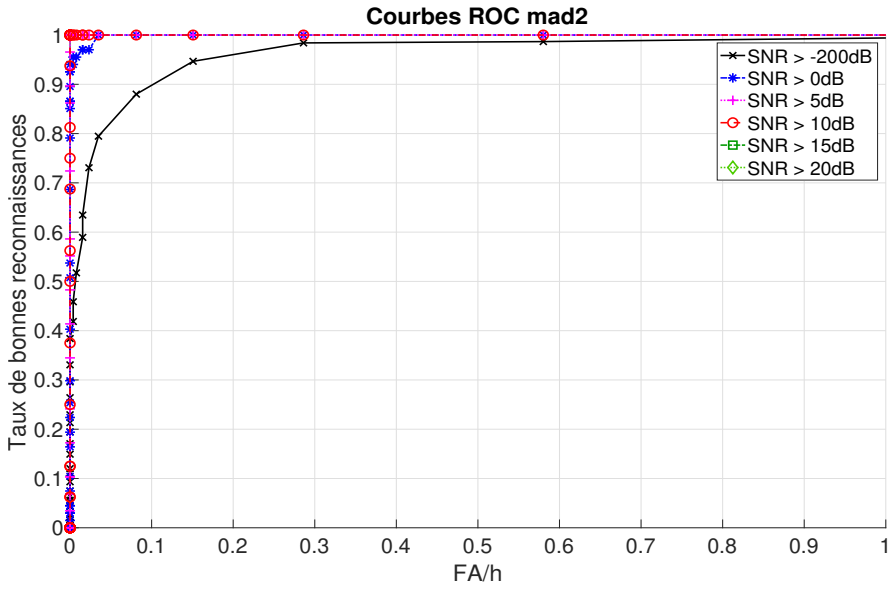


FIGURE 4.16 – Courbes ROC associées à la classe Mad2 avec les performances par SNR où le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des Mad2.

Les mêmes conclusions que pour la classe Mad1 restent valable pour la classe Mad2 (*cf.* figure 4.16) à savoir que les résultats sont bons pour des vocalises dont le SNR dépassent 0 dB. Pour le cas des vocalises avec un SNR plus faible, c'est surtout le fait de perdre la structure de la vocalises à identifier plus que la valeur du SNR qui conditionne la capacité de la méthode à bien reconstruire et donc à bien reconnaître la vocalise considérée.

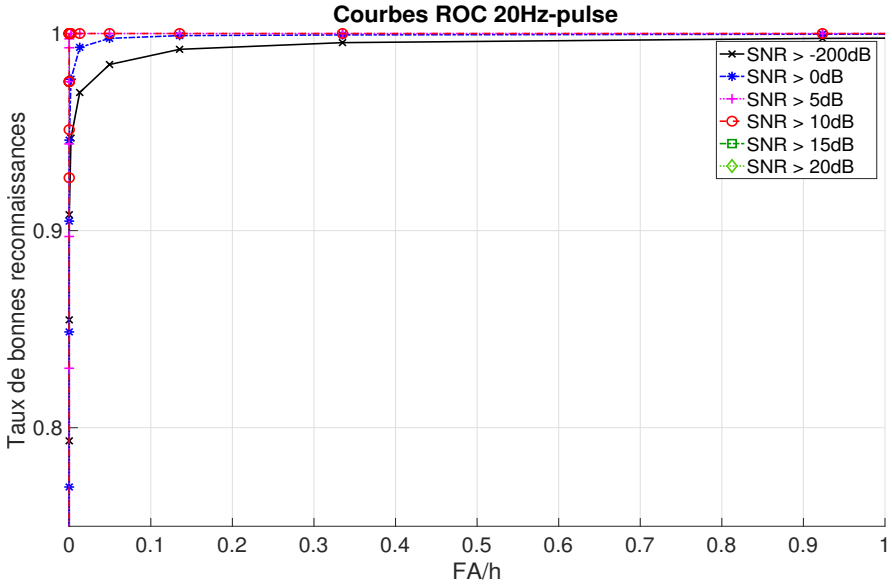


FIGURE 4.17 – Courbes ROC associées à la classe 20Hz-pulse avec les performances par SNR où le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des 20Hz-pulse.

Dans le cas des 20Hz-pulse (*cf.* figure 4.17), nous observons de très bonnes performances (sur des fichiers qui ne contiennent que des 20Hz-pulse) qui s'améliorent avec l'augmentation du SNR.

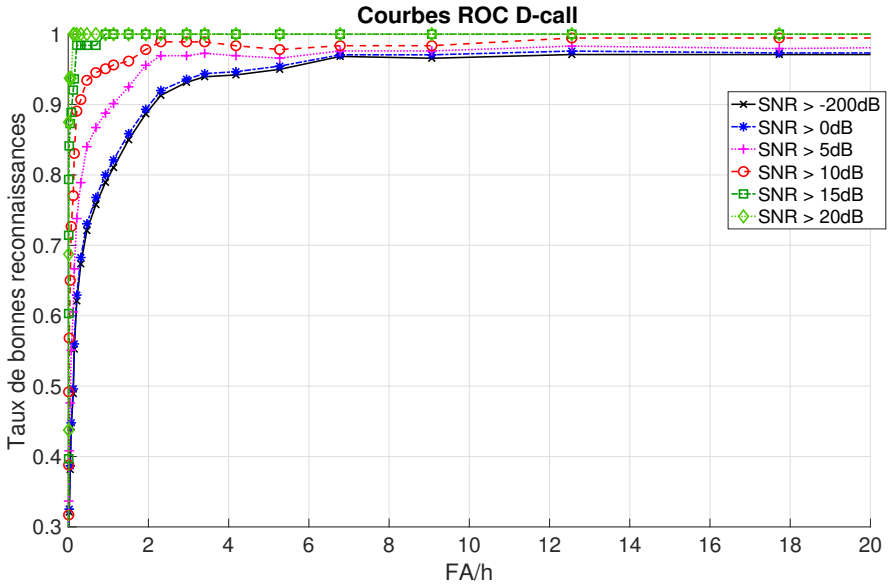


FIGURE 4.18 – Courbes ROC associées à la classe D-call avec les performances par SNR où le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des D-call.

Pour terminer cette section sur les performances dans le cas où l'environnement correspond aux vocalises recherchées, nous notons que les performances pour les D-call sont plus basses que celles obtenues pour les autres classes. Une explication possible de cette baisse de performance est, pour nous, que l'apprentissage du dictionnaire des D-call n'est pas suffisamment représentatif de la classe considérée.

4.2.3 Résumé des résultats

Afin d'avoir une description plus « pratique » de nos résultats, nous affichons ci-dessous, en complément des courbes ROC, des tableaux qui résument les performances de la détection (et donc de la reconnaissance) obtenues pour des taux de fausses-alarmes (FA) de l'ordre de la journée, de la demi-journée et de l'heure pour chaque plage de SNR en plus de donner la répartition des vocalises.

FA / **SNR**	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	91.3292	91.9950	98.3172	317
> 0dB	99.0200	99.1736	99.5868	242
> 5dB	99.4186	99.4186	99.4186	172
> 10dB	99.0909	99.0909	99.0909	110
> 15dB	100.0000	100.0000	100.0000	3

TABLEAU 4.1 – Performances de détection et bonne reconnaissance des Z-call quand le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des Z-call.

FA / **SNR**	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	78.8933	84.3146	97.6175	321
> 0dB	100.0000	100.0000	100.0000	68
> 5dB	100.0000	100.0000	100.0000	27

TABLEAU 4.2 – Performances de détection et bonne reconnaissance des Mad1 quand le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des Mad1.

FA / **SNR**	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	80.7272	88.2020	99.4026	375
> 0dB	100.0000	100.0000	100.0000	67
> 5dB	100.0000	100.0000	100.0000	29
> 10dB	100.0000	100.0000	100.0000	16

TABLEAU 4.3 – Performances de détection et bonne reconnaissance des Mad2 quand le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des Mad2.

FA / **SNR**	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	97.0197	97.5816	99.4546	3242
> 0dB	99.3013	99.4821	99.9212	2847
> 5dB	100.0000	100.0000	100.0000	1107
> 10dB	100.0000	100.0000	100.0000	41

TABLEAU 4.4 – Performances de détection et bonne reconnaissance des 20Hz-pulse quand le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des 20Hz-pulse.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	36.0968	44.9657	79.6912	380
> 0dB	36.5781	45.5652	80.7538	375
> 5dB	45.3025	55.9696	89.2562	294
> 10dB	62.2361	73.3745	95.2750	183
> 15dB	86.2205	89.3950	100.0000	63
> 20dB	93.7500	94.7463	100.0000	16

TABLEAU 4.5 – Performances de détection et bonne reconnaissance des D-call quand le détecteur multiclasses est uniquement appliqué sur des fichiers contenant des D-call.

4.2.4 Discussion

Dans ce cadre où les fichiers sont associés aux classes de vocalises considérées, nous constatons que la prise en charge de la détection dans la mise en œuvre de notre détecteur multiclasses permet de conserver des résultats de détection et bonne reconnaissance satisfaisants par rapport à notre contexte. Cela signifie qu'il est possible pour un utilisateur, qui a des *a priori* sur ces fichiers en termes de présence de vocalises, d'appliquer le détecteur multiclasses afin de détecter, identifier et annoter les données. Néanmoins, ces résultats comportent un *a priori* fort puisque les données de test choisies, même si elles contiennent des interférences et divers bruits de fonds, ne semblent pas mettre la méthode en difficulté. Ainsi, afin d'éprouver notre méthode, nous testons à présent notre détecteur multiclasses dans les trois situations suivantes : la base de test contient tous les types de vocalises, la base de test contient en plus des bruits océaniques et enfin la base de test contient également des bruits sismiques.

4.2.5 La méthode est testée sur toutes les classes de vocalises

Voici ci-dessous les résultats où toutes les vocalises sont représentées pour toutes les classes et donc à travers toutes les fenêtres d'observation.

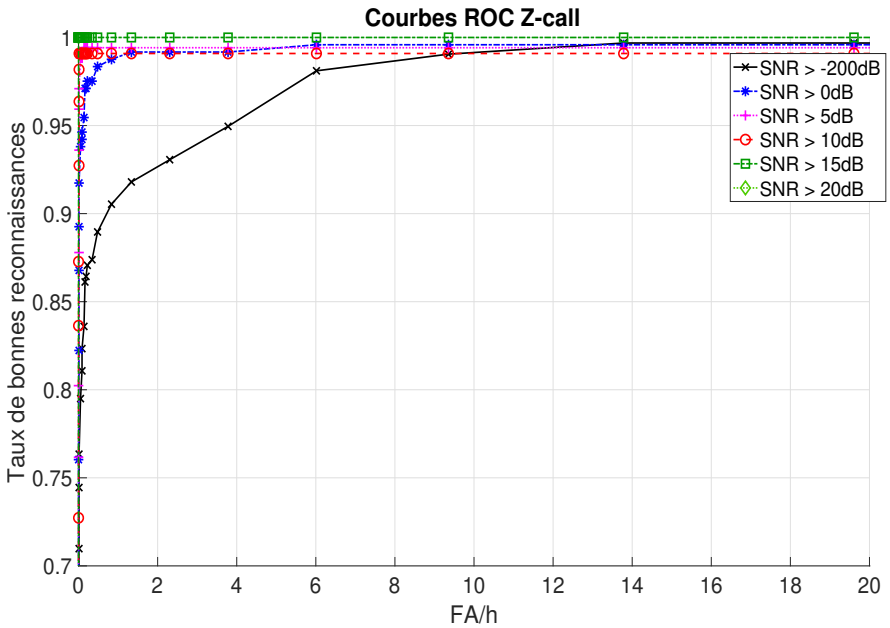


FIGURE 4.19 – Courbes ROC associées à la classe Z-call avec les performances par SNR en présence de toutes les classes de vocalises.

Le fait d'avoir augmenté le nombre de fichiers contenant des signaux structurés fait, dans notre

cas, baisser les performances globales de détections des Z-call (*cf.* figure 4.19). Néanmoins, notre détecteur multiclasses reste performant pour des vocalises dont le SNR est supérieur à 5dB, ce qui permet de satisfaire, dans le cadre de cette base de données, les contraintes de notre contexte. Du point de vue de la détection, il faut comprendre que plus le SNR est bas et plus la « forme » de la vocalise considérée est altérée. De ce constat vient le fait que détecter cette vocalise altérée va également entraîner la détection d'autres « structures » qui n'appartiennent pas à la classe de signaux considérée.

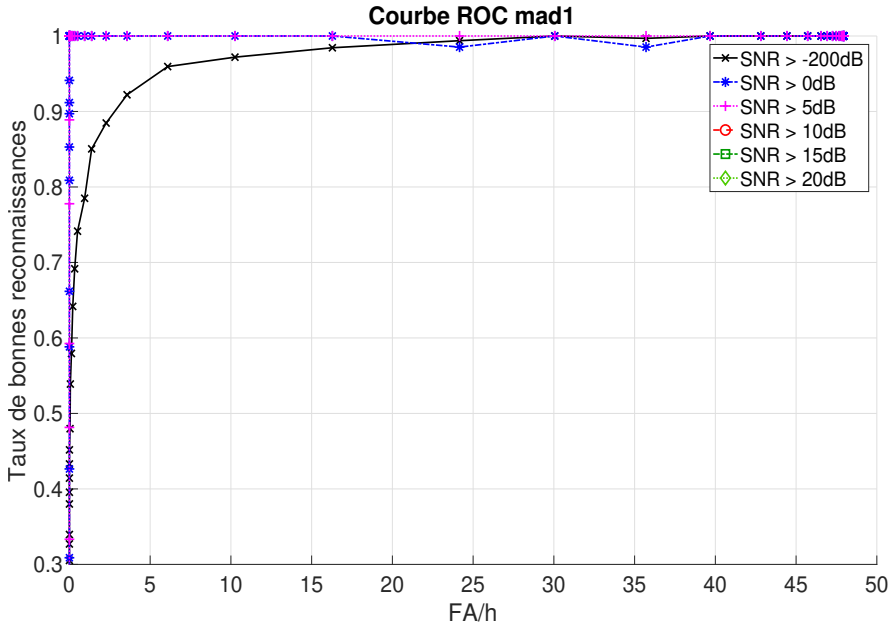


FIGURE 4.20 – Courbes ROC associées à la classe Mad1 avec les performances par SNR en présence de toutes les classes de vocalises.

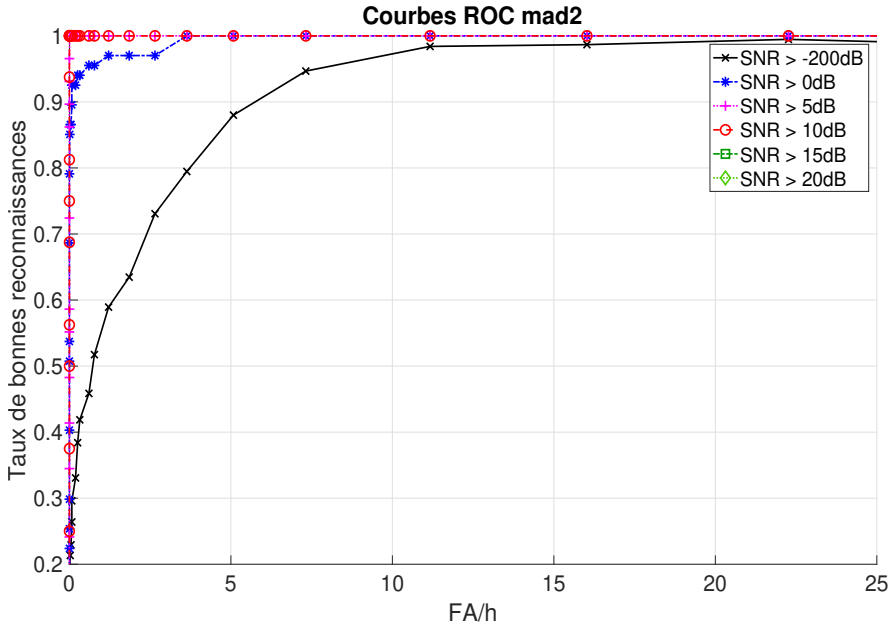


FIGURE 4.21 – Courbes ROC associées à la classe Mad2 avec les performances par SNR en présence de toutes les classes de vocalises.

Les classes de signaux Mad1 et Mad2 ont également des performances globales qui diminuent (*cf.* figures 4.20 et 4.21) avec l'ajout de ces nouvelles données. A nouveau, pour les vocalises dont le SNR est suffisamment élevée (ici supérieur à 0dB), notre méthode a des résultats satisfaisants.

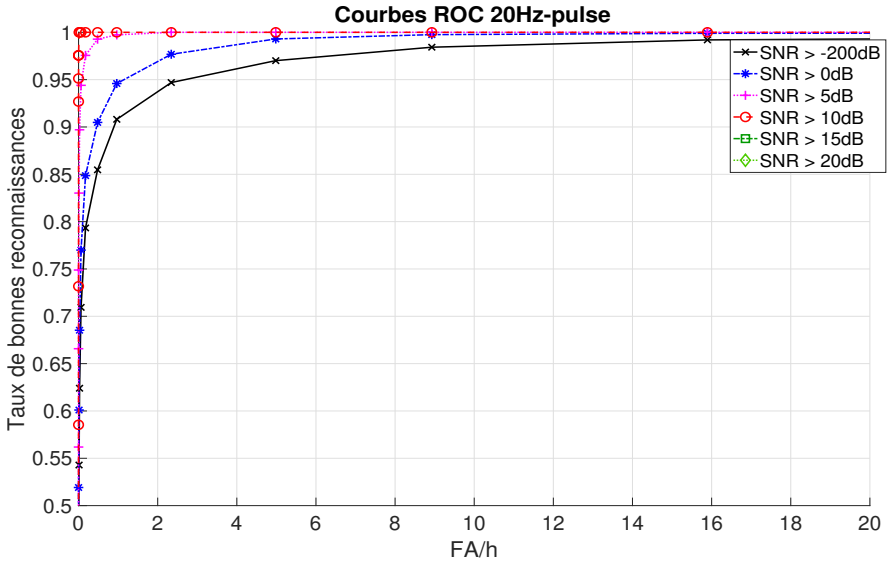


FIGURE 4.22 – Courbes ROC associées à la classe 20Hz-pulse avec les performances par SNR en présence de toutes les classes de vocalises.

Les performances sur la classe des 20Hz-pulse se trouvent également diminuées avec l'ajout des nouvelles données (*cf.* figure 4.22). Comme nous l'avons noté lors de la mise en œuvre de notre méthode, la fenêtre d'observation des 20Hz-pulse est très spécifique et encourage fortement l'algorithme SRC à identifier des 20Hz-pulse tout au long des valeurs du SINR.

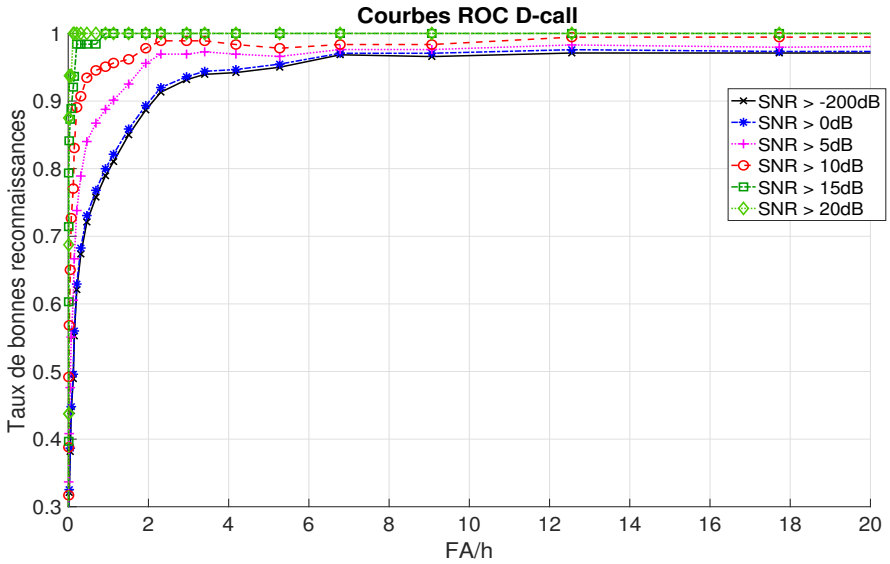


FIGURE 4.23 – Courbes ROC associées à la classe D-call avec les performances par SNR en présence de toutes les classes de vocalises.

Il est intéressant de noter que les performances sur les D-call ne changent pas. Le fait que les performances n'évoluent pas est dû au fait qu'il n'y a pas d'autres « structures », à travers cette fenêtre, qui ressemblent le plus à des D-call.

4.2.6 Résumé des résultats

De la même façon que précédemment, nous présentons les résultats complémentaires dans les tableaux ci-dessous.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	78.7791	80.8822	90.9547	317
> 0dB	93.3326	94.1650	98.8974	242
> 5dB	98.4412	98.8372	99.4186	172
> 10dB	99.0909	99.0909	99.0909	110
> 15dB	100.0000	100.0000	100.0000	3

TABLEAU 4.6 – Performances de détection et bonne reconnaissance des Z-call quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	48.5534	55.3753	79.3703	321
> 0dB	100.0000	100.0000	100.0000	68
> 5dB	100.0000	100.0000	100.0000	27

TABLEAU 4.7 – Performances de détection et bonne reconnaissance des Mad1 quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	22.3885	29.8427	55.4367	375
> 0dB	86.5672	92.5373	96.2901	67
> 5dB	100.0000	100.0000	100.0000	29
> 10dB	100.0000	100.0000	100.0000	16

TABLEAU 4.8 – Performances de détection et bonne reconnaissance des Mad2 quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	66.1561	72.4710	90.9034	3242
> 0dB	72.2499	78.4254	94.6665	2847
> 5dB	91.7671	94.9748	99.7356	1107
> 10dB	100.0000	100.0000	100.0000	41

TABLEAU 4.9 – Performances de détection et bonne reconnaissance des 20Hz-pulse quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	36.0968	44.9657	79.0409	380
> 0dB	36.5781	45.5652	80.0947	375
> 5dB	45.3025	55.9696	88.8359	294
> 10dB	62.2361	73.3745	95.1062	183
> 15dB	86.2205	89.3950	100.0000	63
> 20dB	93.7500	94.7463	100.0000	16

TABLEAU 4.10 – Performances de détection et bonne reconnaissance des D-call quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises.

4.2.7 Discussion

Nous avons testé notre détecteur multiclasses sur l'ensemble des classes de vocalise. Le constat est que, même si globalement les performances diminuent, notre méthode conserve des résultats satisfaisants dans le cas où les vocalises sont supérieures à 0 dB, excepté pour la classe des D-call où les performances ne changent pas et sont à améliorer.

4.2.8 Courbes ROC en présence de bruit océanique et sismique

Dans cette section nous présentons notre détecteur multiclasses dans la situation où du bruit océanique est ajouté à la base de test. Puis nous ajoutons ensuite du bruit sismique. Le fait d'ajouter à la fois du bruit océanique et sismique nous place dans un cas défavorable mais nous permet d'être cohérent avec notre contexte applicatif. Les résultats sont également présentés par des courbes ROC.

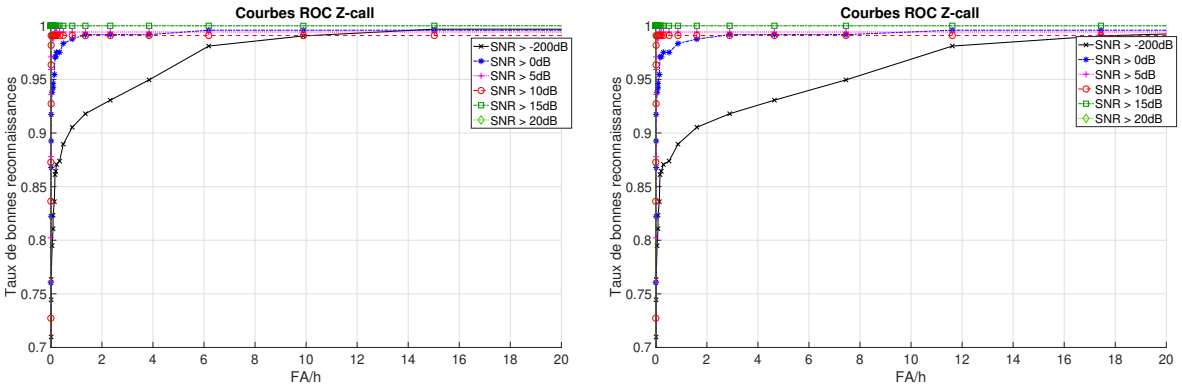


FIGURE 4.24 – Courbes ROC associées à la classe Z-call avec les performances par SNR en présence des bruits océanique (figure de gauche) puis océanique et sismique (figure de droite).

L'ajout de bruit océanique n'a pas grande conséquence sur les performances globales pour la classe des Z-call, ce qui signifie que la fenêtre d'observation associée à cette classe permet d'avoir des résultats satisfaisants par rapport à cette situation. De même, après l'ajout de bruit sismique, le détecteur multiclasses restent robuste et a des résultats satisfaisants par rapport à notre contexte industriel. Ces résultats sont cohérents du point de vue de la reconnaissance de formes, au sens où le bruit provenant de campagnes sismiques contient surtout des bruits impulsifs qui ont une structure bien différentes des Z-call.

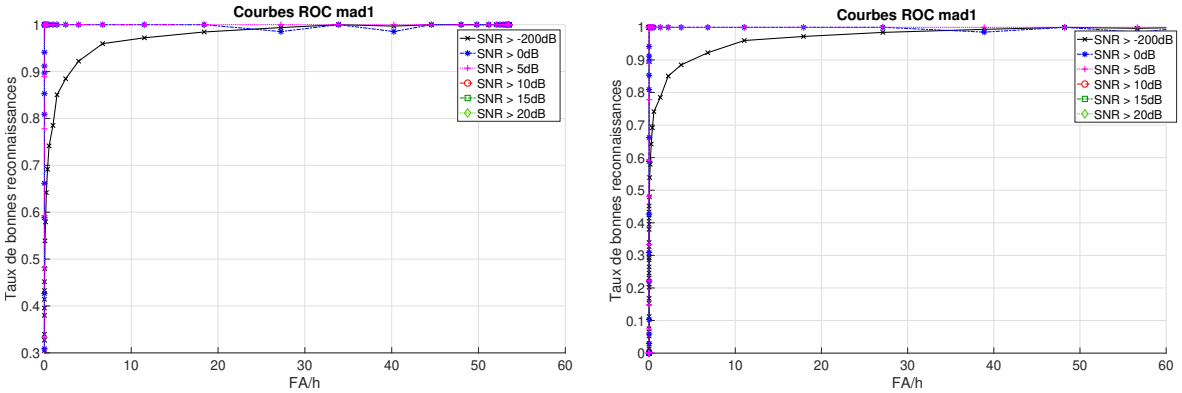


FIGURE 4.25 – Courbes ROC associées à la classe Mad1 avec les performances par SNR en présence des bruits océanique (figure de gauche) puis océanique et sismique (figure de droite).

Pour les mêmes raisons que pour les Z-call, les structures des classes de vocalises Mad1 et Mad2 sont suffisamment différentes pour qu'à partir d'un SNR suffisamment élevé, ici supérieur à

0dB, l'ajout de bruit n'a pas d'influence et justifie de la robustesse du détecteur multiclasses pour ces classes de signaux (*cf.* figures 4.25 et 4.26).

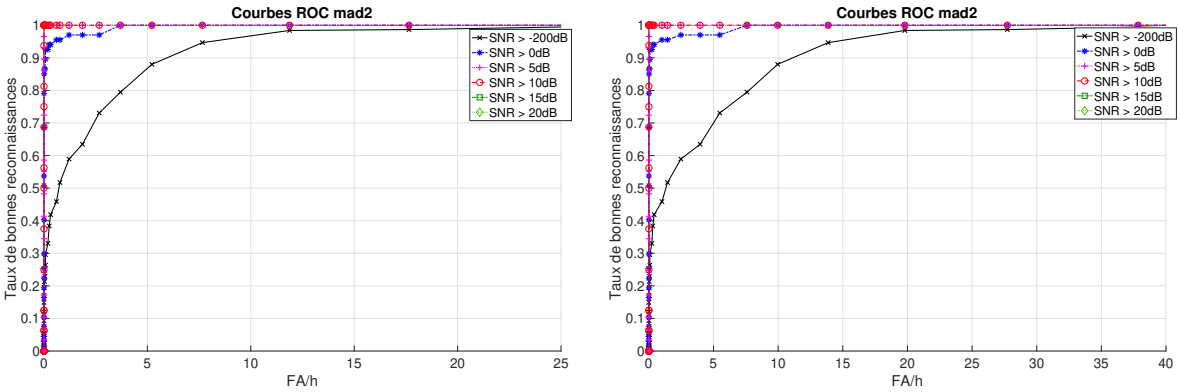


FIGURE 4.26 – Courbes ROC associées à la classe Mad2 avec les performances par SNR en présence des bruits océanique (figure de gauche) puis océanique et sismique (figure de droite).

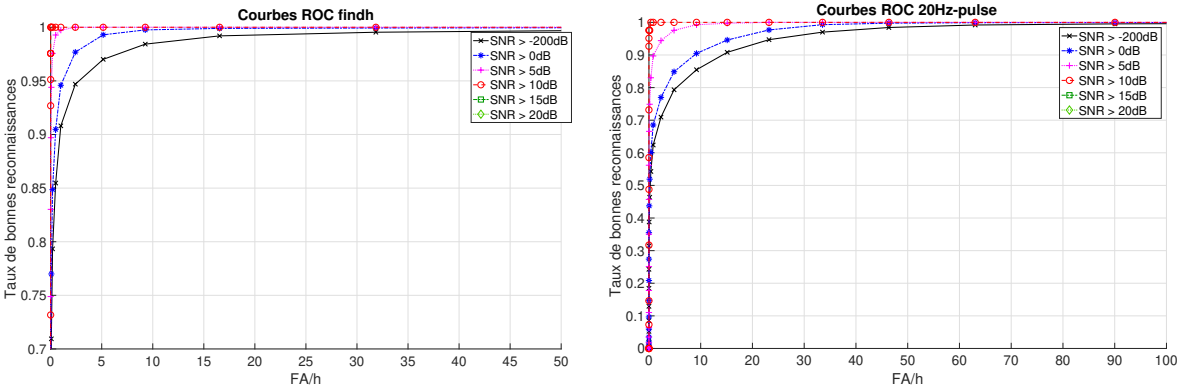


FIGURE 4.27 – Courbes ROC associées à la classe 20Hz-pulse avec les performances par SNR en présence des bruits océanique (figure de gauche) puis océanique et sismique (figure de droite).

Dans le cas des 20Hz-pulse, les performances se trouvent fortement dégradées à partir de l'ajout du bruit provenant de campagne sismique. En effet, dans le cas du bruit océanique, les performances sont satisfaisantes jusqu'à des vocalises supérieures à 5dB (plus de 91% de bonnes reconnaissances pour 1 fausse-alarme par jour *cf.* tableau 4.14), alors que pour le bruit sismique, les performances ne sont intéressantes qu'à partir de vocalises ayant un SNR supérieures à 10dB. Ce comportement s'explique simplement par le caractère impulsif des tirs sismiques présents dans les données qui, à travers la fenêtre d'observation des 20Hz-pulse, ont des structures très similaires aux signaux de la classe considérée.

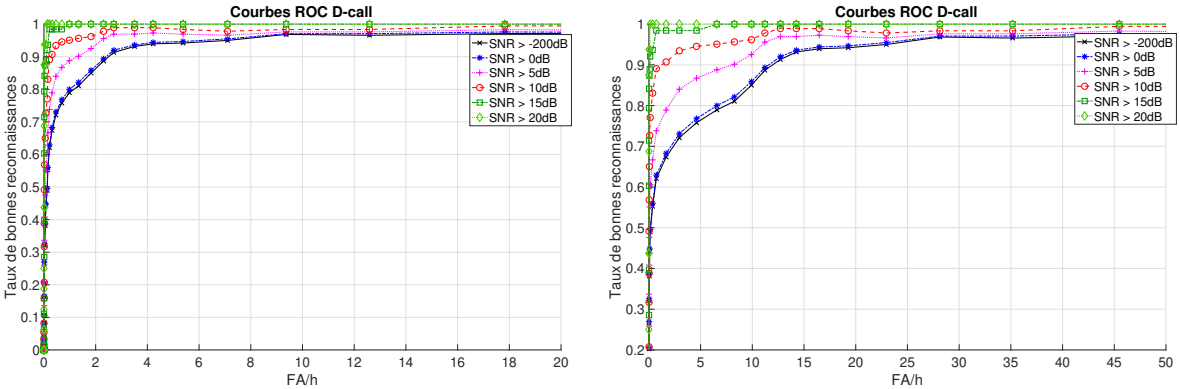


FIGURE 4.28 – Courbes ROC associées à la classe D-call avec les performances par SNR en présence des bruits océanique (figure de gauche) puis océanique et sismique (figure de droite).

Les performances obtenues sur les D-call diminuent mais les effets ne sont pas très marqués. La conséquence de ces résultats par rapport à cette classe de signaux indique que le dictionnaire est réellement robuste aux classes inconnues mais qu'il n'est pas suffisamment représentatif de la classe considérée. Il est alors nécessaire d'agrandir le dictionnaire et/ou de sélectionner de meilleurs représentants de la classe D-call afin d'améliorer les résultats, comme ce qui est présenté dans [SOCHELEAU et SAMARAN, 2017].

4.2.9 Résumé des résultats

De la même façon que précédemment, nous mettons à la suite les résultats complémentaires, tout d'abord lorsque le bruit océanique est ajouté puis lorsque l'ensemble de la base de données est testée. Nous retrouvons les résultats présentés sur les courbes ROC vus plus haut et cela nous permet de justifier la faisabilité de notre détecteur multiclasses.

Résultats avec ajout de bruit océanique

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	78.7791	80.8822	90.9311	317
> 0dB	93.3326	94.1650	98.8896	242
> 5dB	98.4412	98.8372	99.4186	172
> 10dB	99.0909	99.0909	99.0909	110
> 15dB	100.0000	100.0000	100.0000	3

TABLEAU 4.11 – Performances de détection et bonne reconnaissance des Z-call quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises ainsi que sur les fichiers contenant du bruit océanique.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	48.4185	55.0265	78.7173	321
> 0dB	100.0000	100.0000	100.0000	68
> 5dB	100.0000	100.0000	100.0000	27

TABLEAU 4.12 – Performances de détection et bonne reconnaissance des Mad1 quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises ainsi que sur les fichiers contenant du bruit océanique.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	22.3885	29.8427	55.4367	375
> 0dB	86.5672	92.5373	96.2901	67
> 5dB	100.0000	100.0000	100.0000	29
> 10dB	100.0000	100.0000	100.0000	16

TABLEAU 4.13 – Performances de détection et bonne reconnaissance des Mad2 quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises ainsi que sur les fichiers contenant du bruit océanique.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	65.6571	71.9885	90.8214	3242
> 0dB	71.7555	77.9730	94.6013	2847
> 5dB	91.4927	94.7929	99.7299	1107
> 10dB	100.0000	100.0000	100.0000	41

TABLEAU 4.14 – Performances de détection et bonne reconnaissance des 20H-pulse quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises ainsi que sur les fichiers contenant du bruit océanique.

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	36.0968	44.9657	79.0409	380
> 0dB	36.5781	45.5652	80.0947	375
> 5dB	45.3025	55.9696	88.8359	294
> 10dB	62.2361	73.3745	95.1062	183
> 15dB	86.2205	89.3950	100.0000	63
> 20dB	93.7500	94.7463	100.0000	16

TABLEAU 4.15 – Performances de détection et bonne reconnaissance des D-call quand le détecteur multiclasses est appliqué sur tous les fichiers contenant des vocalises ainsi que sur les fichiers contenant du bruit océanique.

Résultats sur toute la base de données

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	78.7791	80.8822	89.2371	317
> 0dB	93.3326	94.1650	98.4200	242
> 5dB	98.4412	98.8372	99.4186	172
> 10dB	99.0909	99.0909	99.0909	110
> 15dB	100.0000	100.0000	100.0000	3

TABLEAU 4.16 – Résultats associés à la classe Z-call sur toute la base de données

FA / SNR	1 FA PAR JOUR	1 FA PAR DEMI-JOURNÉE	1 FA PAR HEURE	NOMBRE DE VOCALISES
> -200dB	48.4185	55.0265	76.7220	321
> 0dB	100.0000	100.0000	100.0000	68
> 5dB	100.0000	100.0000	100.0000	27

TABLEAU 4.17 – Résultats associés à la classe Mad1 sur toute la base de données

FA / SNR	1 FA par jour	1 FA par demi-journée	1 FA par heure	Nombre de vocalises
> -200dB	22.3885	29.7772	45.8035	375
> 0dB	86.5672	92.5373	95.4988	67
> 5dB	100.0000	100.0000	100.0000	29
> 10dB	100.0000	100.0000	100.0000	16

TABLEAU 4.18 – Résultats associés à la classe Mad2 sur toute la base de données

FA / SNR	1 FA par jour	1 FA par demi-journée	1 FA par heure	Nombre de vocalises
> -200dB	37.4442	41.4084	63.4055	3242
> 0dB	42.2087	46.5516	69.5247	2847
> 5dB	64.6532	69.4413	90.2548	1107
> 10dB	97.5610	97.5610	100.0000	41

TABLEAU 4.19 – Résultats associés à la classe 20Hz-pulse sur toute la base de données

FA / SNR	1 FA par jour	1 FA par demi-journée	1 FA par heure	Nombre de vocalises
> -200dB	36.0968	44.9657	63.6596	380
> 0dB	36.5781	45.5652	64.5084	375
> 5dB	45.3025	55.9696	75.3163	294
> 10dB	62.2361	73.3745	89.5552	183
> 15dB	86.2205	89.3950	98.4127	63
> 20dB	93.7500	94.7463	100.0000	16

TABLEAU 4.20 – Résultats associés à la classe D-call sur toute la base de données

4.3 Conclusion et perspectives

Nous avons mis en œuvre un détecteur multiclasses capable de détecter, d'identifier, d'annoter et de donner un coefficient de confiance et/ou ainsi qu'une estimation du SNR à la structure qui est reconnue. Le détecteur multiclasses conserve les bonnes propriétés de SINR-SRC. Notamment, il est tout à fait possible d'envisager de faire de l'apprentissage incrémental semi-supervisé, ce qui reviendrait à créer des dictionnaires « en ligne » et/ou des dictionnaires « jetables ». L'idée est d'être le plus indépendant possible des données qui ne sont pas des vocalises et d'être le plus représentatif possible des vocalises. La question de la définition de la classe est alors pertinente et rejoint cette idée de la représentativité de la classe considérée à travers l'apprentissage de dictionnaire. Dans [SOCHELEAU et SAMARAN, 2017], un apprentissage basé sur un modèle de chirp pour la détection et la reconnaissance de D-call permet d'avoir des résultats légèrement supérieurs à un apprentissage directement effectué sur les données brutes. En effet, les données brutes contiennent nécessairement le signal recherché ainsi que l'environnement dans lequel il est produit. Il convient alors de réfléchir à un modèle de données pour l'apprentissage qui soit véritablement « dé-corrélé » de l'environnement (le bruit de fond) tout en conservant la nature intrinsèque des signaux considérés. Une perspective est de regarder des modèles de débruitage basé par exemple sur les ondelettes et/ou des modèle d'estimation de données qui permettrait de réduire au maximum l'influence de l'environnement sur les données en cherchant à réduire le bruit tout

en conservant la représentativité de la classe considérée. Pour finir, la notion de représentativité est également difficile à définir mathématiquement et mériterait que nous cherchions des critères permettent de prédire et/ou de quantifier la représentativité d'un dictionnaire par rapport à la variabilité de la classe considérée ainsi que par rapport à la taille du dictionnaire à apprendre.

Conclusion générale

« *Dans ce vaste océan, quelle route emprunter ?*
Suivons la rose des vents, laissons-nous guider ! »
— L'auteur

Synthèse des contributions

Tout au long de ce manuscrit, nous nous sommes efforcés de répondre à notre problématique de mise en œuvre d'un algorithme capable de réaliser automatiquement la reconnaissance de vocalises de mammifères marins, avec une contrainte industrielle forte : la gestion des fausses alarmes.

Au travers du **premier chapitre**, nous avons mis en relation la *reconnaissance de formes* et la *bioacoustique*. Puis nous avons présenté les sources sonores du milieu océanique, les mammifères marins et leurs vocalises avec, en parallèle, l'*environnement sismique* de SERCEL. Ces informations nous ont permis de bien définir notre cadre d'étude avec notamment une discussion autour de la démarche scientifique à adopter.

Le **deuxième chapitre** commence par généraliser ce qu'est la *classification* et sa finalité. Ce cheminement nous donne l'occasion d'élargir notre problématique en proposant, non plus uniquement de mettre en œuvre un algorithme de reconnaissance automatique, mais de réfléchir à remplacer ou seconder un expert en reconnaissance de vocalises de mammifères marins. Nous avons alors discuté des solutions proposées dans la littérature bioacoustique par rapport à notre contexte, ce qui nous a amenés à séparer la reconnaissance en deux approches, la première basée sur de la *réduction de dimensions* et la seconde basée sur une *mesure de similarité*. Après avoir explicité les avantages et les inconvénients de ces deux approches à travers un état de l'art de la reconnaissance en bioacoustique, nous avons décidé que la seconde approche correspondait mieux au besoin de notre contexte. Une raison majeure de ce choix est le besoin de prendre en compte les fausses alarmes et plus généralement d'être capable de prendre une décision face aux classes inconnues. La seconde approche est basée sur une mesure qui, par définition, va permettre d'estimer si la donnée observée par le système est plutôt « proche » (classe connue) ou plutôt « loin » (classe inconnue) des classes apprises. Cette propriété permet de « couvrir » tous les cas possibles de notre espace de données d'entrée et notamment, il permet de gérer une grande partie de l'infinie diversité des signaux inconnus. Au contraire, la première approche « oblige » le système à prendre une décision par rapport à un nombre fini de classes apprises. Dans ce cas, la gestion des fausses alarmes ne peut se faire qu'en apprenant aux systèmes l'ensemble infini de classes inconnues, ce qui nous semble *a priori* impossible.

Enfin, en fin de chapitre, nous avons décrit les problématiques liées à la *qualité* d'un travail scientifique et industriel. Nous avons présenté un processus idéal de validation d'une méthode de reconnaissance du point de vue industriel (la notion de qualité) et scientifique (la qualité et le besoin de références). Nos conclusions sont que l'utilisation de différentes bases de données, de différentes mesures de performances et de différentes méthodes par la communauté bioacous-

tique, pas toujours reproductibles, rendent difficile de se positionner par rapport au travail de nos prédécesseurs et n'est généralement pas productif.

Les principales contributions de notre travail de thèse sont développées dans les deux chapitres suivants.

Malgré le constat du chapitre précédent quant à la validation des méthode de reconnaissance, le **troisième chapitre** présente la mise en œuvre et l'évaluation de notre méthode de reconnaissance dénommée SINR-SRC. Les données d'entrées de notre méthode sont considérées comme étant la sortie d'un détecteur

A partir de la notion de mesure de similarité, nous avons introduit l'utilisation des représentations parcimonieuses et l'apprentissage de dictionnaire pour représenter les données et faire de la reconnaissance. Un point fort de ce cheminement est qu'au contraire de la méthode des K-plus-proches-voisins, l'espace de représentation des données (ensemble des combinaisons linéaires des atomes du dictionnaire) permet de « couvrir » un espace plus grand que l'ensemble original de données d'apprentissage.

Après la description méthodologique, nous avons défini un processus de validation de SINR-SRC en commençant par décrire la base de données réelles utilisée et les pré-traitements appliqués sur ces données. Cette base de données a été construite en cohérence avec notre contexte de gestion des fausses alarmes et contient, en plus des vocalises à reconnaître, des données représentatives de bruits océanique et sismique.

Tout d'abord, nous avons proposé de prendre en compte la gestion de la complexité de calcul. Ainsi, nous avons présenté une *option de compression* basée sur l'apprentissage de dictionnaire. Après avoir fixé une nouvelle taille de dictionnaire plus faible que précédemment, cette option construit le nouveau dictionnaire en sélectionnant automatiquement les données d'apprentissage les plus pertinentes. Les tests expérimentaux de cette option montrent que les résultats restent satisfaisants même en divisant par deux la taille du dictionnaire de départ.

Ensuite, afin de prendre en compte la gestion des fausses alarmes, nous avons utilisé le fait que notre méthode est basée sur une mesure de similarité. En conséquence, nous avons été à même de mettre en place une *option de rejet*, le seuillage du SINR, qui permet d'estimer si l'observation, après décomposition à travers le dictionnaire, appartient à une classe connue ou inconnue. De façon plus générale, nous avons réalisé une démonstration de faisabilité de la mise en place possible d'un niveau de confiance qui renseigne à quel taux la méthode connaît, ou non, l'observation considérée. Ce niveau de confiance permet d'envisager de faire de l'auto-apprentissage incrémental semi-supervisé. Par exemple, un utilisateur peut avoir envie d'ajouter une nouvelle classe, c'est-à-dire un nouveau dictionnaire. Le niveau de confiance lui permet alors de mesurer si « l'expérience » donnée au système (les données d'apprentissage) est suffisante par rapport aux performances exigées.

Puis, nous avons mis à l'épreuve notre méthode en testant ces deux options dans le cas de notre base de données. Nous avons ainsi montré que SINR-SRC est une méthode qui donne de très bons résultats (92,1 % de bonnes reconnaissances des vocalises et 97,3 % de bons rejets de bruits sur toute la base de tests) en plus d'être robuste par rapport à ses deux paramètres principaux (la taille du dictionnaire par classe et le degré de parcimonie). Du fait de sa faible quantité de paramètres, elle est facile d'utilisation et notamment elle est modulaire (facilité d'ajouter et d'enlever une classe).

De notre point de vue, un point dur de la reconnaissance est de pouvoir prendre en compte un maximum d'informations discriminantes au service de la reconnaissance. C'est d'ailleurs ce que cherchent à faire les méthodes de réduction de dimensions qui proposent de trouver des sous-espaces avec pour objectif de séparer au mieux les données. Dans notre méthode, plutôt que de réduire l'espace des données d'entrée de façon abstraite, la méthode donne la possibilité d'appliquer toutes les connaissances *a priori* soit directement sur les données d'apprentissage soit sur le dictionnaire. Par exemple, nous avons introduit les propriétés temps-fréquence des vocalises

en filtrant les données d'apprentissage et nous avons également montré, sous forme de preuve de concept, qu'il était possible de prendre en compte la dispersion modale.

Le **dernier chapitre** représente l'évolution de SINR-SRC en prenant en compte, non pas que la reconnaissance, mais également la détection des signaux d'intérêt. Les données d'entrée ne sont plus traitées comme une sortie d'un détecteur, mais au contraire, le signal brut est balayé et traité. De cette façon, nous considérons toute la chaîne de traitement des données brutes jusqu'à la détection et la reconnaissance des signaux d'intérêt. Nous avons ainsi présenté la mise en œuvre de notre méthode, dénommée le *détecteur multiclasses*, et nous avons testé ses performances à travers une base de données similaire à la base de données utilisée pour SINR-SRC. Cette nouvelle architecture permet, au contraire de SINR-SRC, de traiter plusieurs classes simultanément. Comme un expert en reconnaissance de vocalises, elle est capable de détecter et de reconnaître une vocalise d'intérêt. La détection ainsi obtenue permet d'annoter automatiquement les vocalises d'intérêt en donnant un temps de début et un temps de fin de détection, une plage de fréquence, une estimation du SNR et un coefficient de confiance. Les résultats obtenus nous indiquent des pistes d'amélioration. L'apprentissage de certaines classes, comme les D-call, doit être plus représentatif. La reconnaissance des 20Hz-pulse doit être rendue plus robuste aux environnements sismiques. Néanmoins, cette méthode est prometteuse.

Perspectives

Les avancées réalisées durant ces trois années de doctorat nous permettent d'envisager les perspectives suivantes.

L'apprentissage de dictionnaire

L'apprentissage du dictionnaire est un point clé de notre méthode de reconnaissance. Un mauvais dictionnaire ne permettra jamais d'avoir de bons résultats. De cette observation découle de nombreuses perspectives d'amélioration en relation avec l'apprentissage de dictionnaire :

— Le dictionnaire d'une classe considérée a pour objectif d'être le plus représentatif possible. Dans notre détecteur multiclasses, la représentativité d'un dictionnaire par rapport à une classe est traduite par le niveau de confiance. Une perspective serait de commencer par sélectionner les données à prendre en compte (ou à enlever) par le dictionnaire à l'aide du niveau de confiance. De cette façon, le dictionnaire serait appris automatiquement, de façon incrémentale, soit à partir d'une base d'apprentissage fixée, soit en ligne à l'aide d'un opérateur. Par exemple, pour une base d'apprentissage fixée, le dictionnaire est agrandi tant que le niveau de confiance obtenu sur toutes les vocalises de la base d'apprentissage ne dépasse pas 90% de confiance.

— Malgré la construction de notre niveau de confiance comme outil de mesure de représentativité, la notion de représentativité n'est pas bien définie et une perspective fondamentale de la reconnaissance serait d'approfondir cette notion en cherchent à la définir mathématiquement, c'est-à-dire à la traduire en une quantité mesurable.

— Dans notre méthodologie, nous avons choisi de diminuer au maximum le nombre de traitements effectués sur les données ainsi que le nombre de paramètres à gérer. C'est la raison pour laquelle nous avons choisi de travailler directement sur les données réelles brutes. Cependant, même si les résultats sont satisfaisants, il n'est pas justifié qu'apprendre le dictionnaire en se basant sur les données réelles, c'est-à-dire des signaux bruités, permet d'être le plus représentatif possible de la classe considérée. Les perspectives sont, d'une part, de regarder les effets du débruitage de l'observation d'entrée sur les performances de reconnaissance et, d'autre part, d'analyse les effets du débruitage des atomes du dictionnaire sur ces mêmes performances.

— Une perspective dans le même ordre d'idée que le débruitage est de chercher à travailler sur des dictionnaires basés sur des modèles de représentations paramétriques des vocalises considérées. Nous pensons par exemple à l'utilisation de chirplettes, d'ondelettes ou de toutes autres représentations du signal pertinentes. Par exemple, si les données à reconnaître sont des chirps, il serait possible de construire chaque atome comme des chirps de même paramètres temporels et fréquentiels que les vocalises d'apprentissage. De cette façon, les représentants des classes (atomes du dictionnaire) devraient s'affranchir le plus possible de l'environnement des données. Cette perspective a d'ailleurs déjà été initiée dans [SOCHELEAU et SAMARAN, 2017] et donne, dans le cas de ce travail, des résultats de détection de D-call légèrement meilleurs que l'utilisation de données réelles pour l'apprentissage.

— Pour aller plus loin, une évolution de la méthode serait de l'employer à faire l'apprentissage en ligne directement dans le contexte opérationnel. Les perspectives concerneraient la mise en place de « dictionnaires jetables » s'intégrant directement au contexte d'application. Un problème ouvert serait, après avoir exploré le cas d'apprentissage de dictionnaire temporaire supervisé (initialisation par une base d'apprentissage annotée) et semi-supervisé (données sélectionnées « au fil de l'eau » par un expert), d'aborder l'apprentissage non supervisé. Une telle méthode permettrait d'analyser automatiquement l'environnement sonore présent en gérant la détection et l'apprentissage de classes inconnues. De plus, l'aspect jetable n'empêcherait en rien de conserver un suivi statistique de toutes les informations pertinentes du contexte considéré.

— En parallèle de ces considérations, il serait positif de poursuivre le travail d'évaluation en réalisant des tests subjectifs avec des experts en reconnaissance de vocalises de mammifères marins. Ces tests auraient pour but de développer des outils de comparaison entre la méthode et l'être humain, dans la même démarche, ou non, que celle proposée dans [URAZGHILDIIEV et CLARK, 2007].

Perspectives spécifiques au contexte applicatif de SERCEL

L'entreprise SERCEL possède de nombreux systèmes qui ouvrent des perspectives intéressantes.

— Nous avons développé une méthode qui fonctionne pour le cas où l'acquisition des données est mono-capteur. Avec l'architecture multi-capteurs de SERCEL, il serait possible, avec le traitement d'antenne adapté, de faire de la séparation de sources sonores et ainsi d'améliorer le SNR des éventuels signaux d'intérêts avec comme avantage de réduire les potentielles interférences.

— Une application future de la mise en œuvre de notre prototype de détecteur mutli-classes serait d'intégrer notre méthode au logiciel QuietSea. Cette intégration pourra prendre en compte le traitement parallèle des classes et mettre à l'épreuve notre méthode en situation réelle.

— Lors des campagnes sismiques, nous avons accès à beaucoup d'informations comme des *a priori* sur le canal de propagation, les données bathymétriques, etc. Une perspective serait d'essayer de prévoir, grâce à toutes ces informations, à quoi pourrait ressembler la propagation des tirs sismiques produits par le bateau considéré. Notons que cette prédiction ne semble pas envisageable dans le cas où plusieurs campagnes sismiques ont lieu en même temps. Néanmoins, il reste pertinent de considérer la possibilité de créer un dictionnaire dédié à la reconnaissance des tirs sismiques.

— De par son histoire, l'entreprise SERCEL a déjà acquis des données associées à des missions spécifiques. Nous n'avons pas été en mesure d'utiliser ces données pour plusieurs raisons. Elles ne sont pas annotées, elles sont volumineuses et la présence de vocalises de mammifères marins n'est pas garantie. Il est alors intéressant d'effectuer un traitement de ces données par notre détecteur multiclasses. Cela permettrait d'obtenir des statistiques qui pourraient peut-être donner des informations pertinentes sur la répartition des mammifères marins ainsi que sur l'environnement sonore en général. De plus, les données, une fois

annotées, pourraient servir pour tester et/ou comparer d'autres méthodes de reconnaissance directement dans le contexte d'application de SERCEL.

Proposition au service de la communauté bioacoustique

Comme nous l'avons vu à travers ce manuscrit, la communauté bioacoustique ne possède pas encore de processus de validation de méthode de détection et/ou de reconnaissance standardisé complet. Cette lacune rend difficile, pour un nouveau chercheur, l'intégration et la comparaison de ses travaux par rapport à ses prédécesseurs. Nous proposons dès lors de remédier à ce manque en discutant des perspectives suivantes.

— Il nous apparaît urgent que la communauté bioacoustique et les industriels définissent ensemble les besoins actuels, afin d'identifier et de mettre à jour par la suite des critères d'évaluation (performances fixées à mesurer) et des données correspondants à ces différents contextes. C'est ce que commence à faire la communauté bioacoustique qui fournit de plus en plus de données réelles, notamment à chaque *workshop* de la conférence [DCLDE, 2015], ainsi que par l'intermédiaire de la base de données de mobysound présentée dans [MELLINGER et CLARK, 2006], accessible sur le site [MOBYSOUND.ORG]. Nous proposons, dans la continuité de ces actions, de créer un « comité de validation et de valorisation des données » qui aurait pour mission de maintenir les données à jour en plus de valider et de rendre facilement accessibles les nouvelles bases de données. La valorisation des données pourrait se faire via la publication d'articles dédiés aux bases de données.

— Il nous semble pertinent d'explorer la construction de base de données simulées, c'est-à-dire complètement connues mathématiquement, qui permettrait d'interpréter en profondeur les propriétés générales les méthodes proposées. Ces données ne seraient pas utilisées pour la validation. Leur rôle serait de tester facilement et rapidement la mise en œuvre de nouvelles méthodes sur des signaux typiques du contexte considéré avant la validation sur les données réelles validées par le comité.

Liste complète des références

ADAM, O. 2006a, «Advantages of the Hilbert Huang transform for marine mammals signals analysis», *J. Acoust. Soc. Am.*, vol. 120, nº 5, p. 2965–2973. *6 citations pages xv, xvi, 36, 37, 42, et 43*

ADAM, O. 2006b, «The use of the Hilbert-Huang transform to analyze transient signals emitted by sperm whales», *Appl. Acoust.*, vol. 67, nº 11-12, doi :10.1016/j.apacoust.2006.04.001, p. 1134–1143, ISSN 0003682X. *2 citations pages 36 et 43*

AHARON, M., M. ELAD et A. BRUCKSTEIN. 2006, «K-SVD : An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation», *IEEE Trans. Sig. Proc.*, vol. 54, nº 11, p. 4311–4322. *Cité page 59*

ANDRÉ, M., M. VAN DER SCHAAR, S. ZAUGG, L. HOUÉGNIGAN, A. M. SÁNCHEZ et J. V. CASTELL. 2011, «Listening to the deep : live monitoring of ocean noise and cetacean acoustic signals», *Mar. Pollut. Bull.*, vol. 63, nº 1, p. 18–26. *Cité page 33*

ANTITHESIS (**ANT**ILLES **THE**RMICITÉ **SIS**MOGENÈSE). «Campagne inter-disciplinaire pour l'étude du potentiel sismogène de la déformation tectonique et de la migration des baleines le long du segment nord de la marge des Petites Antilles (Guadeloupe - Iles Vierge)», https://ska-france.oca.eu/fr/104-antitesis/antithesis-protection-cetaces. [(en ligne), consulté le 26 mars 2018]. *Cité page 24*

BAHOURA, M. et Y. SIMARD. 2010, «Blue whale calls classification using short-time Fourier and wavelet packet transforms and artificial neural network», *Digit. Signal Process.*, vol. 20, nº 4, p. 1256–1263. *2 citations pages xvi et 48*

BAUMGARTNER, M. F. et S. E. MUSSOLINE. 2011, «A generalized baleen whale call detection and classification system», *J. Acoust. Soc. Am.*, vol. 139, p. 2889–2902. *7 citations pages xix, 40, 42, 43, 46, 66, et 67*

BINDER, C. M. et P. HINES. 2012, «Applying automatic aural classification to cetacean vocalizations», dans *Proc. Meet. Acoust. ECUA2012*, vol. 17, ASA, ISBN 9781622761920, ISSN 1939800X, p. 70029, doi :10.1121/1.4770058. URL http://scitation.aip.org/content/asa/journal/poma/17/1/10.1121/1.4770058. *4 citations pages xv, 33, 38, et 43*

BITTLE, M. et A. DUNCAN. 2013, «A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring», *Proc. Acoust.*, vol. 2013, nº November. *Cité page 46*

BONNEL, J. 2010, *Analyse de la dispersion acoustique UBF (0-150 Hz) pour la surveillance et la caractérisation du milieu marin*, thèse de doctorat, Institut National Polytechnique de Grenoble-INPG. *2 citations pages xv et 21*

BOYD, I., B. BROWNELL, D. CATO, C. CLARKE, D. COSTA, P. G. H. EVANS, J. GEDAMKE, R. GENTRY, B. GISINER, J. GORDON et OTHERS. 2008, «The effects of anthropogenic sound on marine mammals : A draft research strategy», *Eur. Sci. Found. Mar. Board, Ostend. 96pp.* *Cité page 24*

CBI (**C**OMMISSION **B**ALEINIÈRES **I**NTERNATIONALE). 2018, https://iwc.int/cetacea-fr. [(en ligne), consulté le 23 avril 2018]. *Cité page 14*

CLAPHAM, P. J. 1996, «The social and reproductive biology of humpback whales : an ecological perspective», *Mamm. Rev.*, vol. 26, n° 1, p. 27–49. *Cité page 17*

CLAY, C. S. et H. MEDWIN. 1977, *Acoustical Oceanography, principles and applications.*, Wiley-Interscience. *Cité page 21*

CUMMINGS, W. C. 1971, «Underwater Sounds from the Blue Whale, Balaenoptera musculus», *J. Acoust. Soc. Am.*, vol. 50, n° 4B, doi :10.1121/1.1912752, p. 1193, ISSN 00014966. URL http://scitation.aip.org/content/asa/journal/jasa/50/4B/10.1121/1.1912752. *Cité page 19*

DCLDE. 2015, «Detection, Classification, Localisation and Density Estimation of marine mammals using passive acoustics)», http://www.cetus.ucsd.edu/dclde/datasetDocumentation.html. [(en ligne), consulté le 30 avril 2018]. *4 citations pages 52, 62, 64, et 109*

DELARUE, J., B. MARTIN, J. VALLARTA, X. MOUY, J. MACDONNELL, N. E. CHORNEY et D. HANNAY. 2010, *Appendix A. Automated Detection and Classification of Marine Mammal Vocalizations*, JASCO Applied Sciences. *2 citations pages xv et 33*

DZIAK, R. P., J.-Y. ROYER, J. H. HAXEL, M. DELATRE et D. R. B. ET AL. 2008, «Hydroacoustic detection of recent seafloor volcanic activity in the southern Indian Ocean», dans *Trans. Am. Geophys. Union, Fall Meet. T13. San Fr. CA*, p. 1. *Cité page 63*

ELAD, M. 2010, *Sparse and Redundant Representations : From Theory to Applications in Signal and Image Processing*, 1re éd., Springer Publishing Company, Incorporated. *2 citations pages 57 et 58*

ENGAN, K., S. O. AASE et J. HAKON HUSOY. 1999, «Method of Optimal Directions for Frame Design», dans *Proc. Acoust. Speech, Signal Process. 1999. 1999 IEEE Int. Conf. - Vol. 05*, Washington, DC, USA, ISBN 0-7803-5041-3, ISSN 1520-6149, p. 2443–2446, doi :10.1109/ICASSP.1999.760624. URL http://ieeexplore.ieee.org/document/760624/. *Cité page 59*

ERBE, C. 2000, «Detection of whale calls in noise : Performance comparison between a beluga whale, human listeners, and a neural network», *J. Acoust. Soc. Am.*, vol. 108, n° 1, p. 297–303. *Cité page 38*

FORTESCUE, P., S. O. HOLE, R. M. ROBICHAUD, J. R. SESTO, J. THERIAULT, R. HENSLEY, N. WEAVER et B. MAUGHAN. 2005, «Marine Mammals and Active Sonar», cahier de recherche, DEFENCE RESEARCH AND DEVELOPMENT ATLANTIC DARTMOUTH (CANADA). *Cité page 24*

FRIEDMAN, J., T. HASTIE et R. TIBSHIRANI. 2001, *The elements of statistical learning*, vol. 1, Springer series in statistics Springer, Berlin. *Cité page 57*

FUKADA, T., K. TOKUDA, T. KOBAYASHI et S. IMAI. 1992, «An adaptive algorithm for mel-cepstral analysis of speech», dans *Acoust. Speech, Signal Process. 1992. ICASSP-92., 1992 IEEE Int. Conf.*, vol. 1, IEEE, p. 137–140. *Cité page 34*

GEDAMKE, J., D. P. COSTA et A. DUNSTAN. 2001, «Localization and visual verification of a complex minke whale vocalization», *J. Acoust. Soc. Am.*, vol. 109, n° 6, doi :10.1121/1.1371763, p. 3038–3047, ISSN 0001-4966. URL http://asa.scitation.org/doi/10.1121/1.1371763. *Cité page 19*

GILLESPIE, D. 2004, «Detection and classification of right whale calls using an'edge'detector operating on a smoothed spectrogram», *Canadian Acoustics*, vol. 32, n° 2, p. 39–47. *5 citations pages xv, 40, 41, 42, et 43*

HALKIAS, X. C., S. PARIS et H. GLOTIN. 2013, «Classification of mysticete sounds using machine learning techniques», *J. Acoust. Soc. Am.*, vol. 134, nº 5, doi :10.1121/1.4821203, p. 3496–3505, ISSN 0001-4966. URL http://asa.scitation.org/doi/10.1121/1.4821203. *10 citations pages xvi, 38, 40, 43, 44, 48, 49, 60, 64, et 67*

HARLAND, E. J. et M. S. ARMSTRONG. 2004, «The real-time detection of the calls of cetacean species», *Can. Acoust.*, vol. 32, nº 2, p. 76–82. *Cité page 34*

HUANG, N. E., Z. SHEN, S. R. LONG, M. L. C. WU, H. H. SHIH, Q. ZHENG, N.-C. C. YEN, C. C. TUNG et H. H. LIU. 1998, «The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis», *Proc. R. Soc. London A*, vol. 454, nº 1971, p. 903–995. *Cité page 36*

IUCN (**U**NION **I**NTERNATIONALE POUR LA **C**ONSERVATION DE LA **N**ATURE). 2018, http://www.iucn.org/fr/. [(en ligne), consulté le 23 avril 2018]. *2 citations pages 14 et III*

KRAUSE, B. 2008, «Anatomy of the Soundscape : Evolving Perspectives», *J. Audio Eng. Soc*, vol. 56, nº 1/2, p. 73–80. URL http://www.aes.org/e-lib/browse.cfm?elib=14377. *Cité page 13*

LEROY, E. C., F. SAMARAN, J. BONNEL et J.-Y. ROYER. 2017, «Identification of two potential whale calls in the southern Indian Ocean, and their geographic and seasonal occurrence», *J. Acoust. Soc. Am.*, vol. 142, nº 3, p. 1413–1427. *Cité page 64*

LEROY, E. C., F. SAMARAN, J. BONNEL et J. Y. J.-Y. ROYER. 2016, «Seasonal and Diel Vocalization Patterns of Antarctic Blue Whale (Balaenoptera musculus intermedia) in the Southern Indian Ocean : A Multi-Year and Multi-Site Study», *PLoS One*, vol. 11, nº 11, doi :10.1371/journal.pone.0163587, p. e0163 587, ISSN 19326203. *3 citations pages xv, 23, et 64*

LJUNGBLAD, D. K., P. O. THOMPSON et S. E. MOORE. 1982, «Underwater sounds recorded from migrating bowhead whales, B alaenamysticetus, in 1979», *J. Acoust. Soc. Am.*, vol. 71, nº 2, p. 477–482. *Cité page 19*

LOPATKA, M., A. OLIVIER, C. LAPLANCHE, J. ZARZYCKI et J.-F. MOTSCH. 2005, «An attractive alternative for sperm whale click detection using the wavelet transform in comparison to the Fourier spectrogram», *Aquat. Mamm.*, vol. 31, nº 4, p. 463. *4 citations pages xv, 35, 36, et 42*

MAIRAL, J., F. BACH, J. PONCE et G. SAPIRO. 2010, «Online Learning for Matrix Factorization and Sparse Coding», *J. Mach. Learn. Res.*, vol. 11, p. 19–60. *3 citations pages 58, 59, et 69*

MANN, J. 2000, *Cetacean societies : field studies of dolphins and whales*, University of Chicago Press. *2 citations pages xv et 16*

MCDONALD, M. A., J. A. HILDEBRAND et S. MESNICK. 2009, «Worldwide decline in tonal frequencies of blue whale songs», *Endanger. Species Res.*, vol. 9, nº 1, doi :10.3354/esr00217, p. 13–21, ISSN 18635407. *3 citations pages xv, 18, et 47*

MELLINGER, D. K. 2004, «A comparison of methods for detecting right whale calls», *Can. Acoust.*, vol. 32, nº 2, p. 55–65, ISSN 2291-1391. URL http://jcaa.caa-aca.ca/index.php/jcaa/article/view/1588{%}5Cnpapers2://publication/uuid/166D7DD7-F963-496A-8864-33DAEA90C72A. *Cité page 38*

MELLINGER, D. K., C. D. CARSON et C. W. CLARK. 2000, «Characteristics of Minke Whale (Balaenoptera Acutorostrata) Pulse Trains Recorded Near Puerto Rico», *Mar. Mammal Sci.*, vol. 16, nº 4, doi :10.1111/j.1748-7692.2000.tb00969.x, p. 739–756, ISSN 0824-0469. URL http://doi.wiley.com/10.1111/j.1748-7692.2000.tb00969.x. *Cité page 19*

MELLINGER, D. K. et C. W. CLARK. 2000, «Recognizing transient low-frequency whale sounds by spectrogram correlation», *The Journal of the Acoustical Society of America*, vol. 107, nº 6, p. 3518–3529. *3 citations pages xvi, 33, et 46*

MELLINGER, D. K. et C. W. CLARK. 2006, «MobySound : A reference archive for studying automatic recognition of marine mammal sounds», *Appl. Acoust.*, vol. 67, nº 11-12, p. 1226–1242. *2 citations pages 52 et 109*

MELLINGER, D. K., S. W. MARTIN, R. P. MORRISSEY, L. THOMAS et J. J. YOSCO. 2011, «A method for detecting whistles, moans, and other frequency contour sounds», *J. Acoust. Soc. Am.*, vol. 129, nº 6, doi :10.1121/1.3531926, p. 4055–4061, ISSN 0001-4966. URL http://asa.scitation.org/doi/10.1121/1.3531926. *Cité page 33*

MOBYSOUND.ORG. «Base de données pour la recherche en reconnaissance automatique de cris de mammfères marins. Ce site est géré par Sara Heimlich, Holger Klinck et Dans Mellinger au group bioacoustique CIMRS», http://www.mobysound.org. *6 citations pages xvi, 52, 53, 54, 73, et 109*

NG, A. 2011, «Sparse autoencoder», *CS294A Lecture notes*, vol. 72. *5 citations pages xv, xvi, 38, 39, et 44*

ORTIZ, E. G., A. WRIGHT et M. SHAH. 2013, «Face recognition in movie trailers via mean sequence sparse representation-based classification», *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, doi :10.1109/CVPR.2013.453, p. 3531–3538, ISSN 10636919. *Cité page 58*

OSPAR. 2018, «Underwater Noise», \url{https ://www.ospar.org/work-areas/eiha/noise}. *Cité page 13*

PARSONS, E. C. M., A. J. WRIGHT et M. A. GORE. 2008, «The nature of humpback whale (Megaptera novaeangliae) song», *J. Mar. Anim. Their Ecol.*, vol. 1, nº 1, p. 22–31. URL http://www.oers.ca/journal/Volume1/issue1vol1-2008-JMATE.pdf{#}page=23. *Cité page 19*

PATI, Y. C., R. REZAIIFAR et P. S. KRISHNAPRASAD. 1993, «Orthogonal matching pursuit : recursive function approximation with applications to wavelet decomposition», dans *Signals, Syst. Comput. 1993. 1993 Conf. Rec. Twenty-Seventh Asilomar Conf.*, ISBN 0-8186-4120-7, ISSN 1058-6393, p. 40–44 vol.1, doi :10.1109/ACSSC.1993.342465. URL http://ieeexplore.ieee.org/document/342465/. *Cité page 58*

PAYNE, R. S. et S. MCVAY. 1971, «Songs of Humpback Whales», *Science (80-.).*, vol. 173, nº 3997, p. 585–597. *2 citations pages xv et 19*

ROCH, M. A., H. KLINCK, S. BAUMANN-PICKERING, D. K. MELLINGER, S. QUI, M. S. SOLDEVILLA et J. A. HILDEBRAND. 2011, «Classification of echolocation clicks from odontocetes in the Southern California Bight», *J. Acoust. Soc. Am.*, vol. 129, nº 1, p. 467–475. *Cité page 34*

ROCH, M. A., M. S. SOLDEVILLA, J. C. BURTENSHAW, E. E. HENDERSON et J. A. HILDEBRAND. 2007, «Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California», *J. Acoust. Soc. Am.*, vol. 121, nº 3, doi :10.1121/1.2400663, p. 1737–1748, ISSN 0001-4966. URL http://asa.scitation.org/doi/10.1121/1.2400663. *4 citations pages xv, xvi, 34, et 43*

SAMARAN, F. et C. GUINET. 2012, «*Observatoire Acoustique des Grands Cétacés dans l'Océan Austral.*, Rapport Final pour la Fondation Total», . *3 citations pages xv, 17, et 18*

SAMARAN, F., K. M. STAFFORD, T. A. BRANCH, J. GEDAMKE, J.-Y. ROYER, R. P. DZIAK et C. GUINET. 2013, «Seasonal and Geographic Variation of Southern Blue Whale Subspecies in the Indian Ocean», *PLoS One*, vol. 8, nº 8, doi :10.1371/journal.pone.0071561, p. 1–10. *2 citations pages 62 et 63*

SCHARF, L. L. 1991, *Statistical Signal Processing : Detection, Estimation, and Time Series Analysis*, Addison-Wesley, pp. 1 - 524, Reading, Massachusetts. *Cité page 60*

SERCEL. «Sercel», http://www.sercel.com/. [(en ligne), consulté le 23 avril 2018]. *17 citations pages xv, 7, 8, 12, 23, 24, 25, 26, 65, 76, 77, 78, 89, 90, 105, 108, et 109*

SHANNON, R., D. LJUNGBLAD, C. CLARK, H. KATO+, S. RANKIN, D. LJUNGBLAD, C. CLARK, H. KATO, R. S., L. D., C. C. et K. H. 2005, «Vocalisations of Antarctic blue whales, Balaenoptera musculus intermedia, recorded during the 2001/2002 and 2002/2003 IWC/SOWER circumpolar cruises, Area V, Antarctica», *J. Cetacean Res. Manag.*, vol. 7, nº 1, p. 13–20. *Cité page 64*

SHIN, Y., S. LEE, M. AHN, H. CHO, S. C. JUN et H. N. LEE. 2015, «Noise robustness analysis of sparse representation based classification method for non-stationary EEG signal classification», *Biomed. Signal Process. Control*, vol. 21, doi :10.1016/j.bspc.2015.05.007, p. 8–18, ISSN 17468108. URL http://dx.doi.org/10.1016/j.bspc.2015.05.007. *Cité page 58*

ŠIROVIĆ, A., J. A. HILDEBRAND, S. M. WIGGINS, A. SIROVIC, J. A. HILDEBRAND et S. M. WIGGINS. 2007, «Blue and fin whale call source levels and propagation range in the Southern Ocean», *J. Acoust. Soc. Am.*, vol. 122, nº 2, p. 1208–1215. *Cité page 22*

SIROVIC, A., J. A. HILDEBRAND, S. M. WIGGINS, D. THIELE, A. ŠIROVIĆ, J. A. HILDEBRAND, S. M. WIGGINS, D. THIELE, A. \uSIROVIĆ, J. A. HILDEBRAND, S. M. WIGGINS et D. THIELE. 2009, «Blue and fin whale acoustic presence around Antarctica during 2003 and 2004», *Mar. Mammal Sci.*, vol. 25, nº 1, doi :10.1111/j.1748-7692.2008.00239.x, p. 125–136, ISSN 08240469. *Cité page 62*

SOCHELEAU, F.-X., E. LEROY, A. CARVALLO PECCI, F. SAMARAN, J. BONNEL et J.-Y. ROYER. 2015, «Automated detection of Antarctic blue whale calls», *J. Acoust. Soc. Am.*, vol. 138, nº 5, p. 3105–3117. *Cité page 60*

SOCHELEAU, F.-X. et F. SAMARAN. 2017, «Detection of Mysticete Calls : a Sparse Representation-Based Approach», *IMT Atlantique, Research report RR-2017-04-SC*. URL https://hal.archives-ouvertes.fr/hal-01736178/document. *8 citations pages 60, 67, 68, 71, 79, 100, 102, et 108*

SOS GRAND BLEU. 2018, «La législation pour la protection des cétacés par l'association sos grand bleu», http://www.sosgrandbleu.asso.fr/dossiers/la-legislation-pour-la-protection-des-cetaces/. [(en ligne), consulté le 23 novembre 2017]. *Cité page 24*

SOUTHALL, B. L., A. E. BOWLES, W. T. ELLISON, J. J. FINNERAN, R. L. GENTRY, C. R. GREENE JR, D. KASTAK, D. R. KETTEN, J. H. MILLER, P. E. NACHTIGALL et OTHERS. 2008, «Marine mammal noise-exposure criteria : initial scientific recommendations», *Bioacoustics*, vol. 17, nº 1-3, p. 273–275. *Cité page 24*

STAFFORD, K. M., D. R. BOHNENSTIEHL, M. TOLSTOY, E. CHAPP, D. K. MELLINGER et S. E. MOORE. 2004, «Antarctic type blue whale calls recorded at low latitudes in the Indian and eastern Pacific Oceans», *Deep Sea Res., Part I*, vol. 51, p. 1337–1346. *Cité page 62*

STAFFORD, K. M., S. E. MOORE, K. L. LAIDRE et M. P. HEIDE-JØRGENSEN. 2008, «Bowhead whale springtime song off West Greenland», *J. Acoust. Soc. Am.*, vol. 124, nº 5, doi :10.1121/1.2980443, p. 3315–3323, ISSN 0001-4966. URL http://asa.scitation.org/doi/10.1121/1.2980443. *Cité page 20*

STEVENS, K. N., T. M. COVER et P. E. HART. 1967, «Nearest Neighbor pattern classification», vol. I. *Cité page 57*

TAN, L. N., G. KOSSAN, M. L. CODY, C. E. TAYLOR et A. ALWAN. 2013, «A sparse representation-based classifier for in-set bird phrase verification and classification with limited training data», *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, doi :10.1109/ICASSP.2013.6637751, p. 763–767, ISSN 15206149. *Cité page 58*

TERVO O.M. 2011, «Acoustic behaviour of bowhead whales Balaena mysticetus in Disko Bay, Western Greenland», *Tesis Dr.*, , nº April, p. 138. *3 citations pages xv, 19, et 20*

THOMPSON, P. O., L. T. FINDLEY, O. VIDAL et W. C. CUMMINGS. 1996, «UNDERWATER SOUNDS OF BLUE WHALES, BALAENOPTERA MUSCULUS, IN THE GULF OF CALIFORNIA, MEXICO», *Mar. Mammal Sci.*, vol. 12, nº 2, doi :10.1111/j.1748-7692.1996.tb00578.x, p. 288–293, ISSN 1748-7692. *2 citations pages 57 et 62*

TRYGONIS, V., E. GERSTEIN, J. MOIR et S. MCCULLOCH. 2013, «Vocalization characteristics of north atlantic right whale surface active groups in the calving habitat, southeastern united states», *The Journal of the Acoustical Society of America*, vol. 134, nº 6, p. 4518–4531. *2 citations pages xv et 18*

TSANG-HIN-SUN, E., J.-Y. ROYER et J. PERROT. 2016, «Seismicity and active accretion processes at the ultraslow-spreading Southwest and intermediate-spreading Southeast Indian ridges from hydroacoustic data», *Geophys. J. Int.*, vol. 206, nº 2, p. 1232–1245. *Cité page 64*

UNEP (**U**NITED **N**ATIONS **E**NVIRONMENT **P**ROGRAMM). «Scientific synthesis on the impacts of underwater noise on marine and coastal biodiversity and habitats, convention on biological diversity», . *Cité page 24*

UNIVERSALIS. 2017, ««GEOPHYSIQUE », encyclopædia universalis [en ligne], rédigé par michel cara, consulté le 16 mars 2016.», http://www.universalis.fr/encyclopedie/geophysique/. [(en ligne), consulté le 23 avril 2018]. *Cité page 23*

URAZGHILDIIEV, I. R. et C. W. CLARK. 2007, «Detection performances of experienced human operators compared to a likelihood ratio based detector», *The Journal of the Acoustical Society of America*, vol. 122, nº 1, p. 200–204. *Cité page 108*

URAZGHILDIIEV, I. R., C. W. CLARK, T. P. KREIN et S. E. PARKS. 2009, «Detection and Recognition of North Atlantic Right Whale Contact Calls in the Presence of Ambient Noise», *IEEE J. Ocean. Eng.*, vol. 34, nº 3, doi :10.1109/JOE.2009.2014931, p. 358–368, ISSN 15581691. *2 citations pages 60 et 67*

WATKINS, W. A., P. TYACK, K. E. MOORE et J. E. BIRD. 1987, «The 20-Hz signals of finback whales (Balaenoptera physalus)», *J. Acoust. Soc. Am.*, vol. 82, nº 6, p. 1901–1912. *Cité page 19*

WESTWOOD, E. K., C. T. TINDLE et N. R. CHAPMAN. 1996, «A normal mode model for acousto-elastic ocean environments», *J. Acoust. Soc. Am.*, vol. 100, nº 6, p. 3631–3645. *Cité page 71*

WIGGINS, S. S. M., M. A. MCDONALD, L. M. MUNGER, S. E. MOORE et J. A. HILDEBRAND. 2004, «Waveguide propagation allows range estimates for North Pacific right whales in the Bering Sea», *Can. Acoust.*, vol. 32, nº 2, p. 146–154, ISSN 07116659. URL http://jcaa.caa-aca.ca/index.php/jcaa/article/view/1598. *2 citations pages xv et 22*

WORMS (**WO**RLD **R**EGISTER OF **M**ARINE **S**PECIES). http://www.marinespecies.org/index.php. [(en ligne), consulté le 23 avril 2018]. *Cité page 14*

WRIGHT, J., A. Y. YANG, A. GANESH, S. S. SASTRY et Y. MA. 2009, «Robust face recognition via sparse representation.», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, nº 2, doi :10.1109/TPAMI.2008.79, p. 210–227, ISSN 01628828. URL http://www.ncbi.nlm.nih.gov/pubmed/21646680. *2 citations pages 58 et 59*

Annexe A

Liste des mysticètes

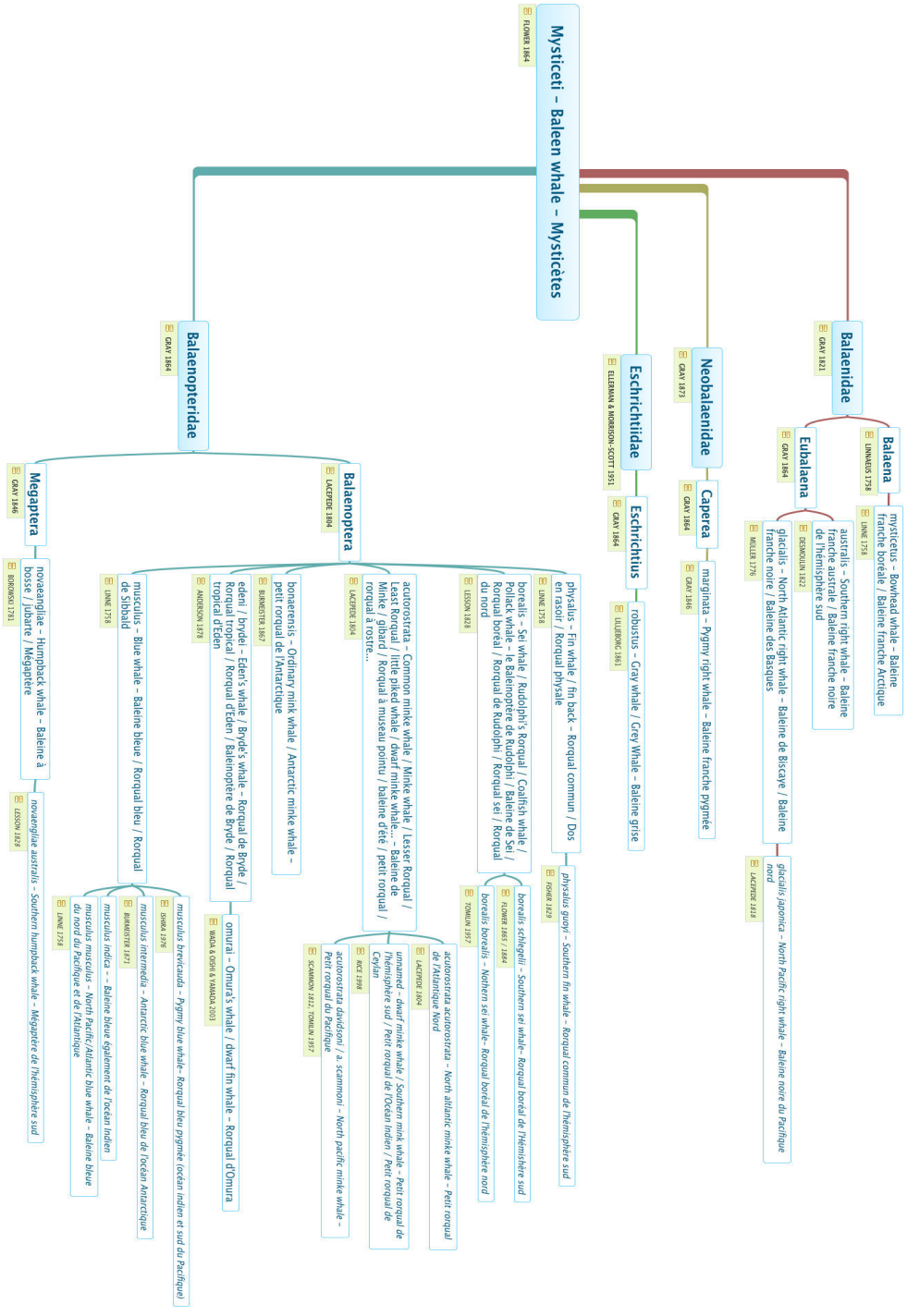


FIGURE A.1 – Liste des mysticètes

Annexe B

Distribution géographique des mysticètes

Nous présentons les différentes distributions d'une majeure partie des mysticètes, extraite de IUCN (**U**NION **I**NTERNATIONALE POUR LA **C**ONSERVATION DE LA **N**ATURE) [2018], ordonnées de la répartition la plus localisée à la moins localisée.

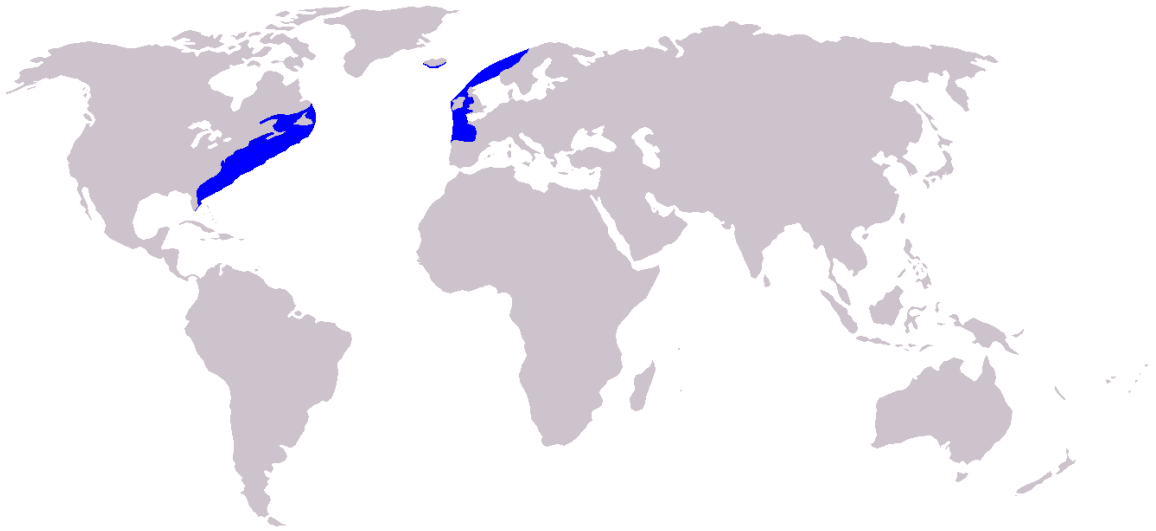


FIGURE B.1 – La répartition de la baleine franche de l'Atlantique Nord

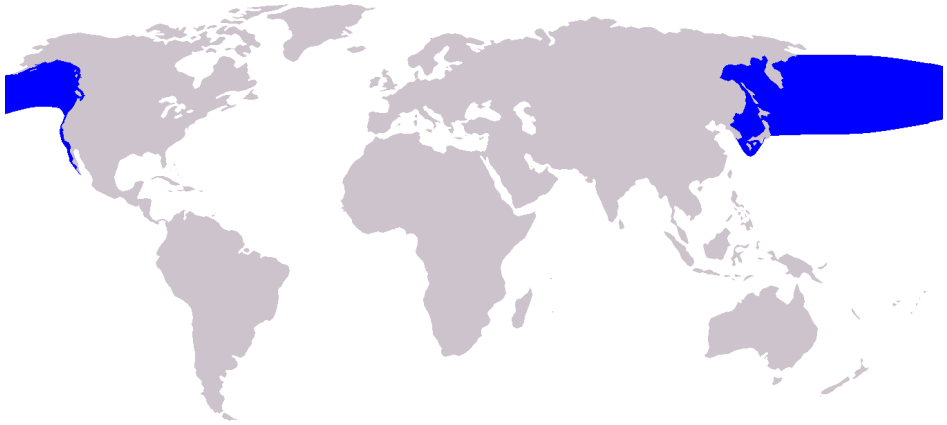


FIGURE B.2 – La répartition de la baleine franche du Pacifique Nord

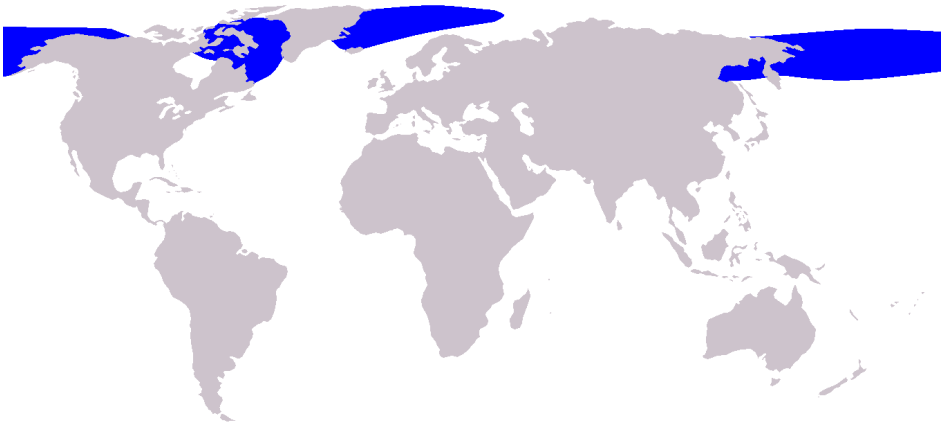


FIGURE B.3 – La répartition de la baleine franche Arctique

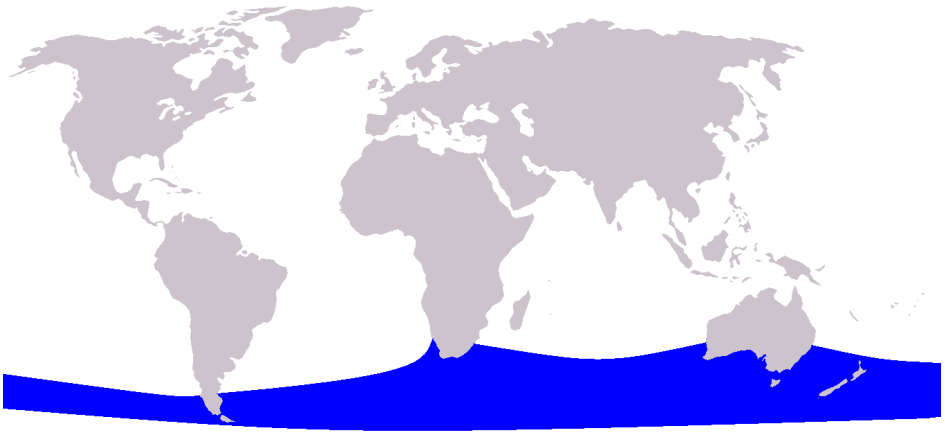


FIGURE B.4 – La répartition de la baleine franche pygmée

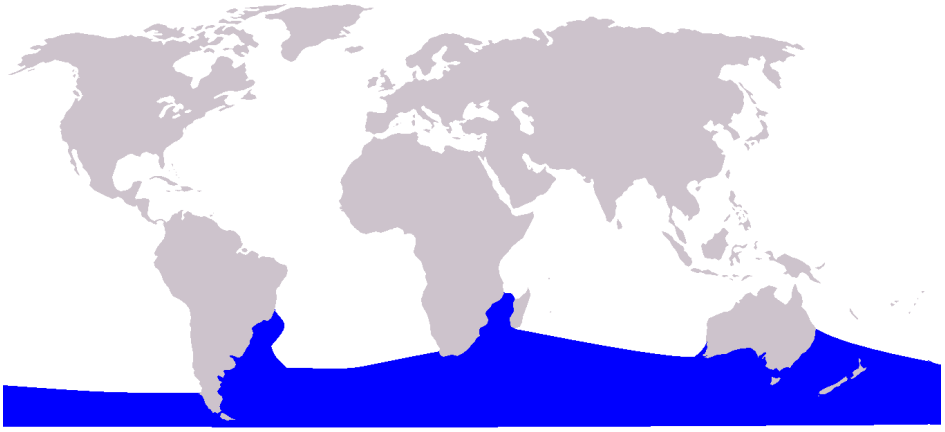


FIGURE B.5 – La répartition de la baleine franche de l'hémisphère sud

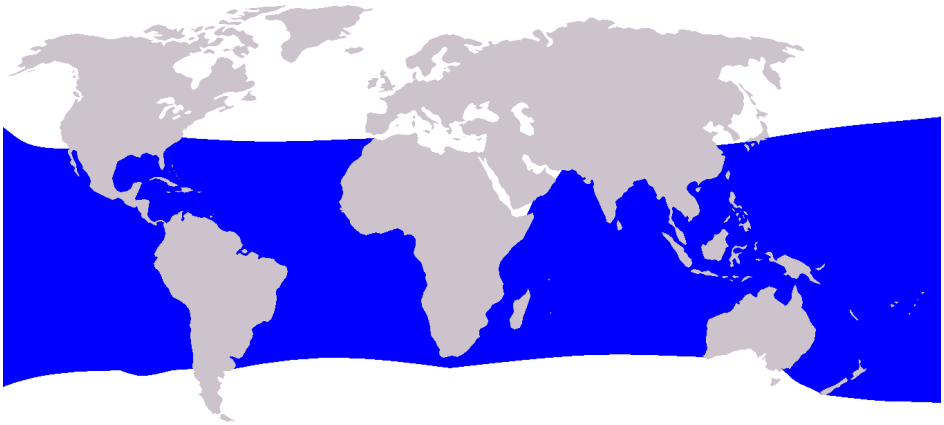


FIGURE B.6 – La répartition du rorqual de Bryde

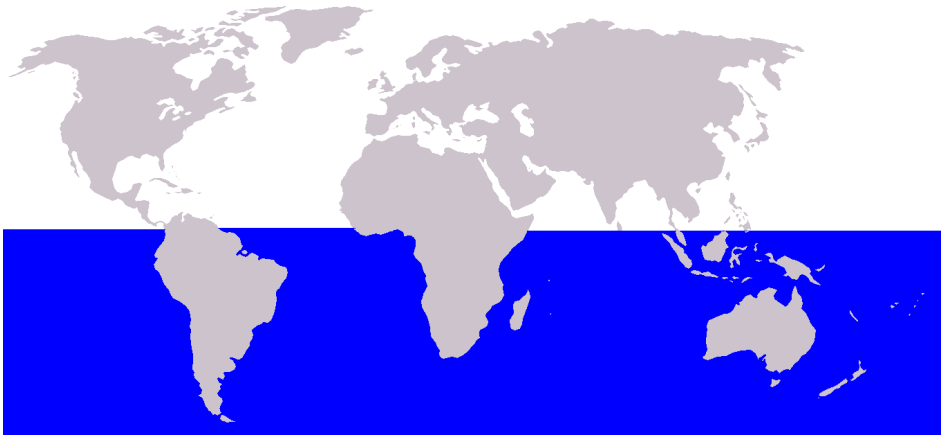


FIGURE B.7 – La répartition du petit rorqual de l'Antarctique

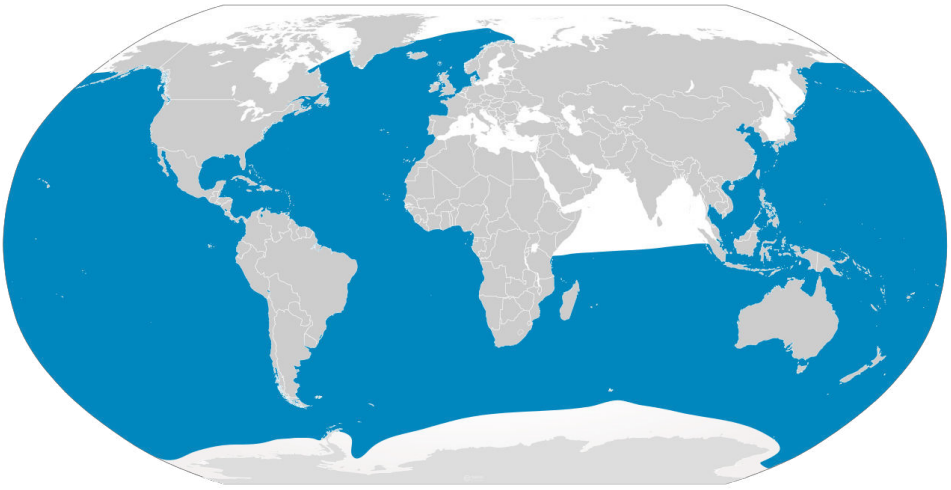


FIGURE B.8 – La répartition du rorqual boreal

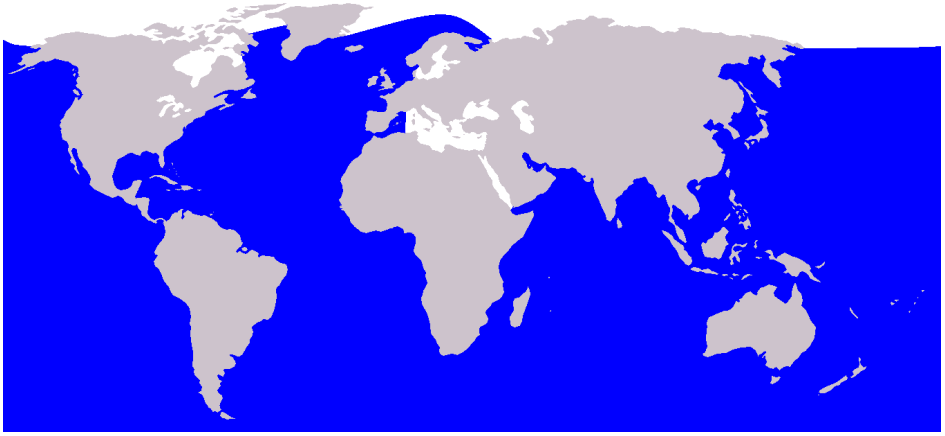


FIGURE B.9 – La répartition de la baleine à bosse

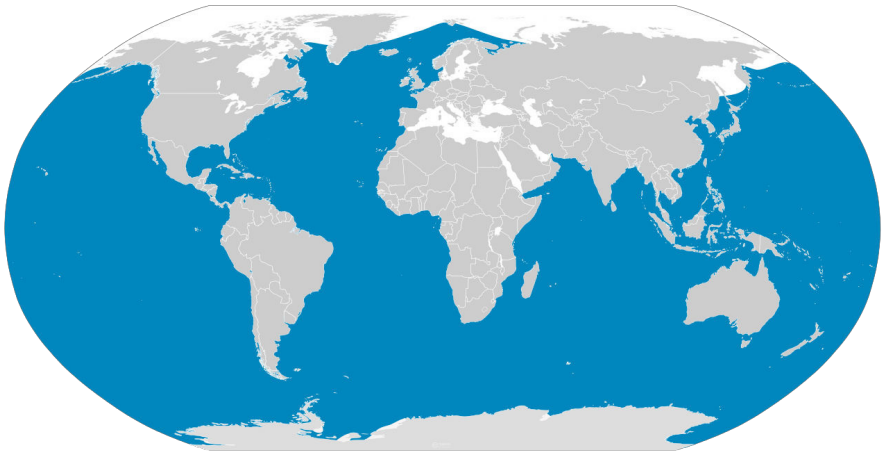


FIGURE B.10 – La répartition de la baleine bleue

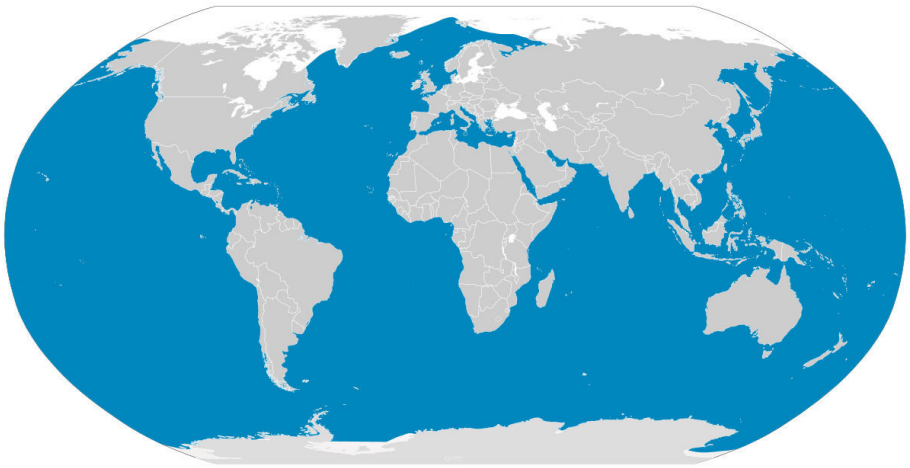


FIGURE B.11 – La répartition du rorqual commun

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

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

Titre : Classification de vocalises de mammifères marins en environnement sismique

Mots clés : reconnaissance, mammifères marins, sismique, représentations parcimonieuses, apprentissage automatique

Résumé : En partenariat avec l'entreprise Sercel, la thèse concerne la mise en œuvre d'algorithmes de reconnaissance des sons émis par les mysticètes (baleines à fanons). Ces sons peuvent être étudiés grâce aux systèmes de surveillance par acoustique passive. L'entreprise Sercel, par ses activités sismiques liées à la prospection pétrolière, a son propre logiciel pour détecter et localiser les sources d'énergie sonores sous-marines. Le travail de la thèse consiste dès lors à ajouter un module de reconnaissance pour identifier si l'énergie détectée et localisée correspond bien à un éventuel mysticète. Les campagnes de tirs sismiques étant onéreuses, la méthode utilisée doit pouvoir réduire la probabilité de fausse alarme, la reconnaissance pouvant infirmer la détection.
La méthode proposée est basée sur l'apprentissage de dictionnaire. Elle est dynamique, modulaire, ne dépend que de peu de paramètres et est robuste aux fausses alarmes. Une expérimentation sur cinq types de vocalises est présentée. Nous obtenons un rappel moyen de 92.1 % tout en rejetant 97.3 % des bruits (persistants et transitoires). De plus, un coefficient de confiance est associé à chaque reconnaissance et permet de réaliser de l'apprentissage incrémental semi-supervisé.
Enfin, nous proposons une méthode capable de gérer la détection et la reconnaissance conjointement. Ce « détecteur multiclasses » respecte au mieux les contraintes de gestion des fausses alarmes et permet d'identifier plusieurs types de vocalises au même instant. Cette méthode est bien adaptée au contexte industriel pour lequel elle est dédiée. Elle ouvre également des perspectives très prometteuses dans le contexte bioacoustique.

Title : Recognition of marine mammal vocalizations in seismic environment

Keywords : Classification, marine mammal, seismic environment, sparse representation, machine learning

Abstract : In partnership with Sercel, the thesis concerns the implementation of algorithms for recognizing the sounds emitted by mysticetes (baleen whales). These sounds can be studied using passive acoustic monitoring systems. Sercel, through its seismic activities related to oil exploration, has its own software to detect and locate underwater sound energy sources. The thesis work therefore consists in adding a recognition module to identify if the detected and localized energy corresponds to a possible mysticete. Since seismic shooting campaigns are expensive, the method used must be able to reduce the probability of false alarms, as recognition can invalidate detection.
The proposed method is based on dictionary learning. It is dynamic, modular, depends on few parameters and is robust to false alarms. An experiment on five types of vocalizations is presented. We obtain an average recall of 92.1% while rejecting 97.3% of the noises (persistent and transient). In addition, a confidence coefficient is associated with each recognition and allows semi-supervised incremental learning to be achieved.
Finally, we propose a method capable of managing detection and recognition together. This "multi-class detector" best respects the constraints of false alarm management and allows several types of vocalizations to be identified at the same time. This method is well adapted to the industrial context for which it is dedicated. It also opens up very promising prospects in the bioacoustic