Antennes microphoniques intelligentes : localisation de sources acoustiques par Deep Learning

Hadrien Pujol

HESAM UNIVERSITÉ

le cnam

ÉCOLE DOCTORALE SMI

Laboratoire de Mécanique des Structures et des Systèmes Couplés

THÈSE

présentée par : **Hadrien PUJOL**
soutenue le : **22 octobre 2020**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée au : **Conservatoire national des arts et métiers**

Discipline : **Sciences pour l'Ingénieur**

Spécialité : **Acoustique**

Antennes microphoniques intelligentes : Localisation de sources acoustiques par Deep Learning

THÈSE dirigée par :
M. Alexandre GARCIA Professeur des Universités, Cnam Paris

et co-encadrée par :
M. Éric BAVU Maître de conférence HDR, Cnam Paris

JURY :

M. Manuel MELON	Professeur des Universités, LAUM, Le Mans Université	Rapporteur
M. Emmanuel VINCENT	Directeur de Recherche, LORIA, INRIA Nancy-Grand Est	Rapporteur
M. Antoine DELEFORGE	Chargé de Recherche, LORIA, INRIA Nancy-Grand Est	Examinateur
M. Laurent GIRIN	Professeur des Universités, GIPSA-LAB, Université Grenoble-INP	Examinateur
Mme Rozenn NICOL	Ingénieur de Recherche HDR, Orange Labs, Lannion	Examinateur
M. Nicolas THOME	Professeur des Universités, CEDRIC, Cnam Paris	Examinateur

THÈSE

À tous mes professeurs, qui ont eu la patience de répondre à toutes mes questions durant mes études, et qui m'ont donné le goût des sciences et de la recherche.

Remerciements

Je voudrais en premier lieu remercier mes directeurs de thèse, Alexandre GARCIA et Éric BAVU. Grâce à leur encadrement, ces trois années sont passées avec une facilité déconcertante. Ils ont accepté de me faire confiance pour ce travail, et m'ont renouvelé cette confiance à chaque moment difficile, et face à chaque problème ils avaient des solutions. Ils ont toujours été avars de reproche et généreux en encouragements. Cette thèse restera dans ma mémoire comme un moment de stimulation scientifique et de bonne humeur grâce à eux.

Mes remerciements vont aussi à tous mes collègues, qui ont partagé tant de repas à la cantine, et de conseil dans les couloirs. En particulier aux courageux membres de la salle des doctorants, qui ont servi à la fois de canards en plastique pour que je puisse exposer mes idées, et de conseillers artistiques pour améliorer la forme de mes présentations. J'espère que la relève saura conserver la bonne ambiance et l'entraide, toutes deux nécessaire à nos réussites respectives sur ces chemins tortueux allant vers la soutenance.

Puis vient le tour de remercier ma famille, toujours prête à écouter mes interrogations scientifiques, et à utiliser les repas du dimanche pour essayer de proposer des solutions ou de nouvelles piste d'exploration. L'intérêt qu'ils ont tous portés à mon sujet m'a motivé et quelque part rendu fière de mon travail.

Enfin merci à ma femme d'avoir été là pendant ces trois années, pour m'épauler dans cette tache mais aussi pour me faire penser à autre choses à chaque fois que ce n'est plus le moment de travailler.

Loin d'être exhaustifs, mes remerciements vont aussi à toutes les personnes qui de près ou de loin ont été là pour moi. Donc en un mot comme en 213 :

Merci !

Résumé

Cette thèse de doctorat propose d'explorer un nouveau paradigme pour la localisation de sources acoustiques : l'apprentissage supervisé. Contrairement aux méthodes plus conventionnelles développées par la communauté scientifique depuis plusieurs décennies, cette approche ne nécessite pas forcément l'utilisation explicite du modèle de propagation des ondes acoustiques, ni de modèles de signaux sous-jacents. Au contraire, une architecture de Deep Learning est ici proposée afin d'extraire les informations pertinentes pour la localisation de sources acoustiques, directement depuis un jeux de signaux microphoniques temporels bruts.

Pour être efficace, cette approche, baptisée BeamLearning, nécessite la constitution de jeux de données conséquents et réalistes, afin d'entraîner le réseau de neurones profond associé. Ces jeux de données peuvent être constitués de deux manières complémentaires. La première repose sur l'exploitation de simulations numériques réalistes de la propagation acoustique, en utilisant en particulier un formalisme des sources images pour réaliser une auralisation multicanale de signaux émis par des sources à localiser dans un environnement réverbérant. La seconde, permettant de constituer des jeux de données expérimentaux de manière reproductible et efficace, repose sur l'utilisation du spatialisateur 3D par ambisonie d'ordres élevés disponible au LMSSC, et qui permet la captation automatisée, pendant plusieurs heures, de sources spatialisées, avec n'importe quelle topologie d'antenne – à condition qu'elle soit suffisamment compacte pour occuper la zone du *sweet spot* de la synthèse ambisonique.

Ces jeux de données permettent alors d'optimiser les variables d'apprentissage d'un réseau de neurones profond développé spécifiquement pour la tâche de localisation de sources au cours de cette thèse, et reposant en particulier sur des couches neuronales de convolutions à trous séparables en profondeur. L'architecture de ce réseau de neurones profond original a été conçue en s'efforçant de dresser un pa-

rallèle entre les opérations issues du monde de l'apprentissage par Deep Learning, et les opérations de traitement du signal communément utilisées par les algorithmes de localisation de sources acoustiques conventionnels reposant sur l'utilisation de modèles de propagation d'ondes sonores.

L'approche BeamLearning, est une méthode qui offre divers avantages par rapport aux méthodes conventionnelles. Tout d'abord, à travers les nombreuses situations testées, les résultats observés démontrent que ses performances de localisation sont *a minima* équivalentes aux performances d'algorithmes de localisation de sources éprouvés, et les dépassent même assez largement en environnement réverbérant et bruité. Par ailleurs, la possibilité d'entraîner le réseau de neurones sur des signaux captés depuis une antenne réelle, permet un étalonnage implicite des capteurs, ainsi que la prise en compte, elle aussi implicite, de la diffraction du corps et du support de l'antenne. Enfin, le temps nécessaire à l'estimation de la position d'une source, est très inférieur à celui des méthodes classiques, ce qui permet d'envisager une détection en temps réel de la position d'une source, en deux ou trois dimensions.

Mots-clés : DOA 2D et 3D, Deep Learning, Convolution à trous, Base de données, Méthode des sources images, Retards fractionnaires, Spatialisation ambisonique, Antennes compactes

Abstract

This PhD thesis proposes to explore a new paradigm of supervised learning for the localization of acoustic sources. On contrary to more conventional methods developed by the scientific community for several decades, this approach does not require the explicit use of the acoustic wave propagation model, nor of underlying signal models. On the contrary, we propose a Deep Learning architecture in order to extract the relevant information for the localization of acoustic sources, directly from a set of raw temporal microphone signals.

To be efficient, this approach, called BeamLearning, requires the constitution of consistent and realistic data sets, in order to train the associated deep neural network. These data sets can be constituted in two complementary ways. The first one is based on the exploitation of realistic numerical simulations of acoustic propagation, using in particular a formalism of image sources to achieve a multi-channel auralization of signals emitted by sources to be located in a reverberant environment. The second one, which enables experimental data sets to be constituted in a reproducible and efficient manner, is based on the use of the 3D spatializer available at the LMSSC, which allows the automated capture, for several hours, of spatialized sources with any microphone array topology – provided that it is compact enough to occupy the ”sweet spot” zone of ambisonic synthesis.

These datasets allow then to optimize the learning variables of a deep neural network, which has been specifically developed for the source localization task during this PhD thesis. This deep neural network is based in particular on separable depthwise atrous convolutional layers. The deep neural network has been tailored with a strong parallel in mind between the common cells used in Deep Learning applications, and the signal processing operations used by conventional acoustic source localization algorithms based on the use of sound wave propagation models.

The BeamLearning approach is a method that offers various advantages over conventional methods. First of all, through numerous situations tested, the observed results show that its localization performances are at least equivalent to the performances of proven source localization algorithms and even largely exceed their capabilities in reverberant and noisy environments. Moreover, the possibility of training the neural network on signals picked up from a real microphone array allows an implicit calibration of the sensors, as well as taking into account, implicitly the diffraction of the body and the support of the antenna. Finally, the time required to estimate the position of a source is much less than that of conventional methods, making it possible to encounter real-time detection of the position of a source in two or three dimensions.

Keywords : 2D and 3D DOA, Deep Learning, atrous convolutions, Sound source Datasets, Image source method, Fractional delays, Ambisonic spatialization, Compact microphone arrays

Zusammenfassung

Diese Doktorarbeit schlägt vor, ein neues Paradigma für die Lokalisierung von akustischen Quellen zu erforschen : das überwachte Lernen. Im Gegensatz zu konventionelleren Methoden, die von der wissenschaftlichen Gemeinschaft über mehrere Jahrzehnte hinweg entwickelt wurden, erfordert dieser Ansatz nicht unbedingt die explizite Verwendung des Ausbreitungsmodells für akustische Wellen oder der zugrunde liegenden Signalmodelle. Im Gegenteil, hier wird eine Deep Learning-Architektur vorgeschlagen, um die für die Lokalisierung von akustischen Quellen relevanten Informationen direkt aus einer Reihe von rohen zeitlichen Mikrofonsignalen zu extrahieren.

Um effizient zu sein, erfordert dieser als BeamLearning bezeichnete Ansatz die Erstellung konsistenter und realistischer Datensätze, um das zugehörige tiefe neuronale Netz zu trainieren. Diese Datensätze können auf zwei komplementäre Arten zusammengestellt werden. Die erste basiert auf der Nutzung realistischer numerischer Simulationen der akustischen Ausbreitung, wobei insbesondere ein Formalismus von Spiegelschallquellen verwendet wird, um eine mehrkanalige Auralisierung von Signalen durchzuführen, die von Quellen ausgesendet werden, die sich in einer halligen Umgebung befinden. Der zweite, der es erlaubt, experimentelle Datensätze auf reproduzierbare und effiziente Weise zusammenzustellen, basiert auf der Verwendung des am LMSSC verfügbaren ambisonischen 3D-Spacizer hoher Ordnung und erlaubt die automatische Erfassung, über mehrere Stunden, von spatialisierten Quellen mit beliebiger Antennentopologie, solange er kompakt genug ist, um den *Sweet Spot*-Bereich der ambisonischen Synthese zu besetzen.

Diese Datensätze erlauben es dann, die Lernvariablen eines tiefen neuronalen Netzes zu optimieren, das speziell für die Aufgabe der Quellenlokalisierung in dieser Arbeit entwickelt wurde und insbesondere auf neuronalen Schichten von tief trennbaren *atrous convolution* basiert. Die Architektur dieses

originellen tiefen neuronalen Netzes basiert auf einer Parallele zwischen den Operationen, die sich aus der Welt des Deep Learning ergeben, und den Signalverarbeitungsoperationen, die üblicherweise von herkömmlichen Algorithmen zur Lokalisierung akustischer Quellen verwendet werden, die auf der Verwendung von Schallwellenausbreitungsmodellen basieren.

Der BeamLearning-Ansatz schließlich ist eine Methode, die gegenüber konventionellen Methoden verschiedene Vorteile bietet. Zunächst einmal zeigen die beobachteten Ergebnisse durch die zahlreichen getesteten Situationen, dass seine Lokalisierungsleistungen den Leistungen bewährter Algorithmen zur Quellenlokalisierung mindestens entsprechen und diese in einer halligen und geräuschvollen Umgebung sogar weit übertreffen. Darüber hinaus ermöglicht die Möglichkeit, das neuronale Netz an Signalen zu trainieren, die von einer realen Mikrofonarray aufgenommen werden, eine implizite Kalibrierung der Sensoren sowie die implizite Berücksichtigung der Beugung des Körpers und der Antennenhalterung. Schließlich ist die Zeit, die für die Schätzung der Position einer Quelle benötigt wird, viel kürzer als bei herkömmlichen Methoden, so dass eine Echtzeit-Erfassung der Position einer Quelle in zwei oder drei Dimensionen möglich ist.

Schlüsselwörter : 2D- und 3D-DOA, Deep Learning, atrous convolution, Spiegelschallquellenmethode, Fractional Delays, Ambisonic spatialization, Kompakte Mikrofonarrays

Table des matières

Liste des tableaux

Liste des figures

Introduction

Depuis trois décennies, l'essor technologique des antennes microphoniques a accompagné le développement de nombreux algorithmes de traitement du signal acoustique. Le domaine a désormais acquis une telle maturité que ce type de dispositif est maintenant utilisé dans un grand nombre de systèmes industriels et commerciaux dédiés à des tâches de localisation de sources, de suivi de sources en mouvement ou de prise de son, et sont même accessibles au grand public. Une grande partie des travaux de recherches actuels en acoustique fait usage d'antennes microphoniques de topologies diverses, pour réaliser des tâches aussi variées que de l'inspection structurale, de la localisation de sources, du débruitage, de la séparation de sources, de la déréverbération ou de la captation spatialisée. C'est le cœur applicatif des travaux réalisés depuis plus de 20 ans par les chercheurs en acoustique du LMSSC.

Les algorithmes développés par la communauté scientifique atteignent aujourd'hui une maturité importante. Ils sont pour la plupart basés sur un modèle physique du milieu de propagation, ou sur des hypothèses statistiques sur les sources et le signal sonore qu'elles émettent. Le problème majeur auquel la communauté scientifique se heurte concerne la robustesse et la précision de ces méthodes dès que le milieu de mesure est mal connu, ou que les sources et les signaux captés s'écartent des hypothèses posées pour la résolution des problèmes inverses.

Pour pallier ce problème, cette thèse de doctorat vise à développer et proposer de nouvelles techniques reposant sur l'utilisation de méthodes de Deep Learning pour traiter le problème de la localisation de sources sonores. L'un des objectifs visés consiste à minimiser les hypothèses sur les informations pertinentes à extraire des données mesurées par les capteurs de l'antenne microphonique, mais aussi sur le modèle de propagation et l'environnement de mesure.

Ces tâches seront confiées à un réseau de neurones profond développé spécifiquement au cours de cette thèse, qui permettra de proposer des traitements de données massives pour construire un algorithme de localisation évoluant avec l'environnement de mesure, les sources en présence, mais aussi la disposition ou le nombre des capteurs composant l'antenne microphonique. Cette approche émergente permettra ainsi de rendre la localisation de sources plus robuste et moins sensible au modèle sous-jacent, puisque les traitements seront appris et construits par le réseau de neurones afin d'être suffisamment adaptables au milieu de mesure ou aux sources.

Le paradigme d'apprentissage supervisé, en particulier l'approche *BeamLearning* développée, s'ap-

plique particulièrement bien au problème de localisation de sources acoustiques, puisque des modèles de propagation numériques existent. Ces simulations permettent ainsi de concevoir une infinité de situations, et donc d'être en mesure de développer numériquement des jeux de données conséquents et réalistes. Ces bases de données profitent des connaissances théoriques ainsi que de l'expérience de l'équipe d'acoustique du LMSSC, et sont constituées en particulier grâce à un outil de calcul rapide de réponses impulsionnelles d'espaces réverbérants sur processeur graphique conçu spécifiquement pour l'occasion et permettant d'obtenir des bases de données massives et réalistes. De plus, un outil de spatialisation acoustique constitué de 50 haut-parleurs permet de constituer des jeux de données expérimentales de manière automatisée sur des topologies variées d'antennes. Ce spatialisateur synthétise de manière réaliste des champs de pressions parfaitement maîtrisés et correspondant à un très grand nombre de positions de sources grâce au formalisme ambisonique sous-jacent. Cet outil de spatialisation de champs 3D par ambisonie à ordres élevés permet de dépasser les limites inhérentes aux jeux de données constituées de simulations numériques, et d'incorporer à la phase d'apprentissage un auto-étalonnage des capteurs microphoniques composant les antennes intelligentes.

En outre, un effort particulier a été fourni tout au long de ce travail de thèse de doctorat pour interpréter les couches neuronales et les cellules construites au sein du réseau de neurones constitué. Plus spécifiquement, les réseaux de neurones ont été conçus à partir de la mise en parallèle entre les techniques éprouvées dans le domaine du Deep Learning et les méthodes classiques d'imagerie acoustique. Des représentations originales sont donc proposées pour interpréter les mécanismes d'apprentissage et les bons résultats de localisation obtenus grâce à la méthode proposée, que ce soit en 2 dimensions et en 3 dimensions, ainsi qu'en environnement réverbérant.

Enfin, des validations numériques et expérimentales ont été menées pour comparer les performances de l'approche *BeamLearning*, reposant sur des *données*, à des approches de localisation reposant sur une approche de type *modèle*. En particulier, une analyse est menée pour comparer les résultats avec ceux des algorithmes *MUSIC* et *SRP-PHAT*, reconnus pour être efficaces pour résoudre les problèmes de localisation de sources acoustiques en environnement bruité et réverbérant. Les tests menés prouvent qu'*a minima*, les performances de l'approche *BeamLearning* sont équivalentes avec ces deux techniques de référence, et peuvent, dans certaines conditions, offrir une amélioration de plusieurs degrés de pré-

cision. Par exemple, dans le cas de la localisation de signaux vocaux dans une salle réverbérante dont la durée de réverbération vaut 0,5 s, avec un bruit de fond faible ($RSB = 30\ dB$), l'approche *Beam-Learning* offre une résolution de $1,9°$, contre $2,6°$ pour SRP-PHAT et $7,2°$ pour MUSIC. Lorsque du bruit blanc décorrélé est rajouté sur les capteurs pour atteindre un $RSB = 5\ dB$, l'approche *Beam-Learning* reste plus robuste, et atteint une précision de $10,8°$, contre $18,4°$ pour SRP-PHAT et $15,7°$ pour MUSIC.

Cette nouvelle approche de localisation de sources acoustiques est déjà intégrée au projet *Deeplomatics* (ANR), porté par le laboratoire LMSSC. Ce projet propose une approche multimodale et modulaire pour la protection contre l'utilisation illicite de drones aériens. La localisation et le suivi de sources acoustiques peut également trouver des applications pour des systèmes trouvant leur place dans un environnement de vie quotidienne, dans le cas des assistants domestiques, par exemple, où la localisation du locuteur peut permettre une meilleure interprétation des consignes données à haute voix. Une autre application potentielle concerne le cadre de visioconférences, offrant ainsi une méthode permettant à la caméra de s'orienter automatiquement vers les différents protagonistes, lors de leur prise de parole. De manière plus générale, ces travaux trouvent des domaines applicatifs partout où des méthodes de localisation de sources acoustiques sont requis, depuis l'industrie, jusqu'au comptage d'espèces animales en zones protégées. La rapidité de traitement de l'approche proposée permet d'envisager, en parallèle de la localisation, une tâche de reconnaissance de sources acoustiques par intelligence artificielle.

Le présent manuscrit présente donc une synthèse du travail effectué durant cette thèse de doctorat. Il est ordonné en cinq chapitres distincts :

— Chapitre 1 : Le premier chapitre présente la problématique liée à ce travail de thèse : dans le cadre de la localisation acoustique, une zoologie d'algorithmes très importante est disponible dans la littérature. Toutefois, chaque algorithme possède un domaine d'application précis, qui dépend des hypothèses qui ont été faites pour modéliser le problème direct de propagation acoustique ou les modèles de signaux sous-jacents. Or, depuis quelques années, le domaine de du Deep Learning est en plein essor, et des tâches de reconnaissances visuelles ou sonores sont

désormais confiées à une intelligence artificielle intégrée à des dispositifs de plus en plus compacts, avec un succès surpassant parfois les capacités humaines. S'élève alors la question de savoir si ces réseaux de neurones profonds peuvent rencontrer des performances aussi spectaculaires, ou, *a minima*, si l'utilisation du Deep Learning peut offrir un paradigme alternatif pertinent pour des taches de localisation de sources acoustiques.

— Chapitre 2 : Le deuxième chapitre propose une architecture originale de réseau de neurones profonds, qui s'inspire en partie des techniques largement éprouvées par les approches modèles. En faisant un parallèle entre les techniques utilisées classiquement en imagerie acoustique, et les outils disponibles dans le domaine du Deep Learning, l'approche *BeamLearning* propose une architecture principalement constituée de banc de filtres convolutifs à trous, et séparables en profondeur. L'une des particularités de l'approche proposée repose sur le fait qu'elle utilise directement les signaux microphoniques captés par une antenne dans le domaine temporel, sans pré-traitement. Les variables du réseau de neurones profond sont donc optimisées lors de la phase d'entrainement pour extraire les informations pertinentes de ces données brutes multicanales. Enfin, une analyse de ces couches neuronales, vues comme des filtres acoustiques, est menée pour justifier l'utilisation de l'architecture proposée.

— Chapitre 3 : Le troisième chapitre précise la manière dont les jeux de données simulées sont construits pour les besoins de cette thèse. En effet, la qualité des données est un enjeu majeur pour l'entrainement des réseaux de neurones profonds. Différents environnements – en particulier des salles réverbérantes – ont donc été simulées grâce à une méthode de sources images optimisée et adaptée à un calcul sur GPU. Pour assurer une précision temporelle nécessaire à l'utilisation d'antennes compactes, une approche par interpolation de Lagrange est implémentée, afin de prendre en compte des retards inférieurs à la durée d'un échantillon lors de la constitution des réponses impulsionnelles de salles. Pour optimiser les centaines de milliers de coefficients du réseau, un nombre important de données est également nécessaire. C'est la raison pour laquelle les jeux de données sont calculés sur carte graphique, offrant ainsi une parallélisation efficace pour simuler la propagation issue de dizaines ou de centaines de milliers de sources dans une salle réverbérante, en quelques heures seulement.

— Chapitre 4 : Le quatrième chapitre expose comment le dispositif de synthèse physique par ambisonie d'ordres élevés du laboratoire LMSSC est utilisé pour constituer des jeux de données expérimentaux, et ainsi prendre en compte les réponses en fréquences non idéales des capteurs, ainsi que tous les phénomènes de diffraction par le corps de l'antenne et de son environnement proche. Si les réponses en fréquence réelles sont parfois prises en compte dans les approches modèles au prix d'une procédure longue et fastidieuse d'étalonnage individuel des capteurs, la diffraction par le corps de l'antenne est quant à elle plus souvent négligée, faute de modèle analytique pour certaines géométries de structures d'antennes. Or, la constitution d'un jeu de données expérimentales proposée dans ce manuscrit permet d'inclure dans les données toutes ces caractéristiques physiques de l'antenne utilisée, et permet alors un étalonnage implicite des capteurs, lors de la phase d'apprentissage.

— Chapitre 5 : Le dernier chapitre résume les principaux résultats obtenus grâce à l'approche *BeamLearning*, en présentant des situations de complexité croissante, depuis une approche de classification par secteurs angulaires en deux dimensions, jusqu'au problème de régression angulaire en trois dimensions dans une pièce réverbérante, avec un bruit de fond important. Pour chacune des situations, des entraînements du réseau reposant sur des données simulées numériquement ou des données expérimentales sont proposés. Les performances offertes par l'approche *BeamLearning* développée dans cette thèse sont ensuite comparées aux algorithmes MUSIC et SRP-PHAT dans différentes situations représentatives de cas d'usages réalistes. Ce dernier chapitre prouve que lorsque la base de données est correctement constituée, l'approche *BeamLearning* offre une précision de localisation supérieure à ces algorithmes reposant sur des modèles. Enfin, le temps de calcul nécessaire à l'estimation de la position angulaire d'une source acoustique (DOA) est suffisamment court pour que le temps d'enregistrement d'une trame temporelle soit supérieur à son traitement, ce qui suggère que cette approche puisse être aisément exploitée en temps réel pour le suivi de sources mobiles.

Chapitre 1

État de l'art : de l'approche modèle à l'approche apprentissage supervisé

L'intelligence est avant tout une faculté de synthèse d'ordre et d'unité. [...] L'intelligence passe de la considération des êtres et des objets particuliers à celle de leurs qualités communes, de leurs rapports permanents. (*Petit Larousse illustré : nouveau dictionnaire encyclopédique*, 1906)

Contenu du chapitre

1.1 Localisation de sources acoustiques

La localisation de sources acoustiques est un domaine dont les applications sont extrêmement vastes. Elles vont de l'industrie automobile et du transport, aux systèmes de surveillance, en passant par la vie quotidienne où les assistants vocaux domestiques utilisent l'information de position du locuteur pour filtrer les consignes captées et mieux les traiter.

Les algorithmes utilisés classiquement ont été développés à partir de modèles de propagation acoustique ou des modèles de signaux, et reposent sur des hypothèses la plupart du temps fortes ou simplificatrices. Or, ces hypothèses ne sont pas toujours vérifiées lorsque les algorithmes sont utilisés, ce qui peut mener à une dégradation des performances de ces méthodes lorsqu'elles sont utilisées dans une situation ne correspondant pas au cadre pour lequel elles ont été développées. De plus, face à l'étendue des algorithmes disponibles dans la littérature, il peut être parfois difficile de choisir celui qui sera le plus approprié au domaine applicatif voulu et à la situation de mesure *in situ*.

C'est pourquoi ce travail de recherche propose d'explorer l'utilisation des techniques de Deep Learning, désormais largement répandues dans le domaine de l'apprentissage supervisé, en particulier de la reconnaissance d'image, pour construire un réseau de neurones spécifiquement pensé pour résoudre le problème de localisation de sources acoustiques.

Ainsi, ce chapitre présente un état de l'art succinct des méthodes communément utilisées pour la localisation de sources acoustiques, ainsi qu'une rapide vue d'ensemble des techniques d'apprentissage supervisé, afin de mieux situer le cadre dans lequel se situe ce travail.

1.1.1 Système de coordonnées

Dans l'ensemble de ce document, le système de coordonnées sera le système sphérique présenté dans cette section. Les coordonnées du point M sont données par les variables réelles $(r;\ \theta;\ \phi)$. Ces trois variables correspondent respectivement au rayon (m), à l'angle azimutal repéré par rapport à l'axe x et à l'angle d'élévation (exprimés en degré ou en radian). La figure 1.1 reprend ces notations, et le passage depuis les coordonnées sphériques vers les coordonnées cartésiennes s'exprime sous la forme suivante :

$$\begin{cases} x = r \cdot \cos(\theta)\cos(\phi) \\ y = r \cdot \sin(\theta)\cos(\phi) \\ z = r \cdot \sin(\phi) \end{cases} \tag{1.1}$$

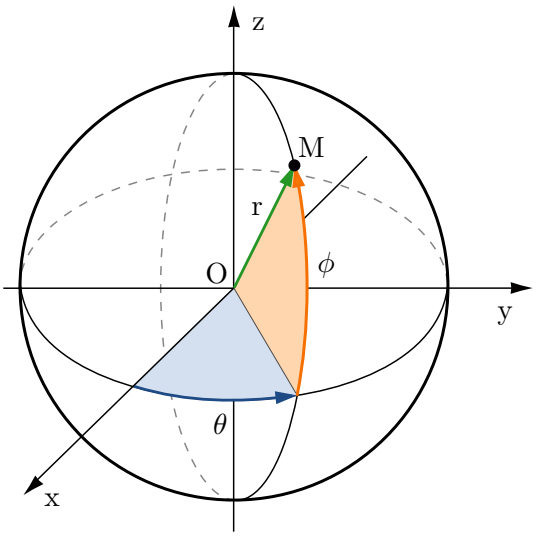


FIGURE 1.1 – Système de coordonnées sphériques

1.1.2 Modélisation du processus de propagation acoustique

Soit un signal $s(t)$ émis par une source en $\vec{r}_s$ à une distance d d'un microphone positionné en $\vec{r}_m$. Si la présence du microphone ne perturbe pas le champ de pression mesuré, alors celui-ci peut s'exprimer, au point de captation, à l'aide de la fonction de Green caractérisant la propagation acoustique entre la source et le capteur dans le milieu de mesure. Si cette fonction solution de l'équation des ondes tridimensionnelle est connue, on peut modéliser la pression acoustique au niveau du microphone comme le produit de convolution entre le signal et cette fonction de Green $G(\vec{r}_m, \vec{r}_s, t)$:

$$p(t) = s(t) \circledast G(\vec{r}_m, \vec{r}_s, t) \tag{1.2}$$

Dans le cas de la propagation en trois dimensions et en champ libre provenant d'une source monopolaire, la fonction de Green $G(\vec{r_m}, \vec{r_s}, t)$ prend la forme simple suivante : :

$$G(\vec{r}_m, \vec{r}_s, t) = \frac{\delta(t - |\vec{r}_m - \vec{r}_s|/c_0)}{4\pi|\vec{r}_m - \vec{r}_s|} \tag{1.3}$$

Cette fonction, relativement simple dans le cas de la propagation en champ libre peut devenir beaucoup plus compliquée dans le cas où des parois réfléchissantes sont prises en compte [1], ou dans le cas de la prise en compte d'un écoulement par exemple [2], mais le formalisme de Green permet de formuler le problème de manière identique, puisque c'est cette fonction qui contient l'information sur le milieu de propagation. Par ailleurs, si plusieurs sources émettent simultanément et que plusieurs microphones

sont utilisés, chaque couple de source-microphone doit être pris en compte. Ainsi, plusieurs fonctions de Green $G(\vec{r}_m, \vec{r}_s, t)$ rentrent en ligne de compte, où les indices s et m représentent respectivement les sources et le microphone en question :

$$p_m(t) = \sum_s s_s(t) \circledast G(\vec{r}_m, \vec{r}_s, t) \tag{1.4}$$

1.1.3 Formation de voies : principe et limitations

Si sur une antenne de M capteurs, on suppose que les signaux captés par les microphones ne diffèrent qu'en amplitude et en phase pour une propagation en champ libre, alors dans ce cas, on peut définir la pression acoustique au niveau de chaque microphone par rapport à un microphone de référence, sous la forme :

$$p_m(t) = \alpha_m s(t - \tau_0 - \delta_m \tau) + b_m(t) \tag{1.5}$$

avec α_m un terme modélisant l'atténuation en amplitude du signal, τ_0 le retard entre le signal de la source et du microphone de référence, $\delta_m \tau$ le décalage temporel introduit entre le microphone m et le microphone de référence de l'antenne et $b_m(t)$ un éventuel bruit de mesure.

Pour toutes les méthodes dérivées de la formation de voies, le principe repose sur un filtrage des signaux microphoniques, suivi d'une somme pondérée de ces voies filtrées. La manière de filtrer et les coefficients de pondération choisis sont sélectionnés en fonction de la géométrie de l'antenne, et de l'objectif visé (minimisation du bruit, minimisation de l'erreur, directivité constante, suppression d'une source dans une direction ...). L'approche la plus simple de cette famille de méthodes est la formation de voies de type "delay and sum" [3,4]. Elle consiste en deux étapes : premièrement, compenser les retards relatifs entre les microphones et le microphone de référence pour une direction donnée, et deuxièmement, sommer tous les signaux ainsi décallés dans le temps. En posant $p'_m(t) = p_m(t + \delta_m \tau)$, et en négligeant les bruit de mesure, on obtient donc la sortie de la formation de voie $F(t)$:

$$F(t) = \frac{1}{M} \sum_1^M p'_m(t) = \alpha_{tot} s(t - \tau_0) \tag{1.6}$$

Un moyen de quantifier la qualité de la localisation de source est d'examiner la figure de reconstruction spatiale de la formation de voies ou *beam pattern* en anglais. Cette figure permet de caractériser pleinement la relation entrée-sortie du système. L'équation 1.6 peut être vue comme un filtre spatial à M points [4]. Cette relation s'exprime simplement dans le cas d'une antenne linéaire, dont les microphones sont régulièrements espacés (figure 1.2). Pour cette situation, comme l'angle d'incidence de l'onde plane est le même pour tous les microphones, le retard relatif entre le microphone 1 et m est :

$$\delta_m \tau = (m-1)\tau_0 = (m-1)d\cos(\theta)/c_0 \tag{1.7}$$

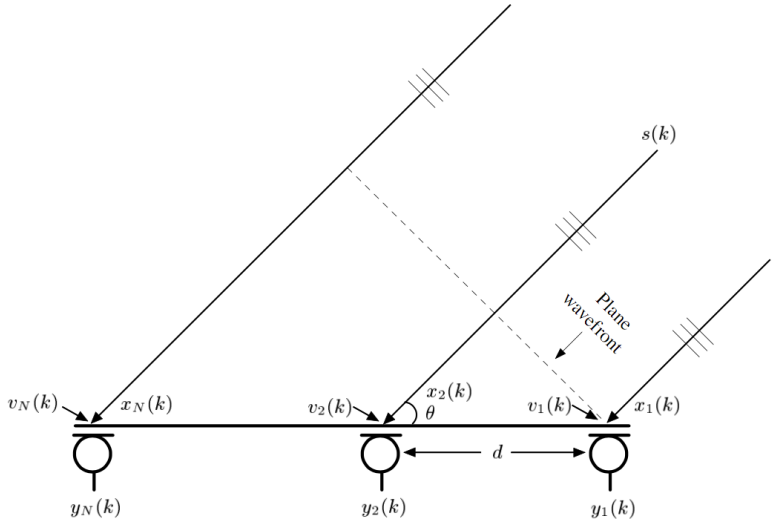


FIGURE 1.2 – Schéma de la captation d'une onde plane en champ lointain par une antenne linéaire

Dans le cas d'un signal monochromatique de pulsation ω, la directivité de la formation de voies pointant dans la direction perpendiculaire à l'axe principal de l'antenne peut être exprimée comme la transformée de Fourier spatiale du filtre [4] :

$$S(\theta, \psi) = \frac{1}{M} \sum_{1}^{M} e^{-j\omega(m-1)d(\cos(\psi)-cos(\theta))/c_0} \tag{1.8}$$

avec $\psi \in [0, \pi]$ l'angle d'observation et $\theta \in [0, \pi]$ l'angle pointant dans la direction de la position réelle

de la source. L'amplitude de la réponse spatiale du filtre est donc la norme de $S(\theta, \psi)$:

$$|S(\theta,\psi)| = \left| \frac{\sin\left(\frac{M\omega d(\cos(\psi)-cos(\theta))}{2c_0}\right)}{M\sin\left(\frac{\omega d(\cos(\psi)-cos(\theta))}{2c_0}\right)} \right| \tag{1.9}$$

Cette réponse est caractéristique de la formation de voie et présente des lobes secondaires d'amplitude non négligeable. La résolution spatiale de ces méthodes est donc limitée par le nombre de capteurs et leur répartition spatiale dans le champ. Par ailleurs, un grand nombre de méthodes de la famille de la formation de voies supposent une propagation en espace libre, afin de simplifier les traitements réalisés sur les voies microphoniques. En milieu réverbérant, cette hypothèse forte mène la plupart du temps à la localisation de sources images ou au déplacement du maximum de localisation. Aussi, des méthodes de localisation à haute résolution ou de décomposition en sous-espaces ont été développées, afin de rendre plus robuste ces approches lorsque l'environnement de mesure est plus complexe.

1.1.4 Méthodes haute résolution

Pour dépasser les limitations inhérentes à la formation de voies, des algorithmes dits à haute résolution ont été proposés par la communauté scientifique. L'objectif est de reformuler le problème en y rajoutant des hypothèses afin d'obtenir une meilleure résolution spatiale. Voici quelques exemples, parmi les plus connu :

— Capon (1969) : La méthode Capon, du nom de son auteur, est fondée sur la recherche d'un vecteur de pondération pour garantir un gain constant de l'antenne dans une direction précise. Cette méthode nécessite de déterminer une matrice de covariance bien conditionnée des signaux mesurés, ce qui suppose que les sources soient stationnaires et fixes. De plus, cette méthode est connue pour avoir des problèmes dans la détection de sources corrélées, noyées dans un bruit de fond parasite.

— MUSIC (1986) [5] : MUltiple SIgnal Classification, est une méthode qui agit à partir de la matrice de covariance. L'idée principale de cette méthode est de séparer les données mesurées en un sous-espace bruit et d'un espace signal, grâce à une extraction des valeurs propres de

la matrice de covariance. Comme la méthode Capon, MUSIC a besoin de la connaissance de cette matrice. Or, comme précédemment, le calcul de cette matrice nécessite un grand nombre de mesures identiques pour être correctement définie. De plus, pour fonctionner, le nombre de sources à retrouver doit être défini à l'avance afin de connaître la taille de l'espace source. Enfin, la méthode utilisant des données fréquentielles, les sources doivent être stationnaires pour être localisées correctement.

— ESPRIT (1989 [6]) : Estimation of Signal Parameters via Rotational Invariant Technique, est un algorithme qui dérive de MUSIC. L'objectif affiché est de réduire les coûts computationnels et de stockage. Pour ce faire, l'antenne utilisée pour capter les signaux sonores doit avoir un invariant spatial. Par exemple les capteurs doivent être appairés deux à deux, et la distance entre chaque doublet de microphones identiques doit être la même pour tous les doublets. Cette hypothèse est donc très forte sur la géométrie de l'antenne, mais permet de gagner en rapidité de calcul.

— CLEAN (1974) : CLEAN est une méthode de post traitement de déconvolution issu de la radioastronomie. Après avoir capté le champ avec la formation de voie, par exemple, les positions des sources sont déterminées de manière itérative et le maximum de niveau de l'image est défini comme la source principale. La réponse du système d'imagerie à la source principale est ensuite soustraite à l'image totale. Ainsi, les lobes secondaires de la source principale sont retirés de l'image. Une deuxième source est ensuite recherchée, et le processus est réitéré. Dans une certaine mesure, on peut voir la méthode CLEAN comme une méthode dite parcimonieuse.

— Méthodes parcimonieuses : Aucune méthode parcimonieuse ne sera ici présentée en particulier, mais plutôt le principe sous-jacent. Comme évoqué plus haut, le problème peut être modélisé comme la recherche d'une solution d'un système y = A.x où y est le vecteur de mesure, x le vecteur d'inconnues représentant les sources, et A la matrice de propagation (appelée dictionnaire). Le principe des méthodes parcimonieuse est de supposer le nombre de composantes non nulles de notre vecteur d'inconnues très faible devant sa dimension. Ce qui revient à supposer le nombre de sources faible par rapport au nombre de positions potentielles. Pour être efficaces,

ces méthodes reposent sur des a priori forts, soit sur le dictionnaire, soit sur le nombre de sources [7, 8], soit sur la géométrie de l'antenne [9].

— SRP-PHAT (2001) [10] : L'approche Steered Response Power - Phase Transform est utilisée spécifiquement pour augmenter la robustesse de la localisation d'un nombre connu de sources dans des environnements réverbérants [11]. Cette approche permet aussi de localiser les sources grâce aux matrices de covariance des signaux. La matrice de covariance croisée généralisée [12] est calculée sur toutes les paires de microphones (SRP). Cette approche nécessite un choix de pondération, qui se porte sur le déphasage entre les microphones, ce qui permet une grande robustesse dans les environnements réverbérants.
— DAMAS (2006) [13] : Deconvolution Approach for the Mapping of Acoustic Sources, ou approche par déconvolution de l'imagerie de sources acoustiques, est une méthode répandue en imagerie haute résolution. Son principe, comme son nom l'indique, repose sur la déconvolution. L'idée est de déconvoluer le résultat de la formation de voie par la réponse de l'antenne. Cette méthode est certes très robuste, mais nécessite un grand nombre d'itérations avant de converger.

Pour resituer les performances de l'approche BeamLearning, les méthodes SRP-PHAT A et MUSIC (ou sa variante dans le domaine ambisonique SH-MUSIC) B seront utilisées dans la suite de ce manuscrit de thèse.

1.1.5 Alternatives à la formation de voies

1.1.5.1 Pression vitesse

D'autres méthodes utilisent, en plus de la mesure de la pression acoustique en un point, celle de la vitesse particulaire. Cette vitesse peut s'obtenir de deux manières différentes. Soit en mesurant directement la vitesse particulaire de l'onde grâce à un capteur à fil chaud, qui consiste à mesurer les variations de températures de deux fils chauffé à plus de $200°C$, soit la vitesse est obtenue en approximant le gradient de pression, par exemple par différences finies en espace [14], ce qui donne accès à la vitesse particulaire. Grâce à cette information supplémentaire il est possible de localiser des sources acoustiques sur des antennes particulièrement compactes [15]. Mais la qualité de la méthode reste tributaire de la précision de la détermination de la vitesse particulaire.

1.1.5.2 Retournement temporel

L'équation de propagation des ondes acoustiques reste invariante par renversement du temps. Si le milieu de propagation est non dissipatif, l'équation des ondes est vérifiée pour les ondes divergentes, mais aussi pour les ondes convergentes. En connaissant les fluctuations de pressions sur une surface fermée $\mathcal{S}$, il est donc possible de créer une onde convergente, qui convergera vers le point d'émission initiale. Ceci reste vrai malgré la présence de réflexion, diffractions ou diffusion au sein du milieu délimité par $\mathcal{S}$ [16, 17]. Toute la difficulté est d'échantillonner le champ sur $\mathcal{S}$ par des capteurs, pour obtenir les informations de pression et de vitesse à ces points [18, 19].

1.1.6 L'apprentissage supervisé, une nouvelle possibilité ?

La localisation de sources acoustiques est en réalité un problème souvent résolu inconsciemment par bon nombre d'êtres vivants dont l'homme, depuis la proie jusqu'au prédateur. En effet, depuis notre plus jeune âge, nous avons appris à localiser la direction de provenance d'un son dans n'importe quel environnement, pour nous prévenir des dangers, ou connaître la position d'un locuteur dans une pièce. La plupart des êtres vivants ayant cette capacité n'ont pourtant à leur disposition que deux oreilles. Pour réaliser cette tâche, bien entendu, aucune connaissance des équations régissant la propagation des ondes acoustiques dans le milieu où ils se trouvent n'est requise.

Malgré ce nombre minimal de capteurs, l'homme peut localiser de manière robuste une source dans le plan azimutal avec une précision allant de 1° pour une source sinusoïdale à 1 000 Hz face à lui, jusqu'à 10° si cette source est latérale [20]. En élévation, la précision dépend énormément des caractéristiques spectrales de la source. Dans le cas d'un bruit blanc, la précison peut aller jusqu'à 4°, mais dans le cas de voix inconnue, elle tombe à 17° [21]. Cette capacité a été acquise par l'expérience multisensorielle (visuelle, auditive et proprioceptive), et est donc intimement liée aux indices acoustiques propres à l'auditeur [22]. Plusieurs caractéristiques des signaux entendus sont utilisés pour remonter à la position de la source. Tout d'abord, les différence interaurales de temps (ITD) et d'intensité (ILD) entre ses deux oreilles [23]. Ces indicateurs permettent en majorité de localiser dans le plan, mais laissent des ambiguïtés en élévation et en azimut, appelées cône de confusion. Pour lever ces ambiguïtés, des indices spectraux sont utilisés. En effet, le pavillon de l'oreille étant très asymétrique,

des interférences tantôt constructives, tantôt destructives se créent et modifient le spectre des signaux en haute fréquence. Ils constituent d'ailleurs des indices prédominants pour la localisation en élévation et pour la discrimination avant-arrière [24, 25]. Enfin, des indices dynamiques comme le mouvement relatif de la tête par rapport à la source finissent de lever les ambiguïtés [26]. L'homme apprend donc à interpréter les sons qu'il entend pour localiser la position des sources acoustiques.

Or, depuis quelques années, grâce à l'apprentissage supervisé, les performances des systèmes de vision assistée par ordinateur ont tellement évolué qu'elles surpassent même parfois les capacités humaines sur certaines tâches [27]. Il se pose donc la question de savoir s'il serait possible d'apprendre aussi à un ordinateur à reconnaître la position d'une source acoustique à l'aide de microphones, en lui permettant d'exploiter librement à partir des données brutes, des indices de localisation présents dans les signaux microphoniques.

1.2 L'apprentissage supervisé

Ces dernières années, les progrès en terme de puissance de calculs sont tels, que réaliser des milliers d'opérations simultanément sur un ordinateur est désormais chose courante. En particulier, le développement des librairies de calculs sur processeurs graphiques (GPU) ont permis la parallélisation massive des opérations, fournissant ainsi la puissance de calcul nécessaire aux méthodes d'apprentissage. De plus, le coût de stockage d'une information numérique n'a cessé de baisser, ce qui a permis d'accumuler un quantité gigantesque d'informations. Celles-ci peuvent concerner un grand nombre de domaines depuis la reconnaissance d'image, jusqu'à la captation audio, en passant par toutes les données environnementales (température, pression ...).

Ainsi un grand nombre de bases de données publiques ont été constituées dans ces différents domaines, et peuvent être utilisées en libre accès. On peut citer par exemple dans le domaine de la reconnaissance d'image MS-COCO [28], CIFAR-10 et CIFAR-100 [29] ou ImageNet [30]. Dans le domaine de la reconnaissance de signaux sonores, des bases de données ont également été constituées pour des tâches majoritairement dédiées à la reconnaissance sonore, notamment : Audioset [31], LibriSpeech [32], DCASE [33] et UrbanSound8K [34]. Ces bases de données ont permis de développer

une nouvelle manière d'aborder les problèmes scientifiques, cette fois-ci sous l'angle des données, et non des modèles. L'ensemble des méthodes développées sur ce type d'approche basée sur les données sont le plus souvent classées sous le terme générique d'intelligence artificielle, même si la variété des techniques démontre que celle-ci relève plus du champ disciplinaire que des méthodes à proprement parler.

Nom	Complexité de l'application	Taille de $\mathcal{D}$	Commentaire
Arbres de décision	Faible	Petite	Séquence de décisions binaires
Machines à vecteur de support (SVM)	Moyenne	Moyenne	Séparer l'espace des exemples par des hypers plans et en déduire des classes
Réseaux convolutifs (CNN)	Grande	Grande	Balayage des données par des noyaux de convolution

Tableau 1.1 – Comparaison de différentes familles de modèles d'apprentissage supervisé

De manière générale, l'optimisation statistique du modèle peut être faite de deux manières : de manière supervisée et non supervisée. Dans ce travail, seule l'approche supervisée sera étudiée. Dans ce type d'apprentissage, l'objectif est de trouver une application surjective entre un ensemble d'entrée x et un ensemble de sortie y, étant donné un ensemble d'exemples de couples antécédent-sortie parfaitement connus $\mathcal{D} = (x_i, y_i)_{i=1..N}$ [35]. Les fonctions mathématiques utilisées sont optimisées grâce à une approche statistique. Un sous ensemble de $\mathcal{D}$ est tiré aléatoirement. Une première application est utilisée par l'algorithme sur les antécédents. Les sorties ainsi obtenues, $\tilde{y}$, sont comparées aux sorties attendues y. Puis, l'application est corrigée itérativement afin de minimiser les erreurs entre $\tilde{y}$ et y. La forme de l'application recherchée dépend avant tout du problème à résoudre, mais aussi du nombre d'exemple de $\mathcal{D}$ disponibles. Un tableau récapitulatif des grandes familles d'approches supervisées est donné dans le tableau 1.1. Chacune d'entre elles est présentée succinctement ci-dessous :

— Les arbres de décision : Les arbres de décision reposent sur une vision séquentielle des données. Chaque élément x de la base de données est représenté par un vecteur multidimensionnel $\{x_1; x_2; ...x_n\}$ correspondant à l'ensemble de variables descriptives de l'élément. Chaque nœud interne de l'arbre correspond à un test fait sur une des variables x_i. La fonction de ce nœud sépare l'espace des sorties en deux ou plusieurs sous espaces. L'espace des sorties est donc successivement réduit jusqu'à obtenir la classe de l'antécédent (dans le cas de la classification) ou un réel (dans le cas d'une régression) [36, 37]

— Les machines à vecteurs de support (SVM) : Dans le cas de la classification à deux catégories, l'objectif des SVM est de trouver *la meilleure* surface qui sépare l'espace des sorties en deux catégories. La fonction à optimiser compare donc la position de la sortie par rapport à une surface, qui évolue au fur et à mesure des itérations jusqu'à converger vers la surface optimale selon le critère choisi. [37, 38]

— Les réseaux convolutifs : Dans le cadre de l'apprentissage profond (*deep learning*), les réseaux convolutifs sont une succession de fonctions de convolution, appelées *couches*, entrecoupées de fonctions non linéaires dites d'activation (*activation layers*). Généralement une couche est une convolution multidimensionnelle de dimension d'entrée N_e vers une dimension de sortie N_s parfois différente. Chaque couche peut être vue comme une projection des données dans un autre espace jusqu'à l'arrivée dans l'espace des sorties. [37, 39]

Dans le cadre de cette thèse, l'approche envisagée se veut la plus physique possible : la majorité des opérations mathématiques qui seront utilisées pour développer le modèle choisi devront être explicables aux vues des caractéristiques physiques du problème à résoudre. Une volonté forte est donc affichée quant à l'interprétabilité des résultats qui seront obtenus. Or, la plupart des méthodes de traitement d'antenne sont basées sur un filtrage des signaux microphoniques et leur recombinaison afin d'obtenir la grandeur de sortie. En particulier, dans le domaine temporel, un filtre peut être vu comme le retourné temporel d'un noyau de convolution au sens où il est utilisée pour les réseaux de neurones profonds. L'approche par réseaux convolutifs en apprentissage supervisé est donc physiquement semblable à l'approche par filtre temporel communément utilisé en acoustique.

Dans le cas général, cette approche nécessite une grande diversité d'exemples parfaitement connus pour optimiser les noyaux de convolution utilisés dans le réseau de neurones. Pour le problème de localisation de sources acoustiques qui nous intéresse ici, le problème direct peut être modélisé grâce au formalisme des fonctions de Green, comme présenté en partie 1.1.2. Grâce à ce formalisme, il est tout à fait possible de constituer des jeux de données conséquents et valides physiquement, contrairement à bon nombre de domaines d'application de l'apprentissage supervisé. Dans le domaine de la reconnaissance d'image par exemple, l'une des difficultés rencontrées repose sur les jeux de données disponibles, pour lesquels il est difficile d'envisager l'utilisation d'images de synthèse suffisament réalistes afin d'entrainer un réseau de neurones. De plus, nous disposons au laboratoire LMSSC d'une sphère de haut-parleurs et d'une suite logicielle permettant de réaliser de la spatialisation sonore grâce au formalisme ambisonique qui ont été développés lors de la thèse de Pierre Lecomte [40–44]. Cet outil de spatialisation permettra ainsi de soumettre des antennes microphoniques à un grand nombre de champs contrôlés afin de constituer des jeux de données de grandes dimensions. Les champs de pression 3D générés par le spatialisateur pourront par ailleurs être synthétisés, ou correspondre à une restitution d'enregistrements ambisoniques existants. Cette approche permet ainsi une flexibilité pour l'apprentissage, qui serait inaccessible dans d'autres conditions. Ces enregistrements pouvant être automatisés, nous avons en plus à disposition une base de données mesurées fiable et conséquente.

À mon sens, pour réussir à mettre en œuvre un algorithme de Deep-Learning, il faut travailler sur trois axes interconnectées [45, 46]. Le premier est l'objectif que doit remplir l'algorithme. Dans notre cas de localisation de sources acoustiques, il faut connaître la précision angulaire attendue, le type d'environnement dans lequel sera utilisé l'antenne microphonique, les types de signaux à localiser, *etc.* Le cahier des charges de l'algorithme est primordial car c'est lui qui dictera quelle base de données constituer. En effet, la base de données est le deuxième axe sur lequel travailler. Pour que l'apprentissage des coefficients du réseau de neurones soit optimal, il faut un jeu de données constitué d'un grand nombre d'exemples étiquetés. Il faut donc que ces exemples soient le plus représentatifs possible des situations réelles où sera utilisé l'algorithme. Enfin, le réseau est le troisième axe à développer. La structure du réseau de neurones ainsi que son dimensionnement dépendent grandement de l'utilisation qui en sont fait. Le travail qui a été fourni durant ces trois années a donc été itératif.

L'approche a d'abord été validée en deux dimensions, avec des bases de données simulées pour des signaux monochromatiques, puis étendu progressivement à la localisation à 3 dimensions en environnement réverbérant, qui a nécessité des simulations complexes, et une approche par Deep Learning conçue spécifiquement pour ce problème : le *BeamLearning*.

1.2.1 Utilisation de l'intelligence artificielle en acoustique

Le domaine de l'apprentissage supervisé s'est énormément développé ces dernières années, et a permis de confier des taches comme la vision assistée par ordinateur, la traduction de texte, ou la reconnaissance vocale à des intelligences artificielles de plus en plus performantes. Ces technologies sont suffisamment matures pour pouvoir être proposée dans la vie de tous les jours, et les méthodes associées à l'apprentissage profond motivent une communauté scientifique internationale grandissante, tant sur des aspects théoriques qu'applicatifs.

L'émergence des techniques de Deep Learning en intelligence artificielle a révolutionné la manière d'appréhender les problèmes. Jusqu'à présent, en particulier en acoustique, la manière de résoudre les problèmes consistait avant tout à modéliser grâce à des approches toujours plus précises les phénomènes physiques étudiés. Cette modélisation repose sur des hypothèses simplificatrices, qui s'affinent en même temps que la compréhension du phénomène ou des limites sous-jacentes des modèles développés. Cette approche a ainsi menée à une profusion de méthodes de traitement du signal acoustique et de problèmes inverses, relevant essentiellement d'améliorations itératives d'approches existantes. Au contraire, dans le domaine de l'intelligence artificielle, la compréhension complète de la physique du problème n'est pas indispensable, puisque celle-ci est contenue dans les données qui permettent d'optimiser l'algorithme construit par l'intelligence artificielle au cours de son entrainement. En effet, l'objectif de l'apprentissage profond est d'optimiser un algorithme à partir d'un grand nombre de données pour faire ressortir les informations pertinentes à la résolution du problème. Mais si l'on combine la connaissance *a priori* des phénomène physiques régissant le problème avec la capacité des réseaux de neurones profonds de faire ressortir d'une multitude de données des informations pertinentes, alors les résultats obtenus ne peuvent qu'en être meilleurs.

Un des domaines d'application de l'intelligence artificielle dans l'acoustique est la synthèse ou la

modification d'environnements sonores [47,48]. Les applications peuvent être la modification de l'écho entendu dans un enregistrement pour simuler un changement de lieu dans un jeu vidéo, ou la séparation depuis une piste audio enregistrée sur une ou plusieurs voies des différents instruments de musique qui y jouent [49]. Enfin il est désormais possible, grâce à l'apprentissage supervisé, de synthétiser une voix humaine et lui faire lire un texte de manière totalement naturelle, tant du point de vue de l'intonation, que du point de vue du timbre de la voix. Une architecture de réseau de neurones a offert récemment un bond de performances dans ce domaine : l'architecture Wavenet [50].

Une autre application de l'acoustique qui a su tirer partie du Deep Learning, est la reconnaissance vocale. Tous les assistants domestiques ou les smartphones ont désormais la capacité de comprendre des instructions de leur propriétaire. Ce domaine de recherche est en particulier développé au laboratoire LMSSC, et un réseau de neurones, TimeScaleNet [51], optimisant des filtres biquadratiques, inspirés du traitement du signal numérique et des modèles de perception auditive, y est développé. Une partie de ce réseau de neurone comporte des similitudes avec celui proposé pour l'approche BeamLearning, qui a pour vocation de localiser des sources acoustiques.

La localisation de sources acoustiques est également un domaine qui est, depuis très récemment, étudié à travers le prisme de l'apprentissage supervisé. La section suivante présente succinctement les différentes approches proposées dans la littérature pour ce problème.

1.2.2 Cas de la localisation de sources acoustiques

La localisation de sources acoustiques est un domaine de recherche qui ne s'est approprié que récemment les outils d'intelligences artificielles. Même si on trouve des travaux de recherches sur le sujet dès le début des années 1990 [52], et sporadiquement jusqu'aux années 2015 [53–56] c'est surtout à partir des années 2017, en particulier avec l'essor du Deep Learning, que la communauté scientifique associe localisation de sources et intelligence artificielle. Si la plupart des publications sur le sujet utilisent de l'apprentissage supervisé [57–87], Hu *et al.* ont prouvé que des réseaux peu ou non-supervisés pouvaient aussi être utilisés [88, 89]. Mais comme la majorité des travaux, ce manuscrit présente uniquement le cas d'un apprentissage parfaitement supervisé.

Si le travail proposé se concentre sur de la localisation de sources en milieu aérien, la localisation de sources sous-marines a aussi profité de l'essors de ces techniques [65, 66]. De nombreux parallèles peuvent être tirés entre la localisation dans ces deux milieux, mais les spécificités de propagation de l'onde acoustique dans le milieu marin, en particulier les gradient de salinité, font qu'il n'est plus possible de considérer le milieu homogène. C'est pourquoi seule des sources en milieu aérien seront traitées ici.

L'objectif est dans ce cas de retrouver la position d'une source acoustique, qui peut être un locuteur ou une source quelconque, à partir des seules informations captées par une antenne microphonique. Sans avoir vocation à proposer une revue bibliographiques exhaustive sur ce sujet, cette section présente un aperçu des méthodes proposées récemment dans la littérature scientifique sur ce sujet.

L'intérêt porté par l'utilisation du Deep Learning à la tache de localisation de sources acoustiques s'illustre par exemple par la participation au challenge LOCATA [90], qui regroupe un corpus de données enregistrés par différentes antennes de microphones dans différentes situation (combinaison de cible statique ou mouvantes, seule ou multiples, enregistrées sur des antennes statiques ou en mouvement), d'un auteur proposant un algorithme de réseau de neurones profonds [91]. Cette tendance se confirme entre autre par la publication dans l'édition d'un numéro spéciale sur la localisation de sources acoustiques du journal IEEE en 2019 [92], des travaux de quatre auteurs proposant de travailler eux aussi avec du Deep Learning [62–64, 68]. C'est donc dans cette dynamique que s'est inscrit, depuis 2017, ce travail de thèse de doctorat.

L'objectif de déterminer la position d'une source acoustique est commun à tous les auteurs,mais la manière d'exprimer cette position, quant à elle, diffère. Tout d'abord, le nombre de sources acoustiques recherchées n'est pas forcément le même. La majorité des auteurs se concentre sur la localisation d'une seule source (et ce sera le cas pour le travail présenté ici), quand d'autres proposent au contraire de retrouver la position de plusieurs sources [67–72].

De plus, la manière de rendre compte de la position de la (ou des) sources acoustiques diffère. L'approche la plus commune consiste à voir le problème de localisation comme un problème de classi-

fication, où l'espace est découpé en différentes zones distinctes. L'estimation de la position se fait alors en estimant quelle zone de l'espace contient la source. Ces zones peuvent être des secteurs angulaires, et ainsi représenter des portions d'angles azimutaux [73–76], ou combiner les angle azimutaux et en élévation pour définir des portions de sphère [69]. Mais il est possible d'aller plus loin et d'estimer en plus la distance séparant la source de l'antenne, comme proposé dans [63]. Lorsqu'une précision angulaire importante est nécessaire, l'approche par classification n'est plus suffisante, et une approche par régression, donnant une valeur chiffrée de la position, est nécessaire [77–79]. Dans une étude récente [72], les auteurs proposent d'utiliser les deux approches conjointement pour déterminer plus facilement la position de plusieurs locuteurs dans une pièce, d'abord de manière grossière à l'aide d'une approche de classification, puis plus finement dans la zone déterminée à l'aide d'une approche de régression. En ce qui concerne l'approche du BeamLearning, proposée dans ce manuscrit, les deux approches seront étudiées.

Quelque soit la grandeur de sortie choisie pour estimer la position de la source acoustique, le choix du type de données à fournir au réseau de neurones est primordial, car les seules informations dont disposera le réseau de neurones sont celles qui lui seront fournies en entrée. Suivant les auteurs, les données issues des microphones de l'antenne sont plus ou moins traitées en amont du réseau, dans l'idée de lui fournir la représentation la plus adéquate de la scène sonore. Généralement, les auteurs proposent d'utiliser en entrée du réseau de neurones une représentation dans le domaine de Fourier des signaux captés. Les informations peuvent être alors des informations de phases uniquement [67], d'amplitude et de phase [68, 69] ou de puissance [70, 73, 80]. Une autre représentation, moins souvent utilisée, propose d'utiliser, à l'instar de certaines méthodes reposant sur des modèles de propagation (voir section 1.1.4), une matrice de covariance obtenue à partir de données microphoniques [81–83]. Enfin, certains auteurs proposent de transposer les données temporelles de pression dans le domaine ambisonique [63, 79, 87]. Au contraire, pour ce travail de thèse de doctorat, aucune transformation *a priori* n'est effectuée, comme pour les travaux des auteurs [72, 84–86], et une approche de *joint feature learning* est proposée, pour que la transformation des données de pression vers la représentation la plus adaptée par le réseau de neurones fasse partie intégrante de la phase d'apprentissage. Cette approche, également connue sous le nom de *end-to-end learning*, représente un pan de plus en plus important des recherches dans de nombreux domaines de l'apprentissage profond.

Le travail effectué pour cette thèse de doctorat se démarque des autres travaux contemporains par l'explicabilité de la structure du réseau de neurone profond proposée. Sans rien diminuer des performances de l'approche exposée dans ce manuscrit, le fait de comprendre physiquement l'apport de chaque couche du réseau a permis de choisir et de justifier chaque opération. Les très bonnes performances de l'approche BeamLearning sont en outre validées par un grand nombre d'exemples issus de simulations, mais aussi de mesures expérimentales. Ces nombreux exemples sont obtenus grâce aux méthodes proposées dans les chapitres 3 et 4, qui permettent la constitution d'un nombre d'exemples rarement atteint dans la littérature. En revanche, l'objectif de localisation de sources acoustique ce limite ici à une seule source, sans chercher à l'identifier. Mais les temps de calculs réduits de l'approche, présentés au chapitre 5, laisse la porte ouverte à ces axes de développement traité dans la littérature.

Chapitre 2

BeamLearning, un réseau de neurones modulaire pour la localisation de sources

En essayant continuellement on finit par réussir. Donc : plus ça rate, plus on a de chance que ça marche. (*Les Shadocks*, Jacques Rouxel, 1968)

Contenu du chapitre

Comme expliqué précédemment, un réseau de neurones est une succession d'opérations mathématiques dont les variables d'apprentissage sont optimisées grâce à une procédure d'entrainement du réseau à l'aide d'une base de données d'apprentissage. Tout ce qui se réfère aux bases de données sera présenté ultérieurement. On suppose donc dans tout ce chapitre, que nous disposons d'une base de données constituée, pour se concentrer uniquement sur la description de l'approche par Deep Learning proposée et sur l'architecture du réseau de neurones profond développé dans le cadre de cette thèse de doctorat. Les différentes couches du réseau BeamLearning sont présentées en détail dans ce chapitre.

2.1 Architecture du réseau de neurones profond pour l'approche BeamLearning

2.1.1 Présentation générale

L'un des objectifs principaux visés au cours du développement du réseau de neurones profond avec l'approche BeamLearning est d'éviter tant que faire se peut une approche de type "boîte noire". Le premier critère sur lequel nous avons basé ce développement est donc l'interprétabilité des traitements réalisés et appris par le réseau. Ainsi il est possible de dresser formellement un parallèle entre la plupart [1] des couches du réseau construit et les méthodes traditionnellement utilisées en traitement d'antennes en acoustique.

La figure 2.1 présente schématiquement l'architecture globale du réseau de neurones utilisé, qui est divisé en blocs. Le premier bloc représente les données d'entrées brutes, correspondant aux signaux microphoniques mesurés par l'antenne. Le deuxième bloc correspond à une succession de M bancs de filtres qui permettent de projeter les donnés dans des sous-espaces représentatifs pour le problème de localisation, grâce à des noyaux de convolution à trous qui seront décrits dans la suite du manuscrit. Le troisième bloc sert quant à lui à calculer une pseudo-énergie des canaux de sortie de cette succession de bancs de filtres. Enfin, le dernier bloc permet d'exploiter ces pseudo-énergies, afin d'en déduire la position de la source ayant émis le champ de pression capté par l'antenne microphonique.

Chacun de ces blocs est détaillé plus amplement dans la suite du document. En particulier, les couches neuronales seront décrites précisément en ce qui concerne le nombre de voies, la taille et les

1. en dehors des couches de non-linéarités, inhérentes à l'apprentissage profond

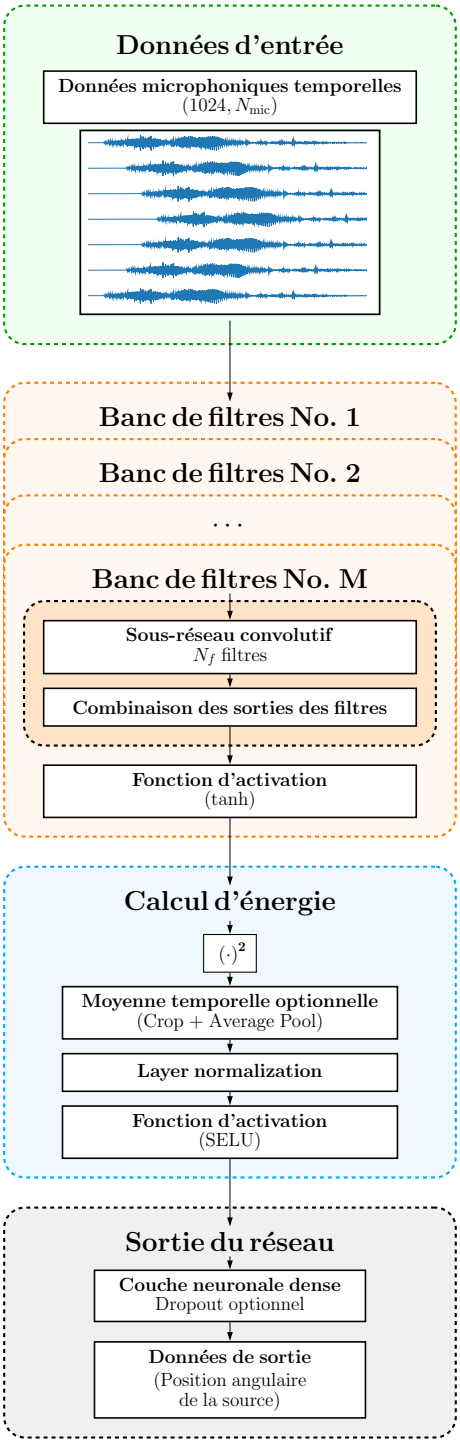


FIGURE 2.1 – Architecture générale du réseau développé pour l'approche BeamLearning

caractéristiques des filtres correspondants. Il est essentiel de noter que le dimensionnement du réseau a été obtenu au cours de cette thèse de doctorat par améliorations itératives, et qu'il n'existe pas à ce

jour de méthode standardisée pour dimensionner un réseau de neurones profond. L'approche utilisée a consisté à l'obtention d'un compromis entre le temps de calcul pour la convergence de l'apprentissage et la précision obtenue en sortie de la détermination de la position de la source.

2.1.2 Données d'entrée du réseau

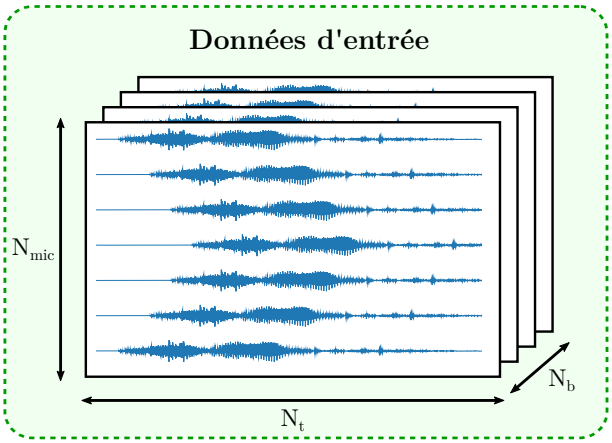


FIGURE 2.2 – Représentation schématique de la matrice de données d'entrée lors de la phase d'apprentissage hors-ligne. Lors d'un apprentissage classique, on a $N_t = 1024, N_{mic} = 7, N_b = 100$.

Les algorithmes de la littérature exploitent généralement comme données d'entrées, des signaux pré-traités, comme par exemple en utilisant soit la covariance des signaux [71, 77, 78, 81–83], ou leur représentation spectrale, soit les informations contenues dans le module, ou/et la phase [62, 67, 69, 79]. Au contraire, ici, nous proposons d'utiliser des signaux temporels bruts. En effet, les différentes convolutions utilisées pour traiter les données à un type de représentation dans le bloc suivant, ont justement pour but de projeter les données temporelles dans un espace le plus approprié au problème posé. Ainsi, on ne contraint pas *a priori* les données en les pré-traitant. Cette approche, communément appelée "Joint Feature Learning", représente un pan de recherche de plus en plus important pour les applications en acoustique du Deep Learning, et repose sur le fait qu'un choix a priori de représentation pour les données d'entrée peut masquer des caractéristiques que pourrait extraire le réseau de neurones par lui-même. En choisissant de conserver les données sans leur adjoindre un pré-traitement ou un changement de représentation, le réseau de neurones profond construit alors une représentation intermédiaire adaptée au problème posé. Par ailleurs, grâce à cette approche sans pré-traitement, les latences entre le moment où le signal est émis, et le moment où la position de la source est retrouvée

sont minimisées et il est possible de rester sur une approche de traitement en temps réel des données. Cette caractéristique est fondamentale pour la plupart des utilisation de la localisation de sources (localisation de locuteur en visio conférence, surveillance de site sensibles...).

Le réseau construit pour l'approche BeamLearning est utilisé dans deux configurations différentes : pour l'apprentissage, et pour l'inférence après "gel" des variables d'apprentissage du réseau. Pour l'inférence, les données d'entrée peuvent être collectées en temps réel ou être pré-enregistrées. Pour l'apprentissage, le jeu de données doit correspondre à un ensemble "entrée-étiquette de référence". Les données sont collectées et étiquetées dans un premier temps, puis utilisées hors ligne.

Comme expliqué dans la section 1.2, dans le cadre de cette thèse de doctorat, les jeux de données utilisés pour l'apprentissage et l'inférence ont été conçus soit par simulation numérique, soit expérimentalement, par la synthèse physique ambisonique. La sphère de spatialisation Spherbedev [42] présentée au chapitre 4 est l'outil de spatialisation permettant d'obtenir des jeux de données expérimentales sur les antennes microphoniques testées. Compte tenu de la dimension de la sphère de spatialisation et de l'ordre de décomposition en harmoniques sphériques, le champ de pression obtenu par synthèse ambisonique est valide dans la bande de fréquences [100, 4000] Hz au centre de la sphère [42] pour un "sweet spot" de l'ordre de 10 centimètres de diamètre correspondant à la taille caractéristique des antennes utilisées dans le cadre de cette thèse. Par conséquent, afin de pouvoir comparer les résultats d'apprentissage issus de données simulées numériquement et de données obtenues expérimentalement grâce à cette synthèse, tous les signaux en entrée du réseau sont systématiquement filtrés à l'aide d'un filtre FIR déterministe passe-bande à phase nulle d'ordre 99 [93]. Ce filtre est implémenté sur GPU, de manière à l'intégrer aux opérations du réseau de neurones. En revanche, le noyau de convolution de ce filtre est gelé et ne fait pas partie des variables d'apprentissage comme le sont ceux des bancs de filtres présents dans les couches suivantes.

Quelque soit l'approche choisie, la dimension de la matrice contenant les valeurs d'entrée est $N_t \times N_{mic}$, comme le montre la figure 2.2. où N_t est la taille du premier indice de la matrice, correspondant à une trame de 1024 échantillons correspondant à une durée de 21 à 23 ms de signal échantillonné (à 44.1 kHz ou 48kHz suivant les expériences). N_{mic}, quant à lui, est la dimension correspondant au

nombre de canaux, c'est à dire au nombre de voies microphoniques de l'antenne. Pour tous les résultats présentés dans cette thèse de doctorat, ce nombre de voies microphoniques varie entre 7 et 19. Comme expliqué précédemment, l'approche BeamLearning proposée est conçue pour être adaptable à la géométrie de l'antenne ainsi qu'au nombre de capteurs microphoniques la composant. Elle reste valide quelque soit la taille de la matrice d'entrée. En revanche, le choix du nombre d'échantillons temporels utilisés représente un compromis afin de correspondre à une durée caractéristique suffisante pour estimer la position de la source, tout en étant adaptée à l'obtention d'un nombre suffisant de périodes de signal en basses fréquences, sans avoir un impact trop important sur l'espace mémoire alloué en RAM GPU par les données traversant le réseau au cours de l'entrainement par lots.

Enfin, lorsque l'approche hors-ligne est utilisée pour l'apprentissage, les matrices d'entrées ne sont pas traitées une par une, mais par lots, ou *(mini)-batchs*. Les calculs sont alors parallélisés sur GPU, et une statistique peut être effectuée pour la rétropropagation des erreurs et l'optimisation des variables d'apprentissage du réseau. Dans notre cas, la dimension du batch N_b vaut 100. Ainsi, au cours de l'apprentissage, la matrice des données d'entrées contient $N_t \times N_{mic} \times N_b \simeq 10^6$ éléments. Dans la suite du manuscrit, par souci de concision, la dimension du batch d'apprentissage N_b sera volontairement omise dans la plupart des descriptions.

2.1.3 Bancs de filtres

Avec l'avènement du Deep Learning, les réseaux de neurones ayant fait leurs preuves reposent pour la plupart sur des successions de couches neuronales effectuant des opérations mathématiques très similaires, opérant sur les représentations obtenues par les couches précédentes. La 2e partie de BeamLearning, est constituée d'une superposition de bancs de filtres (cf. fig 2.1) dont chacun est un ensemble de convolutions séparables en profondeur, entrecoupées de connections résiduelles, de normalisations et de fonction d'activation non linéaires (cf. fig 2.3). Ces opérations permettent de travailler avec des données **temporelles** assimilables à des signaux sonores, et sont présentées et justifiées en détail dans la suite du document.

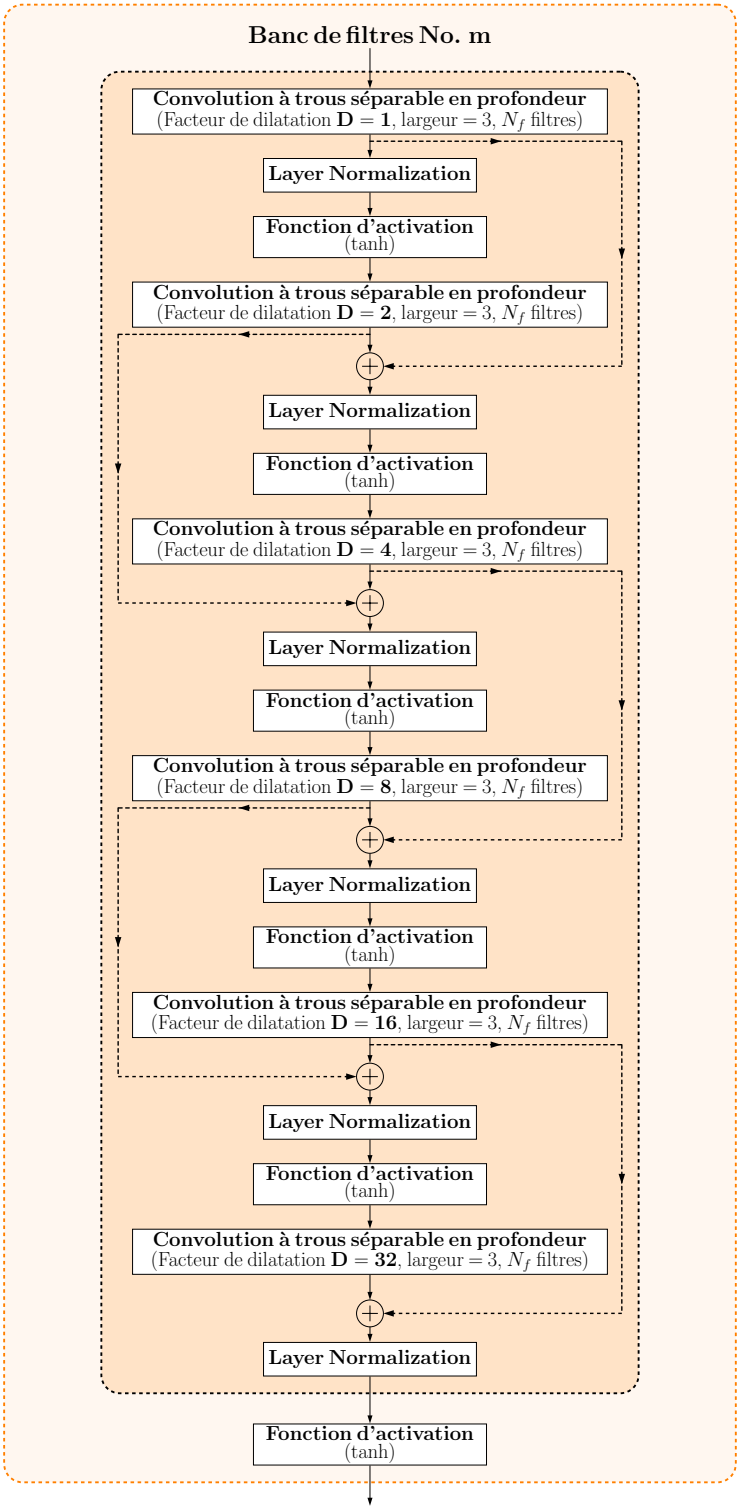


FIGURE 2.3 – Schéma d'un banc de filtre parmi les M bancs du réseau

2.1.3.1 Résumé d'un banc de filtre de BeamLearning

Dans le réseau de neurones présenté, trois bancs de filtres identiques sont mis en cascade comme schématisé en figure 2.1. Ainsi la sortie du banc de filtre 1 (respectivement 2) devient l'entrée du banc de filtre 2 (respectivement 3). Au cours du développement du réseau, l'ajout de bancs de filtres en cascade a permis d'améliorer grandement les performances du réseau. Toutefois, seuls 3 bancs de filtres sont utilisés pour limiter l'espace mémoire alloué lors de l'apprentissage, puisqu'une augmentation supplémentaire du nombre de bancs n'impacte que de manière minime les performances de localisation.

Avant de justifier l'utilisation de chaque opération, le schéma générique d'un banc de filtre est présenté en figure 2.3. Comme indiqué sur la figure 2.1, chacun des M bancs de filtres proposé pour l'approche BeamLearning correspond à une succession de convolutions *pointwise* à trous (voir sec. 2.1.3.4) séparables en profondeur, et de couches convolutives constituées de N_f filtres, avec pour ce travail, $N_f = 128$. Les facteurs de dilatation valent respectivement 1, 2, 4, 8, 16 et 32. Le résultat de chaque convolution est sommé avec le résultat de la convolution précédente grâce à des connections résiduelles (voir sec. 2.1.3.5). Cette somme est ensuite normalisée avec une fonction *layer normalization* (voir sec. 2.1.3.7). Puis, la fonction d'activation tangente hyperbolique est appliquée (voir sec. 2.1.3.6). La sortie non linéaire du banc de filtre m sert ensuite pour l'étape suivante qui peut être le banc de filtre suivant si $m < M$, ou bien un calcul d'énergie dans le cas du dernier banc de filtre ($m = M$).

2.1.3.2 La convolution au sens de l'apprentissage profond

Les cellules neuronales les plus couramment utilisées en Deep Learning pour la vision ou pour l'audio sont basées sur des convolutions (réseaux convolutifs, ou CNN). Lorsqu'il est fait référence au noyau de convolution dans le domaine de l'intelligence artificielle, celui-ci correspond au retourné temporel de la réponse impulsionnelle finie correspondante dans le domaine du traitement du signal. Les données utilisées dans le cas de la vision assistée par ordinateur sont des images. Elles sont donc classiquement exploitées comme des données tridimensionnelles : 2 dimensions dans le plan, et la troisième pour l'encodage des canaux colorimétriques. Les dimensions du noyau de convolution sont donc en général de dimension 2 pour balayer le plan, et sont identiques pour tous les canaux dans le cas de la convolution classique, ou bien différents pour chaque canal dans la cas de la convolution séparable en profondeur. Les noyaux sont aussi bidimensionnels pour la plupart des réseaux ayant pour

données d'entrée des spectrogrammes qui sont généralement traités comme des images dans lesquelles on cherche à reconnaître des motifs. Pour les signaux issus de séries temporelles, les algorithmes de Deep Learning exploitent la plupart du temps des convolutions unidimensionnelles, ou des cellules neuronales récursives de type GRU [94] ou LSTM [95].

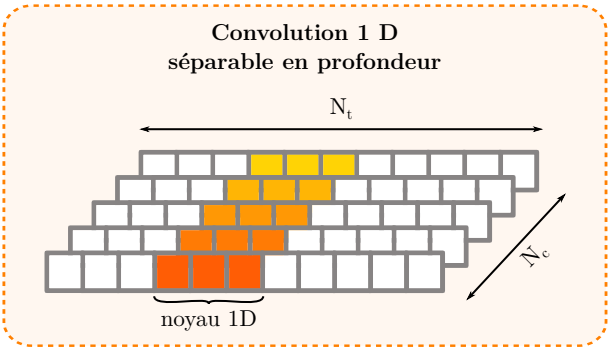


FIGURE 2.4 – Représentation schématique d'une convolution unidirectionnelle séparable en profondeur. Pour l'approche BeamLearning les convolutions sont de largeur 3 avec $N_t = 1024$ et $N_c = 128$.

Dans notre cas, les données temporelles captées ou simulées sont représentées par des tenseurs à seulement deux dimensions : une pour le temps, et une pour les microphones. Plutôt que d'implémenter des convolutions bidimensionnelles avec un seul canal, nous avons fait le choix de convolutions unidimensionnelles avec plusieurs canaux. Les canaux correspondent donc pour la 1$^{\text{ère}}$ couche de convolution aux voies microphoniques. En pratique, les opérations de convolutions sont donc le strict équivalent d'un filtrage des signaux par une succession de réponses impulsionnelles dont les coefficients seraient appris et optimisés pour minimiser une fonction de coût liée au problème inverse posé. Mathématiquement, si on appelle $h_{c,k}$ un noyau appris pour l'opération de convolution sur le canal c et s_c le signal de ce canal, on a alors pour l'échantillon temporel n :

$$y_{c,k}[n] = \sum_{p=0}^{n} h_{c,k}[p] \cdot s_c[p] \tag{2.1}$$

2.1.3.3 Principe de la convolution "pointwise"

Puisque les convolutions utilisées sont séparables en profondeur, les informations d'un canal ne peuvent pas influer sur les résultats provenant des autres canaux. Or, en acoustique, et plus particu-

lièrement dans le domaine de la localisation de sources, la plupart des méthodes de type formation de voies peuvent être exprimées sous la forme *filter and sum*, où l'objectif est de filtrer les signaux microphoniques, puis de les combiner avec des coefficients de pondération adaptés aux contraintes posées (voir 1.1.3). Dans l'optique de créer un réseau de neurones s'inspirant des méthodes classiques, il a été choisi de réaliser une somme pondérée des canaux après filtrage, avec des variables d'apprentissage qui sont, au même titre que les noyaux de convolution, optimisés lors de la phase d'apprentissage. Ainsi, un canal de sortie hérite des informations de tous les canaux d'entrée de la couche précédente [2] . Il est, en outre, possible avec cette méthode d'obtenir un nombre différents de canaux en sortie et en entrée de la couche de convolution, comme présenté en figure 2.5.

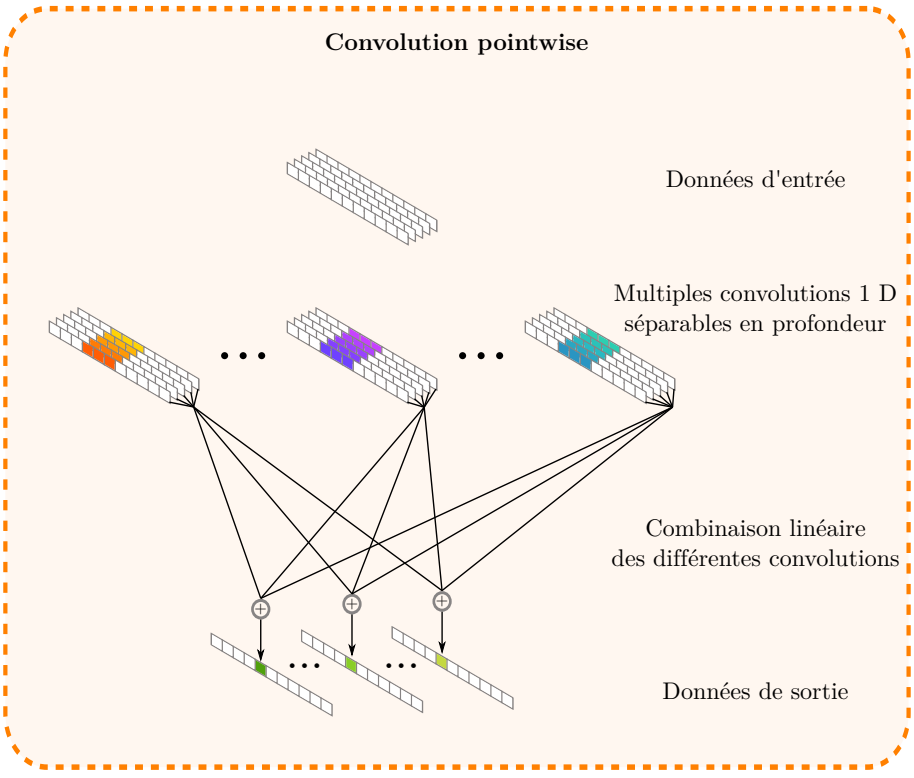


FIGURE 2.5 – Représentation schématique d'une convolution pointwise. Pour l'approche BeamLearning les 128 canaux d'entrée sont chacun filtrés 4 fois de manière différentes, avant d'être recombinés de nouveaux en 128 canaux de sortie.

En notant c l'indice du canal d'entrée et k l'indice du canal de sortie, on obtient donc en sortie la

2. Mathématiquement, ce résultat est équivalent à celui qui aurait été obtenu dans le cas d'une convolution en 2D *temps-canal*. Mais d'un point de vue explicabilité du réseau, une convolution suivie d'une somme paraît plus claire

grandeur $y_{out,k}$, qui correspond au canal numéro k en sortie de la couche, où $a_{c,k}$ est le coefficient de pondération de la $k^{\text{ième}}$ somme des canaux filtrés. On peut donc écrire :

$$y_{out,k} = \sum_{c} a_{c,k} \cdot y_{c,k}$$

Pour BeamLearning, la dimension du noyau de convolution c vaut 3, et est répliquée quatre fois par canal. La dimensions de sortie des canaux k vaut 128. Ainsi, il y a 1 536 variables d'apprentissage différents pour chaque couche de convolutions, et 65 536 coefficients de pondération utilisés pour les combinaisons linéaires.

2.1.3.4 La convolution à trous

Les approches basées sur l'utilisation de méthodes temporelles nécessitent d'exploiter les signaux sur des échelles de temps très variées. En effet, pour l'application qui nous intéresse, les périodes temporelles des signaux varient d'un facteur de 1 à 40, ce qui implique que les noyaux de convolution associés aux filtres appris par le réseau soient adaptés à l'ensemble de ces échelles temporelles. Il est donc primordial de réaliser un compromis sur leur taille, car plus un noyau de convolution sera long et plus il sera en mesure de capter des informations basses fréquences. En revanche, une augmentation sensible des noyaux de convolution mène irrémédiablement à une augmentation du nombre de paramètres à optimiser lors de l'apprentissage, rendant ainsi le réseau plus lourd en terme d'empreinte mémoire, mais également en termes de temps d'apprentissage. Pour donner un ordre d'idée, les noyaux de convolution classiquement utilisés pour des applications de vision assistée par ordinateur excèdent rarement une taille caractéristique de taille de 3×3 pixels, ce qui correspond à seulement 9 coefficients par noyau de convolution. Dans notre cas, une longueur de filtre de 9 coefficients correspondrait à la période d'un signal de fréquence 4 900 Hz échantillonné à 44.1 kHz. En faisant l'hypothèse que la longueur d'un filtre doit être d'au moins 25% de la période du signal temporel filtré pour être efficace et extraire de l'information pertinente,un noyau de convolution présentant une longueur de 9 échantillons ne pourrait pas extraire de l'information efficacement d'un signal en dessous de 1 225 Hz. Pour nos applications, nous nous sommes fixés une gamme fréquentielle utile couvrant la bande fréquentielle de 100 Hz à 4000 Hz. Ainsi, un filtre permettant d'exploiter les fluctuations temporelles d'un signal jusqu'à 100 Hz, aurait un noyau de longueur d'environ 110 échantillons avec une fréquence d'échantillonnage de 44 100 Hz. Mais une telle taille de noyau multiplierait par 10 le nombre de variables d'apprentis-

sage à calculer et impacterait excessivement la mémoire et les performances d'apprentissage du réseau.

Afin de concevoir un réseau sur des données temporelles brutes aux échelles de fluctuations variées, tout en minimisant le nombre de coefficients à apprendre, il est tout à fait possible de spécialiser chaque couche du réseau de neurones pour une échelle temporelle différente. On peut concevoir un réseau avec une première couche convolutive avec des filtres à trois coefficients, puis la couche suivante à cinq coefficients, puis la suivante à 9, etc. Cependant avec cette approche, le réseau de neurones présente *in fine* des filtres avec de nombreuses variables d'apprentissage à optimiser pour les basses fréquences. Pour dépasser cet inconvénient, il est possible d'annuler un certain nombre de coefficients des filtres. Cette approche, initialement proposée par l'approche PixelNet [96] pour des applications de vision assistée par ordinateur et en particulier pour de la segmentation sémantique, consiste à exploiter le principe de convolution *à trous*, ou convolution *dilatée*. Ainsi, il est possible d'utiliser des noyaux de convolution avec peu de coefficients, présentant un large champ réceptif ou taille caractéristique. Chaque dilatation d'un noyau de convolution correspond ainsi à la conception d'un filtre se spécialisant pour des périodes augmentant exponentiellement, tout en minimisant l'impact sur la mémoire et le nombre de coefficients à optimiser dans la phase d'entraînement du réseau de neurones.

La figure 2.6 présente un exemple de réseau avec des couches successives exploitant un noyau de convolution court de seulement 3 coefficients par filtre, avec des facteurs de dilatation successifs augmentant exponentiellement. Pour ces convolutions à trous, on parle de facteur de dilatation pour caractériser l'écart entre le point central du noyau et la prochaine valeur non nulle du noyau. Cette méthode est à mettre en parallèle de la décomposition en ondelettes issue du traitement du signal [97], qui consiste à approcher un signal en superposant une même fonction de référence, dilatée à différentes échelles temporelles. De la même manière, les convolution à trous sont la répétition d'un même schéma convolutif, dilaté à différentes échelles temporelles, afin d'extraire l'information pertinente dans les signaux temporels bruts sur toute la gamme fréquentielle d'intérêt.

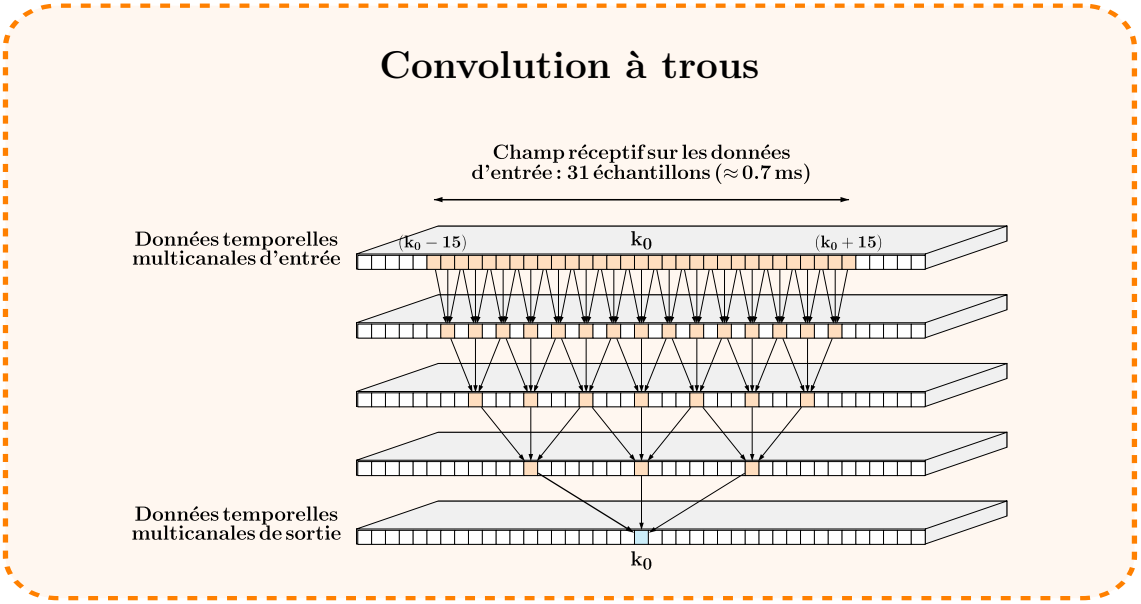


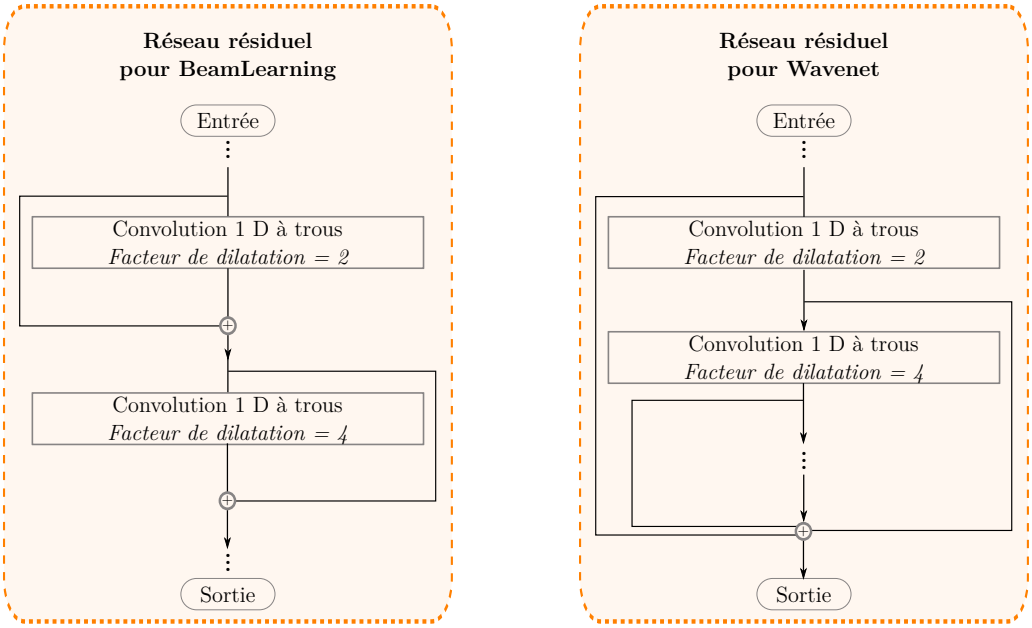
FIGURE 2.6 – Schéma de principe d'une succession de couches convolutives à trous, pour un exemple de facteurs de dilatations successifs égaux à 1,2,4,8. Les flèches représentent les opérations de convolutions, reliant les données d'entrée au données de sortie pour chaque couche. Les données utilisées pour le calcul de la valeur temporelle à l'échantillons k_0 de la couche de sortie sont mis en évidence par les échantillons colorés en orange. La convolution étant séparable en profondeur, chaque canal est filtré indépendamment des autres, pour chaque couche convolutive à trous. Pour l'approche BeamLearning, les noyaux de convolutions sont de largeur 3 points, et les facteurs de dilatation successifs valent 1, 2, 4, 8, 16 et 32.

Dans le domaine du machine learning pour l'audio et l'acoustique, ce type d'approche multi-résolution a été proposée récemment avec succès pour des applications de synthèse vocale avec *Wavenet* [50], de traduction assistée par ordinateur [98], de débruitage de signaux audio [99], ou la reconnaissance de mots [51], et a permis d'obtenir des performances exceptionnelles dans ces domaines.

2.1.3.5 Réseaux résiduels

De manière à favoriser l'émergence de traitements les plus "expressifs" possibles pour l'approche BeamLearning, chaque couche de convolution à trous est complétée par des connexions résiduelles. L'utilisation de ce réseau résiduel permet ainsi aux sorties de chaque couche du sous-réseau de "contourner" la suivante [100]. Ce type de connexions résiduelles a été introduit dans la littérature pour éviter des phénomènes de saturation ou de détérioration de l'apprentissage au cours du calcul direct et de la rétropropagation des gradients pour les réseaux profonds [101, 102]. Ces connexions représentent l'un

des ingrédients clés de certaines méthodes ayant permis de grandes avancées dans le domaine de la reconnaissance d'image [103], comme ResNet [101], Inception-Resnet [104] et ResNext [105].



(a) Présentation du réseau résiduel implémenté pour BeamLearning

(b) Présentation du réseau résiduel implémenté pour Wavenet

FIGURE 2.7 – Représentations schématiques de l'implémentation d'une connexion résiduelle dans les architectures BeamLearning et *Wavenet* [50]

Pour notre application, les connexions résiduelles sont introduites entre chaque couche correspondant à des facteurs de dilatation croissants. On peut ainsi voir cette somme résiduelle comme un moyen d'apporter à chaque couche de convolution des données filtrées correspondant à de l'information extraite des signaux à des échelles temporelles différentes que celle qui est traitée par la couche convolutive suivant la connexion résiduelle. Cette approche a également été proposée pour le réseau WaveNet [50], mais la topologie des connexions résiduelles est légèrement différente de celle proposée pour l'approche BeamLearning. L'équipe de DeepMind ayant proposé le réseau WaveNet ont quant à eux exploité les connexions résiduelles entre sous-réseaux : plutôt que de sommer l'entrée et la sortie de chaque couche de convolution à trous, les convolutions à trous se font en cascade et les sorties des convolutions effectuées sont ensuite sommées pour être traitées par la suite du réseau (voir fig. 2.7(b)). Dans le cadre du développement de l'approche BeamLearning, nous avons testé les deux solutions pour

notre problème de localisation de sources, et ces tests ont révélé qu'il était plus efficace d'appliquer les connexions résiduelles entre chaque facteur de dilatation comme proposé dans [101], plutôt qu'entre chaque sous-réseau, comme proposé par [50].

2.1.3.6 Choix de la fonction d'activation : tanh

Jusqu'à présent, les opérations présentées ci-dessus sont linéaires [3] . Ainsi, les modèles proposés restent linéaires puisque construits à partir de combinaisons d'opérations linéaires. Pour avoir accès à des modèles non-linéaires, il faut donc rajouter des fonctions d'activation non-linéaires, outils clés des mécanismes sous-jacents aux techniques d'apprentissage par réseaux de neurones dès leur apparition avec le perceptron dans les années 50 [106], inspirés des connexions neuronales biologiques. En effet, une non-linéarité bien choisie favorise la rétropropagation des erreurs à travers le réseau, et l'activation sélective des sorties de chaque couche neuronale [39].

Ce rôle est donc joué par des fonctions d'activation, qui sont placées après chaque couche convolutive dans le sous-réseau "banc de filtres". Elles permettent également d'assurer que la sortie de chaque couche convolutive est bien dans le domaine de définition optimal pour la couche convolutive suivante. Elles empêchent ainsi l'augmentation exponentielle de la valeur des données au fur et à mesure des couches, en conservant les données de sortie dans l'intervalle $[-1, 1]$ par exemple.

Enfin, l'approche recherchée pour cette partie du réseau de neurones est toujours une approche *filtre and sum*. Or, en acoustique, les données temporelles sont à moyenne nulle. Afin de conserver cette propriété physique des grandeurs associées au problème, il est donc primordial que la fonction d'activation utilisée dans le réseau de neurones associé à l'approche BeamLearning soit elle aussi centrée sur 0. Parmi les fonctions d'activations communément utilisées en Deep Learning, seules la tangente hyperbolique et la sigmoïde recentrée sur 0 répondent à ce critère. Pour nombre de problèmes traités, la fonction tanh a permis d'obtenir des performances accrues [107]. Cette tendance s'est vérifiée lors de nos tests pour le problème spécifique de localisation de sources acoustiques, et le choix s'est tourné

3. L'opération de normalisation présentée en section suivante (2.1.3.7) ne l'est pas non plus. Mais elle reste cependant une opération linéaire à un changement de variable prêt, qui ne change pas fondamentalement la physique des données. Elle change seulement l'échelle des valeurs.

naturellement vers cette fonction d'activation, pour le sous-réseau décrit dans cette section :

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

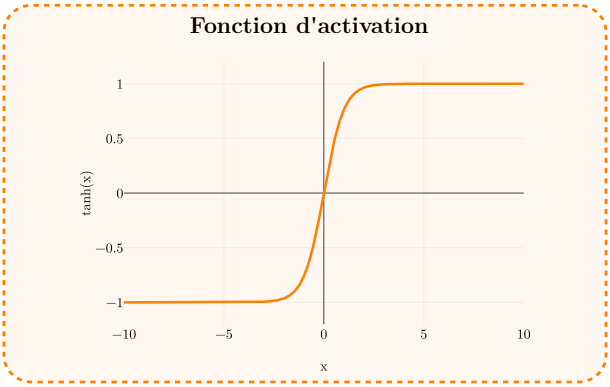


FIGURE 2.8 – Fonction d'activation non linéaire utilisée pour les bancs de filtres dans BeamLearning : Tangente hyperbolique (tanh)

L'inconvénient de la fonction choisie, est qu'elle aplatit fortement les entrées élevées en valeurs absolues (phénomène de seuillage). Pour éviter de perdre en expressivité, il faut donc absolument éviter que les grandeurs manipulées à travers le réseau atteignent des valeurs trop élevées. C'est en partie pour cette raison, qu'il est important de normaliser les données avant chaque non-linéarité.

2.1.3.7 Normalisation des variables d'apprentissage

Une analyse de la valeur des coefficients lors de l'apprentissage a démontré que plus le nombre d'itérations augmentait, plus la valeur des coefficients augmentait. Avoir des noyaux de convolution aux coefficients élevés n'est pas en soit un problème puisque les résultats étaient satisfaisants, mais ces valeurs entraînent une augmentations de la valeur numérique des sorties. Or, comme exposé précédemment, les fonctions d'activations tanh sont utilisées dans le réseau pour aider à la rétropropagation des erreurs grâce aux gradients [39], et apporter de la non linéarité au modèle. Compte tenu de la forme de cette fonction d'activation, une trop forte valeur des grandeurs de sortie des couches convolutive mène à un phénomène de seuillage des signaux filtrés par les différentes couches. Afin d'éviter ce phénomène, nous avons donc exploité une méthode de normalisation en amont des fonctions d'activation.

Les méthodes de normalisation sont traditionnellement utilisées dans le domaine du Deep Learning

pour stabiliser l'apprentissage et réduire le nombre d'itérations nécessaires pour qu'un réseau converge. Par ailleurs, pour le problème de localisation de sources qui nous intéresse ici, l'objectif est également d'obtenir des performances de localisation qui soient peu impactées par le niveau sonore des sources émettrices. Les apprentissages réalisés au cours du développement de l'approche BeamLearning ont révélé que ces stratégies de normalisation apportaient effectivement une robustesse accrue aux variations de niveaux sonores des données d'entrées.

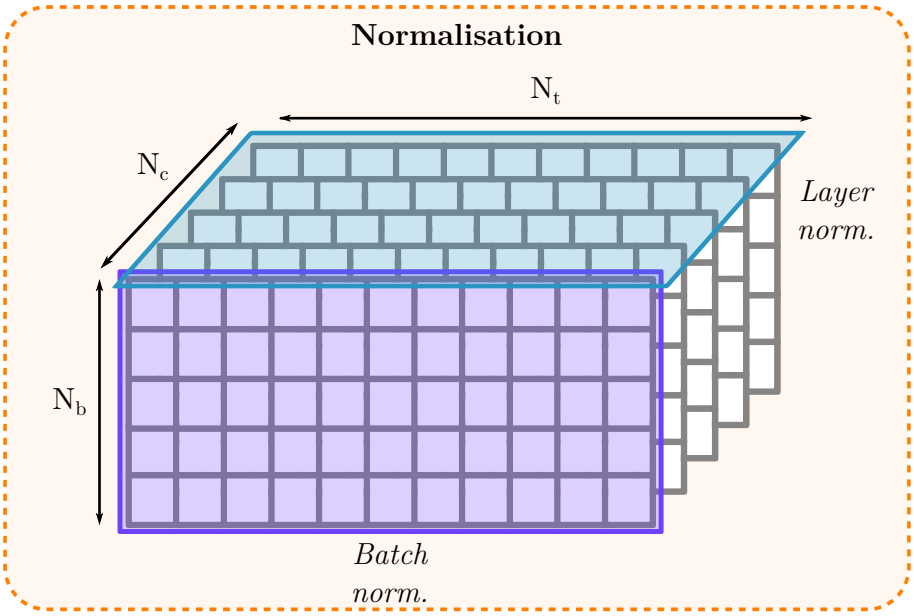


FIGURE 2.9 – Représentation schématique de la normalisation par batch (*batch normalization*) [108] et de la normalisation de la couche (*layer normalization*) [109]. Pour plus de lisibilité, la dimension sur les canaux est omise.

Il existe plusieurs manières d'effectuer un processus de normalisation pour un réseau de neurones profond. Les trois manières les plus communes sont la normalisation des poids (*weight normalization*) [110], la normalisation par batch (*batch normalization*) [108], et la normalisation de la couche (*layer normalization*) [109]. Pour plus de clarté, un schéma est présenté en figure 2.9. Comme son nom l'indique, la normalisation des poids ne modifie pas les entrées de chaque couche de neurones. Or, pour obtenir une robustesse aux variations de niveau sonore d'entrée, il faut agir sur les entrées des couches convolutives du réseau. La différence entre la normalisation par *batch* ou par couche est uniquement la direction selon laquelle la normalisation est effectuée. Nos données d'entrées sont de dimension 3 : une dimension pour le *batch* indicée par $_b$, une dimension pour le temps indicée par $_t$, et

une dimension pour les canaux indicée par $_c$. Dans les deux cas, la normalisation s'effectue en prenant en compte deux directions, dont l'une est la dimension du temps. Dans cette direction, les données n'ont pas besoin d'être recentrées puisque des signaux de pression sont par essence de moyenne nulle. En revanche, changer la dispersion statistique des données permet de s'affranchir du niveau sonore des données. Dans le cas de la normalisation par *batch*, les moments statistiques sont calculés selon la dimension du temps et du batch :

$$\begin{cases} \mu_b = \frac{1}{N_b \cdot N_t} \sum_{b}^{N_b} \sum_{t}^{N_t} x_{b,t,c} \\ \sigma_b = \frac{1}{N_b \cdot N_t} \sum_{b}^{N_b} \sum_{t}^{N_t} (x_{b,t,c} - \mu_b)^2 \\ \tilde{x}_{b,t,c} = \frac{x_{b,t,c} - \mu_b}{\sqrt{\sigma_b^2 + \epsilon}} \end{cases} \tag{2.2}$$

Au contraire, lors de la normalisation de la couche, la deuxième dimension utilisée est celle des canaux. On a alors :

$$\begin{cases} \mu_t = \frac{1}{N_c \cdot N_t} \sum_{c}^{N_c} \sum_{t}^{N_t} x_{b,t,c} \\ \sigma_t = \frac{1}{N_c \cdot N_t} \sum_{c}^{N_c} \sum_{t}^{N_t} (x_{b,t,c} - \mu_t)^2 \\ \tilde{x}_{b,t,c} = \frac{x_{b,t,c} - \mu_t}{\sqrt{\sigma_t^2 + \epsilon}} \end{cases} \tag{2.3}$$

Pour que le BeamLearning soit efficace, il est nécessaire que les données traitées par la partie convolutive du réseau aient une statistique équivalente entre toutes les couches du sous-réseau de bancs de filtres pour ne pas donner plus d'importance à une couche de convolution plutôt qu'à une autre. En outre, pour s'affranchir du niveau sonore des sources à localiser, la normalisation par couche est la plus pertinente, ainsi, cette normalisation est utilisée après chaque couche convolutive. Enfin, du point de vue de l'implémentation dans le réseau, nous avons choisi d'effectuer la normalisation après la connexion résiduelle, comme le montre la figure 2.3. De cette manière, la somme des résultats de

chaque couche est successivement normalisée à chaque étape du réseau et sert d'entrée à la couche suivante, tout en évitant un aplatissement trop important des données par la fonction d'activation présentée en section 2.1.3.6.

2.1.4 Représentation pseudo-énergétique

La plupart des méthodes de traitement d'antennes appliquées à la localisation de sources exploitent une sortie homogène à une grandeur énergétique afin de déterminer la position de la source par recherche de maximas dans les cartes spatiales obtenues en sortie [3]. L'objectif de cette partie du réseau de neurones, proposé pour l'approche BeamLearning, est donc de transformer les données temporelles en sortie du sous-réseau précédent, qui sont homogènes à des grandeurs de pression, en des données quadratiques pseudo-énergétiques (voir fig. 2.10).

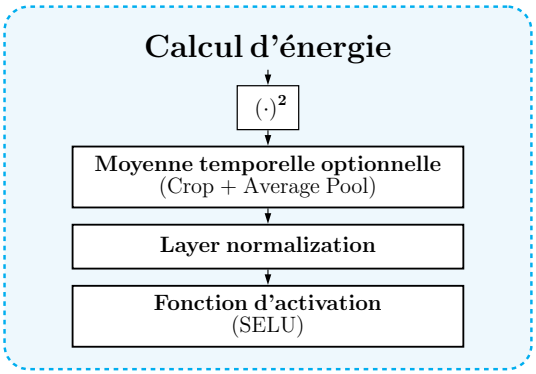


FIGURE 2.10 – Représentation schématique du calcul pseudo-énergétique du réseau pour l'approche BeamLearning.

Contrairement au sous-réseau constitué de bancs de filtres à optimiser par apprentissage, le sous-réseau chargé de la représentation pseudo-énergétique est majoritairement composé d'opérations déterministes, permettant ainsi d'obtenir une grandeur quadratique moyenne, que l'on peut assimiler à une pseudo-énergie dans chacun des *canaux* constitué par les N_f filtres précédents. Ces représentations pseudo-énergétiques sont ensuite recentrées autour d'une moyenne nulle grâce à la fonction de normalisation vue précédemment en section 2.1.3.7. La fonction d'activation SELU [111] présentée en section 2.1.4.2, est utilisée en sortie de cette couche de normalisation afin de favoriser la rétropropagation des gradients pendant la phase d'apprentissage.

2.1.4.1 Calcul d'une moyenne temporelle

L'objectif de cet partie de BeamLearning étant de calculer une énergie, il faut moyenner les signaux après les avoir élevés au carré. Cette moyenne se fait donc sur la dimension temporelle, pour chaque canal. Dans le domaine de l'intelligence artificielle, cette réduction de dimension par moyenne, s'appelle *average pool*, et est réalisée par la statistique moyenne sur un batch.

D'un point de vue algorithmique, les données temporelles ont été filtrées plusieurs fois. Or, les signaux étant finis, la sortie de chaque convolution n'est valide au sens physique que pour $N_t - N_{noy} + 1$ éléments temporels, avec N_t la dimension temporelle d'entrée, et N_{noy} la taille du noyau de convolution, dans le cas où N_{noy} est impair, comme c'est le cas pour les noyaux proposés ici. Or, d'un point de vue algorithmique, pour conserver une architecture modulaire, il est plus facile de garder des sorties aux mêmes dimensions que les entrées, et avant de moyenner, de supprimer les éléments non valides. Dans notre cas, les données temporelles passent donc de 1 024 échantillons par canal, à 832 éléments, compte tenu du nombre de couches utilisées et des facteurs de dilatation choisis pour le sous-réseau de bancs de filtres. Cette opération appelée *crop*, est schématisée dans la figure 2.11. Ainsi, en sortie du sous-réseau appelé *calcul d'énergie* de BeamLearning, une représentation pseudo énergétique est calculée sur chaque canal.

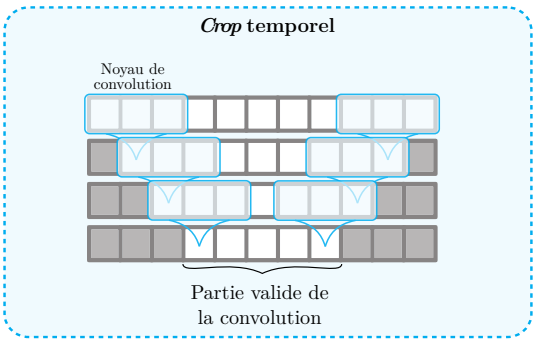


FIGURE 2.11 – Représentation schématique de la sélection valide des convolutions grâce à un *crop*.

2.1.4.2 La fonction d'activation SELU

Dans le domaine de la reconnaissance d'image, les données sont généralement positives. Les fonctions d'activations qui sont utilisées ont donc généralement deux comportements distincts, l'un pour les valeurs positives et l'autre pour les valeurs négatives. En particulier, la fonction ReLU (*Rectified*

Linear Unit) $ReLU(x) = x^+ = max(0, x)$ est commune à de nombreux réseaux [112]. Si cette fonction évite que les gradients soient évanescents (*vanishing gradient*), le fait d'annuler toutes les valeurs négatives de la grandeur d'entrée, peut mener à ce que des parties du réseau de neurones deviennent inexpressives, c'est pourquoi des variantes de cette fonction ont été développées. En particulier, on peut citer les fonctions Leaky ReLU [113] ou le ELU (*Exponential Linear Unit*) [114] ou encore plus récemment le Swich [115]. Dans cette partie du réseau traitant des données énergétiques, donc positives, l'utilisation de la fonction d'activation non linéaire tanh n'est plus pertinente. Le choix de la fonction d'activation s'est porté sur une autre variante de la fonction ReLU : la fonction dite *Self Exponential Linear Unit*, notée SELU. Elle est en particulier utilisée pour ses propriétés de normalisation des poids du réseau [111] et est définie par :

$$SELU(x) = \begin{cases} \lambda \cdot x & \text{si : } x > 0 \\ \lambda \cdot (\alpha \cdot e^x - \alpha) & \text{si : } x \leq 0 \end{cases} \tag{2.4}$$

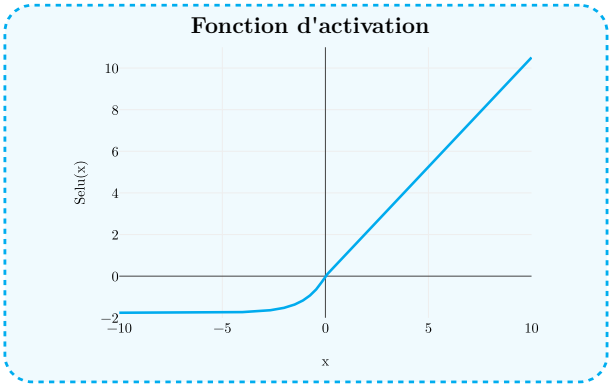


FIGURE 2.12 – Fonction d'activation non linéaire utilisée pour la normalisation du réseau : SELU.

D'après Klambauer *et al.* [111], lorsque $\lambda \approx 1,0507$ et $\alpha \approx 1,67326$, la fonction d'activation permet d'introduire des propriétés de normalisation des coefficients, contrairement à la normalisation des données vue en section 2.1.3.7. De plus, cette fonction d'activation permet à la fois de limiter la croissance trop importante des gradients lors de la rétropropagation des erreurs et d'éviter la disparition de ces gradients. Le fait d'utiliser cette fonction d'activation permet aussi d'accélérer le temps de calcul. Cette accélération permet de gagner jusqu'à un ou deux jours sur un apprentissage d'une dizaine de jours d'après les différents essais effectués pour l'approche BeamLearning.

2.1.5 Sortie du réseau de neurones pour l'approche BeamLearning

En fin de réseau, la succession de couches neuronales denses (*full connected* ou FC) permet de combiner les N_f sorties du sous-réseau en charge du calcul pseudo-énergétique. La sortie finale, qui représente l'estimateur de la position de la source, est obtenue soit par une approche de classification (une classe par zone de l'espace), soit par une approche de régression (position X, Y, Z de la source). La couche de FC permet donc de faire une combinaison linéaire des énergies des canaux, et ainsi d'obtenir N_{out} valeurs. Le nombre de valeurs dépend de la représentation de la sortie du réseau voulu. Schématiquement, dans le cas de la classification à 8 classes, la grandeur de sortie du réseau sera un vecteur à 8 éléments. Dans le cas de la régression la grandeur de sortie du réseau correspondra à un scalaire.

De manière à réaliser la phase d'apprentissage, une fonction de coût permettra de quantifier l'erreur entre la sortie obtenue à l'itération courante de l'apprentissage par rapport à la position de référence de la source. Cette fonction de coût prend donc comme arguments le *label* de la donnée d'entrée et le résultat obtenu en sortie finale du réseau. À partir de cette fonction de coût, les opérations de rétropropagation sont ensuite réalisées, afin de mettre à jour les valeurs numériques des variables d'apprentissage à optimiser dans le réseau pour l'itération suivante, grâce à l'algorithme Adam (voir sec. 2.1.6.2). Le choix de la fonction de coût dépendant de l'approche utilisée (classification ou régression) et du problème (2D ou 3D), il sera discuté ultérieurement en section 2.2.

2.1.6 Optimisations statistiques et temps caractéristiques

Pour la phase d'entraînement, l'ensemble des valeurs des variables d'apprentissage du réseau doivent être initialisées avant d'être optimisées itérativement. L'optimisation des valeurs des variables d'apprentissage se fait grâce aux calculs de gradients de la fonction totale résultant des compositions de toutes les opérations mathématiques constituant les couches du réseau de neurones. Les gradients sont calculés depuis la fonction de coût finale jusqu'aux premières convolutions, en passant par toutes les non linéarités. Cette phase de calcul, appelée rétropropagation, permet de calculer la valeur du gradient pour l'ensemble des éléments d'un lot d'apprentissage, à une itération donnée. Les valeurs de l'ensembles des gradients étant obtenus pour une itération donnée, il s'agit ensuite de mettre à

jour les valeurs des variables d'apprentissage, de manière à converger au plus vite vers un ensemble permettant de minimiser statistiquement l'erreur obtenue par la fonction de coût, sans observer de phénomène de sur-apprentissage. Avec l'avènement des réseaux de neurones profonds, les algorithmes classiques de descente de gradients représentent un défi à eux seuls, puisque le problème posé est un problème d'optimisation de dimension très élevée.

Pour cela, la communauté scientifique a développé un grand nombre de méthodes d'optimisations adaptées à ce type de problème, ayant pour objectif d'éviter des problèmes d'extinction, d'explosion de gradients, et la gestion des points de selles. Parmi elles, on peut citer plusieurs classes d'algorithmes, comme la Stochastic Descent Gradient, les algorithmes basés sur les moments, et les algorithmes adaptatifs, comme l'optimiseur « Adam » qui sera décrit plus en détail dans la suite du manuscrit. Enfin, au vu du très grand nombre d'opérations à effectuer, tous les calculs sont parallélisés sur des processeurs graphiques (GPU) grâce à la librairie Python *Tensorflow* [116]. Cette librairie a la particularité de pouvoir construire un graphe computationnelà partir d'un code Python et de l'exécuter soit sur CPU, soit sur GPU, de manière indifférenciée pour l'utilisateur. Le travail fourni pour développer l'approche BeamLearning s'est donc fait en langage Python, sans se soucier de l'implémentation sur GPU ou d'optimisation CUDA, qui sont prises en charge par la librairie TensorFlow.

2.1.6.1 Initialisation des variables d'apprentissage

Pour la première itération de l'entrainement, toutes les variables d'apprentissage du réseau associé à l'approche de localisation par BeamLearning sont initialisés grâce à une loi de distribution normale centrée sur 0. En revanche, le deuxième paramètre caractérisant les lois normales, l'écart-type, dépend du rôle de la variable à initialiser dans le réseau. Nous avons fait le choix d'initialiser toutes les variables d'apprentissage des couches convolutives en utilisant le processus d'initialisation de He [117], qui consiste à initialiser selon un loi aléatoire pilotée par le nombre d'entrée et de sortie de la couche neuronale. Même si elle a été conçue initialement pour exploiter l'allure de la fonction d'activation RELU, cette approche est connue pour améliorer les performances d'entrainement avec plusieurs types de fonctions d'activation par rapport à une simple loi aléatoire. Pour tous les autres variables d'apprentissage (pointwise, FC...), l'écart type vaut 0,1, nos tests ayant montré que l'initialisation de He n'apportait pas d'amélioration sensible pour ces portions du réseau.

2.1.6.2 Optimisation des variables d'apprentissage

Une des raisons pour lesquelles les réseaux de neurones profonds permettent aujourd'hui d'atteindre des résultats de grande qualité – malgré leur complexité – réside dans les progrès réalisés pour le mécanisme d'optimisation des variables d'apprentissage. Dans la configuration présentée dans ce travail, $1,2 \times 10^6$ variables doivent être optimisées à chaque itération : le choix de la méthode d'optimisation est donc primordiale. La méthode choisie est la méthode Adam [118], pour *Adaptative Momentum Estimation*.

Avant de décrire cet optimiseur plus en détail, il est nécessaire de dresser un cadre général sur les techniques associées aux descentes de gradient. Pour des raisons de clarté de l'exposé, les notations utilisées ici ne présentent une optimisation que sur une seule variable Θ, mais il est essentiel de garder en tête qu'en pratique, ce problème d'optimisation est de grande dimensionnalité, Θ représentant en fait un ensemble pouvant atteindre des millions de variables. Pour un algorithme de descente de gradient, la mise à jour de la variable Θ se fait classiquement à partir de la valeur initiale de la variable et du pas de l'optimisation multiplié par le gradient de la fonction de coût choisie [4] , $\nabla_\Theta \mathcal{F}(\Theta; (x, y))$, qui mesure les performances du réseau de neurones :

$$\Theta_{t+1} = \Theta_t - \eta \nabla_\Theta \mathcal{F}(\Theta; (x, y)) \tag{2.5}$$

Cette fonction de coût dépend à la fois de la variable à optimiser, mais aussi de l'exemple choisi x pour optimiser les variables, ainsi que du *label* y de cet exemple. L'optimisation se fait à chaque itération grâce au gradient calculé à partir de l'ensemble des données d'apprentissage. Dans ce cas, cette méthode est appelée *batch gradient descent* (BGD). Comme cette approche est longue et souvent redondante, il est possible de ne calculer $\nabla_\Theta \mathcal{F}$ que pour un unique couple d'exemple / *label* (méthode *stochastic gradient descent* : SGD). Enfin, un compromis peut être trouvé en ne calculant le gradient de la fonction de coût que sur une petite partie d'exemples (méthode *mini-batch gradient descent*), qui permet d'allier une bonne représentation statistique du gradient, tout en garantissant une grande vitesse de calcul [119]. Dans notre cas, la taille du mini-batch est de 100 exemples. Dans ce manuscrit, le terme de *batch* employé correspond en réalité au *mini-batch*.

4. Une discussion sur le choix de cette fonction de coût a lieu en section 2.2.

Pour favoriser la stabilité de l'apprentissage, le pas choisi entre chaque itération peut être variable. Dans le cadre de ce travail, le pas est exponentiellement décroissant avec le nombre d'itérations. Plus l'apprentissage avance, donc potentiellement plus il est proche du point de convergence, plus les pas sont petits. Pour ne pas surcharger les notations, cette variabilité sera implicite.

Pour augmenter la vitesse de convergence de l'algorithme, il est possible d'ajouter une notion *d'inertie* à l'équation, en utilisant la notion de moment, exprimé $\nu_t = \sum_{\tau=1}^{t} \eta \nabla_\Theta \mathcal{F}$. Ce moment caractérise l'historique des évolutions des gradients. Comme l'indique l'équation 2.5, le moment correspond à une sommation pondérée des différentes valeurs précédentes du gradient. Par conséquent, il est possible de mettre à jour la variable Θ d'autant plus que ses variations antérieures ont été importantes. L'optimisation se fait alors grâce à la formule :

$$\Theta_{t+1} = \Theta_t - \eta \cdot g_t + \nu_t \tag{2.6}$$

où $g_t = \nabla_\Theta \mathcal{F}$ pour l'itération t. Enfin, l'algorithme Adam [118] propose d'adapter le pas d'optimisation en fonction du nombre d'itérations t, de la valeur des gradients et des moments de l'itération présente et précédente. La mise à jour de la variable Θ est alors définie par :

$$\Theta_{t+1} = \Theta_t - \frac{\eta \cdot \tilde{m}_t}{\sqrt{\tilde{\nu}_t} + \epsilon}, \tag{2.7}$$

$$\text{avec : } \begin{cases} \tilde{m}_t = \frac{m_t}{1-(\beta_1)^t} \\ \tilde{\nu}_t = \frac{\nu_t}{1-(\beta_2)^t} \end{cases} \text{ , et : } \begin{cases} m_t = (1-\beta_1) g_t + \beta_1 m_{t-1} \\ \nu_t = (1-\beta_2) g_t^2 + \beta_2 \nu_{t-1} \end{cases}$$

où les coefficient $\beta_1 = 0,9$ et $\beta_2 = 0,99$ ne servent qu'à corriger un léger biais de m_t et μ_t observé par les auteurs. Le coefficient ϵ quant à lui, est un coefficient infinitésimal qui assure que la division dans

l'équation 2.7 ne soit jamais nulle.

Grâce à cet algorithme, déjà disponible dans la librairie de calcul *Tensorflow* les variables d'apprentissage sont mises à jour au fur et à mesure de l'apprentissage.Comme dit précédemment, il est à noter que cette explication n'est ici formellement faite que sur l'optimisation d'une seule variable, alors que l'algorithme doit optimiser dans notre cas plus d'un million de variables et à *fortiori* plusieurs millions de dérivées partielles pour les gradients.

2.2 Choix de la représentation de l'espace angulaire

Pour traiter le problème de localisation angulaire d'une source par machine learning, deux approches de représentations angulaires peuvent être utilisées : l'une consiste en un partitionnement de l'espace angulaire en subdivisions, afin de déterminer l'appartenance de la source à l'une de ces portions d'espace angulaire. Cette manière d'aborder le problème est traité en apprentissage supervisé avec l'approche par classification. Une autre manière de rendre compte de la position de la source est d'estimer directement ses coordonnées. L'espace des solutions est dans ce cas un *continuum*, potentiellement à plusieurs dimensions. Dans ce cas, l'approche par régression est utilisée. L'approche de régression et de classification pour la localisation de sources ont été comparées en particulier dans l'article de Tang *et al.* [79], et seront toutes les deux développées dans la suite de ce manuscrit, pour des problèmes de détermination d'angles d'arrivée (DOA en anglais) à 2 dimensions, ou à 3 dimensions, que ce soit en espace clos ou en environnement ouvert.

2.2.1 Classification angulaire à deux dimensions

Au cours du développement de l'approche BeamLearning pendant ma thèse de doctorat, le problème de localisation de sources a tout d'abord été traité comme un problème de classification, ce qui permet de simplifier le problème en restreignant la localisation à des zones de l'espace azimutal. Même si la communauté scientifique ne s'est intéressée à la localisation de sources acoustiques par machine learning que depuis très peu de temps, l'approche par classification est assez commune [67, 68, 70]. Pour ce type de problème de classification angulaire, l'espace angulaire est découpé en N sous-espaces, chacun correspondant à une classe différente (cf. figure 2.13).

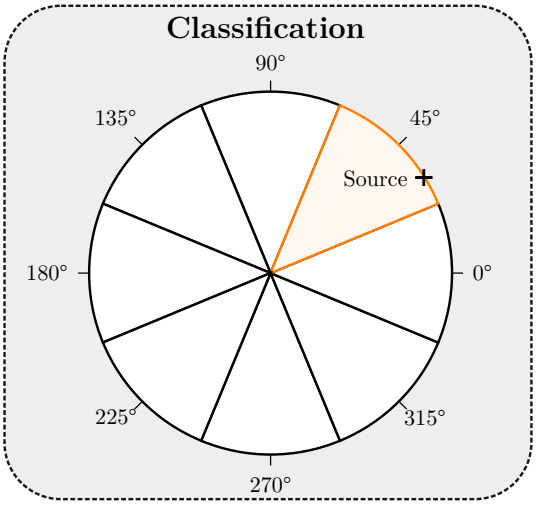


FIGURE 2.13 – Découpage en sous-espaces angulaires pour le problème de classification : exemple pour 8 classes. La sortie correspondante est ici un vecteur de longueur 8, encodant la probabilité d'appartenance de la source à chaque secteur angulaire.

Avec cette approche de classification, la sortie du réseau, qui est le secteur angulaire estimé $\tilde{\theta}$, correspond à un vecteur orthonormé à N dimensions avec N le nombre de classes possibles. Ce vecteur contient les différentes probabilités d'appartenance de la source aux différents secteurs angulaires. Le vecteur estimé $\tilde{\theta} = [0.1, 0., 0.75, 0.05]$ correspond par exemple à un source ayant une probabilité de 75% d'être située dans le $3^{\text{ème}}$ secteur angulaire.

Dans le cas optimal, la fonction de répartition est un vecteur pouvant être assimilée à une fonction de Dirac : la sortie pour la classe à laquelle appartient la source vaut 1, et toutes les autres valeurs valent 0, comme le *label*. Ce cas idéal n'étant pas systématiquement atteint, les N probabilités représentées par les valeurs du vecteur de sortie $\tilde{\theta}$ prennent des valeurs entre 0 et 1, et celle qui prend la valeur la plus élevée correspond à l'estimateur de la position de la source, cette estimation étant d'autant plus certaine que la valeur est proche de 1. Pour une meilleure expressivité du résultat, la fonction *softmax* (fonction exponentielle normalisée) est appliquée à chaque composante $\tilde{\theta}_i$ du vecteur $\tilde{\theta}$:

$$\mathcal{S}(\tilde{\theta}_i) = \frac{e^{\tilde{\theta}_i}}{\sum_{1}^{N} e^{\tilde{\theta}_k}} \tag{2.8}$$

Pour ce type de problème, la fonction de coût $\mathcal{F}$ la plus adaptée est l'entropie croisée, qui permet de représenter l'écart entre l'ensemble des sorties du réseau et l'ensemble des labels correspondants pour un batch d'apprentissage. En notant y le label de l'entrée testée, et $\tilde{y}$ la sortie (après application de la fonction *softmax*) effectivement obtenue par BeamLearning dans un problème de classification, la fonction de coût à optimiser est alors définie par la formule suivante :

$$\mathcal{F}(\tilde{y}, y) = -\sum_i \log(\tilde{y_i}) \cdot y_i \tag{2.9}$$

Afin d'illustrer ce type d'approche sur un problème relativement simple et d'en tirer des tendances observables graphiquement, la figure 2.14 présente un exemple de diagramme de directivité [5] en sortie du réseau obtenu après un entrainement pour un problème de localisation 2D, traité comme un problème de classification dans 8 secteurs angulaires. Ce diagramme de directivité permet de mieux comprendre l'allure des vecteurs de sortie après convergence du réseau, et d'interpréter la sélectivité angulaire du réseau.

Sur la figure 2.14, les courbes en couleur représentent la valeur de la sortie de chaque composante du vecteur de sortie, en fonction de l'angle d'arrivée de la source. Les couleurs en arrière plan désignent la prédiction de BeamLearning en fonction de ces différentes sorties. L'analyse de cette figure permet de voir qu'y compris dans les secteurs angulaires qui ne correspondent pas à son domaine de détection, chaque neurone de sortie spécialisé pour une portion angulaire possède des valeurs non nulles. On observe en effet pour chaque neurone de sortie une courbe en trait plein en forme de *fleur* à 8 pétales, où chaque pétale pointe dans la direction correspondante aux 8 secteurs angulaires utilisés pour traiter le problème de localisation par une approche de classification.

Il est intéressant de noter qu'en superposant les diagrammes de directivité de chaque neurone en sortie du réseau comme sur la figure 2.14, on observe ainsi que pour chaque secteur angulaire, le «

5. Le terme de diagramme de directivité est ici utilisé par analogie avec les diagrammes utilisés en électro-acoustique. Plutôt que de représenter la pression mesurée par un microphone en fonction de la position de la source, c'est la dépendance angulaire de la réponse de chaque neurone qui est ici représentée, mais toujours pour des sources positionnées tout autour de l'antenne de captation.

pétale » de directivité ayant la valeur maximale appartient bien au neurone de sortie spécialisé dans cette direction. Cette observation permet de confirmer que chaque neurone répond bien de manière prédominante par rapport aux autres dans une direction donnée.

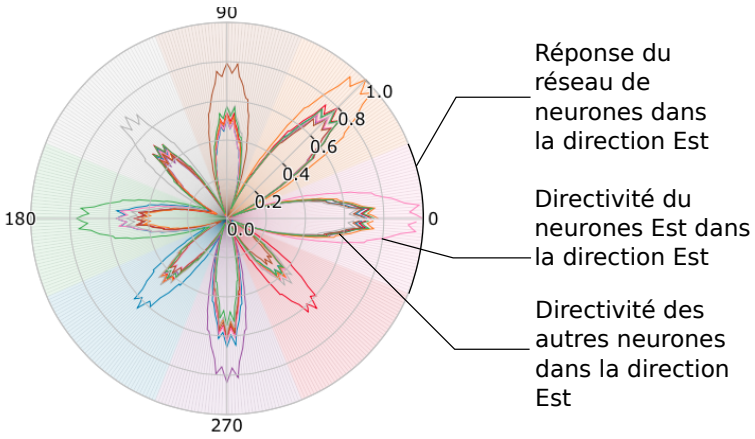


FIGURE 2.14 – Diagramme de directivité de BeamLearning pour une régression à 8 classes. Les courbes colorées représentent la réponse de chaque neurone, alors que la couleur de fond représente la classe estimée par le réseau pour chaque position de source.

En revanche, si on analyse chaque diagramme de directivité des 8 neurones de sortie indépendamment, le lobe de directivité de valeur la plus élevée de chaque neurone ne correspond pas forcément à la zone de l'espace pour laquelle le neurone est spécialisée. Par exemple, pour le neurone spécialisé dans la direction Sud-Ouest (pointant vers 225°), en bleu sur la figure 2.14, la valeur maximale de son diagramme de directivité pointe pourtant dans la direction Nord-Est (pointant vers 45°), sans pour autant que le problème de classification n'en soit affecté. En effet, malgré ce maximum individuel dans la direction 45° pour le neurone pointant à 225°, dans la portion à 45°, le neurone associé à ce secteur angulaire (en orange) prend bien une valeur supérieure à tous les autres neurones.

Pour finir, la forme particulière de *fleur* de la figure de directivité des neurones de sortie provient du fait que pour ce problème de classification, tous les neurones de sortie possèdent une probabilité très faible (directivité tendant vers 0) lorsque la source se trouve dans une direction correspondant à l'intersection entre deux secteurs angulaires. Pour ces positions de sources se situant à la jonction

entre plusieurs classes, le problème de classification est particulièrement ambigu et difficile à résoudre, puisque les indices extraits par le réseau sur la localisation de la source dans les signaux microphoniques sont trop peu différents de part et d'autre de la frontière pour que la sortie soit certaine dans ces zones. Pour toutes ces raisons, il est essentiel, afin d'interpréter le comportement global en sortie du réseau et ses capacités de localisation dans des secteurs angulaires, de représenter toutes les sorties, et de les comparer les unes aux autres.

Comme expliqué plus haut, cette illustration à 8 classes angulaires permet d'extraire et de comprendre graphiquement le comportement de l'approche BeamLearning en classification, mais le choix de 8 classes reste trop faible pour être satisfaisant pour toutes les applications de localisation. Dans ce cas illustratif, les secteurs angulaires possèdent en effet une largeur de 45°, ce qui peut être pertinent dans le cas des assistants personnels par exemple (Homepod, Alexa Assistant, Djingo, Google Home), où une localisation du locuteur peut servir à améliorer la reconnaissance vocale nécessaire à leur fonctionnement. Une autre application où cette simplification du problème de localisation pourrait être pertinente est le cas de la visioconférence, où la caméra pourrait s'orienter automatiquement dans le secteur angulaire du locuteur. Dans ces cas, l'avantage principal de la simplification offerte par le faible nombre de secteur angulaires de classification réside dans le fait que l'entrainement du réseau peut être beaucoup plus rapide (quelques heures) qu'avec une approche de régression comme celle qui sera détaillée dans la sous-section suivante. En revanche, pour des applications nécessitant une résolution angulaire plus précise, comme pour le suivi de trajectoire ou l'inspection structurale, une approche de classification à faible nombre de classes n'est pas suffisante.

Lorsqu'une résolution angulaire plus importante est attendue, une approche naïve pourrait consister à augmenter le nombre de classes pour avoir une meilleure précision angulaire. Cependant, comme discuté sur l'exemple de 8 secteurs angulaires, la limite de l'approche par classification est la frontière séparant deux classes. En effet, comme la source n'est en pratique jamais parfaitement ponctuelle, plus la source se rapproche de la frontière, plus sa position est ambiguë. Or, en augmentant le nombre de classes, le nombre de frontières augmente nécessairement, et avec lui le nombre de cas ambigus. Une fois le nombre de classes devenu important (> 128), l'approche par classification perdrait alors de son intérêt : pour une nombre élevé de classes le problème peut être traité comme un problème de

régression, ce qui est pourtant rarement réalisé dans la littérature [68].

2.2.2 Régression angulaire à deux dimensions

Afin de retrouver la position angulaire précise de la source dans un *continuum* d'azimuts possibles, l'approche par régression est plus appropriée. Avec ce type d'approche, la sortie n'est plus un vecteur dont les composantes représentent la probabilité d'appartenance à une zone de l'espace, mais une valeur continue. Dans notre cas, la valeur voulue est un angle, en degré (ou en radian), dans l'intervalle $[0; 360[$ (respectivement $[0; 2\pi[$) comme la figure 2.15 l'illustre.

Or, un angle est intrinsèquement périodique, ce qui peut éventuellement poser un problème de calcul de l'erreur résiduelle par la fonction de coût si elle est mal choisie. En effet, une estimation d'une position de source à un azimut de 358 degrés ne représente qu'une erreur de 3 degrés lorsque la source est en réalité positionnée à 1 degré d'azimut. C'est la raison pour laquelle il est primordial, dans ce cas, que la fonction de coût associée au calcul de l'erreur d'estimation de l'angle respecte elle aussi cette périodicité. Afin de faciliter la rétropropagation des gradients et le mécanisme d'optimisation associé à l'entraînement, il est en outre nécessaire que la fonction de coût soit différentiable.

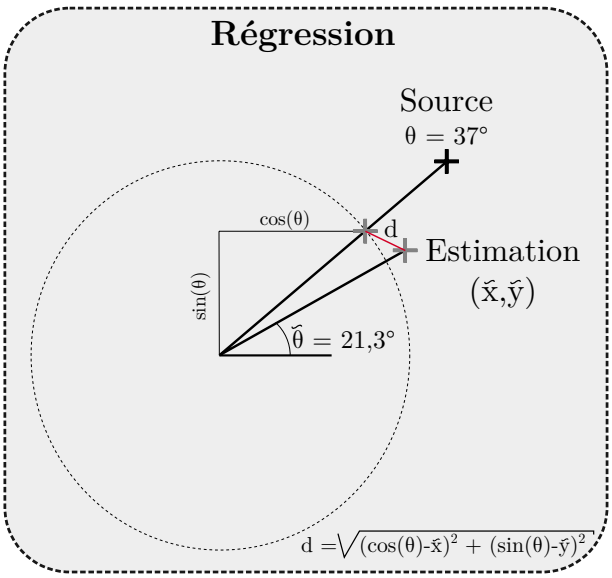


FIGURE 2.15 – Schéma de fonctionnement de la régression pour un problème de DOA 2D : La sortie obtenue est une grandeur continue représentant une position dans le plan (ou dans l'espace en cas de localisation 3D)

Pour satisfaire ces deux critères, plutôt que de ne définir que la grandeur continue $\tilde{\theta}$ en sortie du réseau, une approche pertinente consiste à réaliser une régression par rapport à un *label* (étiquette) représentant la position de la source à localiser ramenée au cercle unité : $(x, y) = (\cos(\theta), \sin(\theta))$, comme représenté sur la figure 2.15. L'estimation angulaire par le réseau est quant à elle obtenue grâce à deux positions $\tilde{x}$ et $\tilde{y}$, qui doivent s'approcher au mieux des valeurs des cosinus et sinus de l'angle θ de la position réelle de la source (voir figure 2.15). En effet, les fonctions cosinus et sinus étant périodiques et $\mathcal{C}^\infty$, il est plus facile de les utiliser dans la fonction de coût de la position angulaire de la source, plutôt que d'utiliser l'angle directement. Par ailleurs, l'interprétation géométrique illustrée sur la figure 2.15 permet de voir que l'étiquette (x, y) représente tout simplement les coordonnées cartésiennes de la position angulaire de la source, ramenées sur le cercle unité. Dans ces conditions, la fonction de coût $\mathcal{F}$ choisie correspond à une simple norme $\mathcal{L}_2$ dans le plan, correspondant à la distance entre le point sur le cercle unité représentant le label de la source et le point estimé $(\tilde{x}, \tilde{y})$ dans le plan :

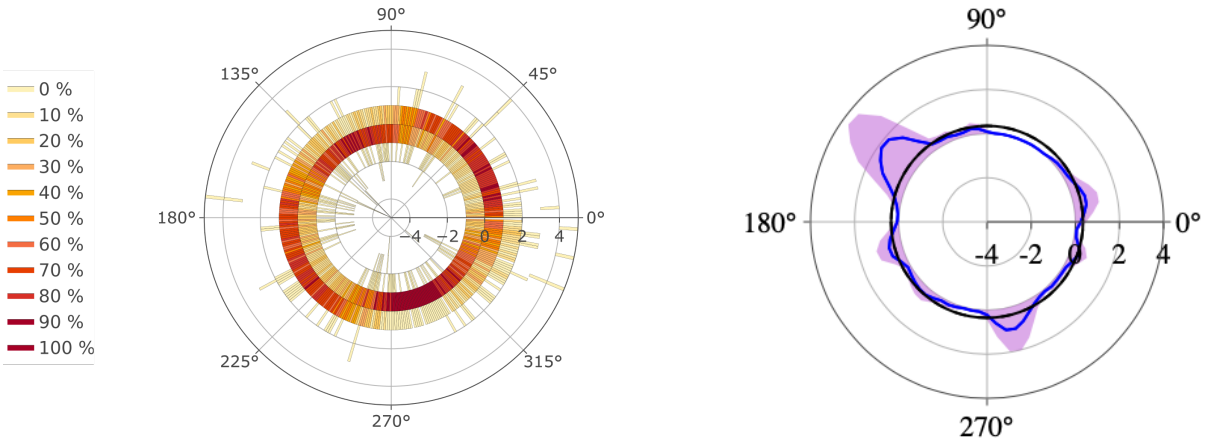
$$\mathcal{F} = \sqrt{(\cos(\theta) - \tilde{x})^2 + (\sin(\theta) - \tilde{y})^2}$$

Il est par ailleurs essentiel de noter que pour l'approche BeamLearning par régression, seule la position de la source est projetée sur le cercle unité, en faisant une hypothèse de champ lointain. En revanche, les positions $(\tilde{x}, \tilde{y})$ estimées en sortie du réseau de neurones profond proposé ne sont pas contraintes à appartenir au cercle unité, mais évoluent librement dans le plan (x,y). L'optimisation des variables d'apprentissage au cours de l'entraînement du réseau de neurones suffit à ce que les positions retrouvées $(\tilde{x}, \tilde{y})$ convergent vers une position *suffisamment proche* du cercle unité. Les entraînements réalisés avec des variantes de cette fonction de coût, cherchant à contraindre les valeurs $(\tilde{x}, \tilde{y})$ sur le cercle unité, ont en effet montré que cette approche complexifiait inutilement la fonction de coût, sans pour autant offrir d'améliorations sensibles, ni en termes de vitesse de convergence du réseau (c'est même une dégradation que nous avons observé), ni en précision de l'estimation angulaire après convergence (sans changement significatif).

Pour analyser la sortie du réseau après convergence de l'apprentissage, l'outil de tracé de dia-

gramme de directivité des sorties pour le problème de classification (voir figure 2.14) n'aurait plus de sens pour le problème de régression qui nous intéresse dans cette section. En effet, avec une approche par régression, il n'y a plus du tout de notion d'intensité de réponse des neurones de sortie en fonction de l'angle de la source, puisqu'avec une approche de régression, la réponse attendue est un nombre réel correspondant à la position de la source – et non plus un vecteur de sortie représentant une probabilité d'appartenance à des secteurs angulaires.

En revanche, il est possible de quantifier finement l'erreur de localisation de la source acoustique après entraînement et convergence du réseau, en traçant un diagramme polaire d'erreur absolue angulaire et de son écart-type. Ce type de visualisation permet d'analyser précisément les performances de localisation, et d'évaluer si certaines portions angulaires présentent des défauts de localisation plus importants que d'autres. À ce titre, la figure 2.16 propose deux types de visualisations possibles.



(a) Histogramme de l'erreur de localisation

(b) Moyenne et écart type de l'erreur angulaire

FIGURE 2.16 – Représentations possible de l'erreur angulaire pour un problème de localisation à 2D. À gauche, représentation de l'histogramme de l'erreur angulaire – à droite, représentation polaire des deux premiers moments statistiques de l'erreur angulaire.

Parmi ces deux types de visualisations, celle qui présente le plus d'informations est l'histogramme des erreurs angulaires **signées**, en fonction de l'angle de la source à localiser (voir figure 2.16(a)). Pour construire ce graphique, un ensemble de 9 600 sources provenant d'exemples non utilisés pendant la phase d'entraînement du réseau ont été tirées aléatoirement. L'ensemble de ces 9 600 données constitue la base de données de test, à différencier de la base de données d'apprentissage, constituée d'exemples

utilisés lors de la phase d'optimisation du réseau [6] .

Dans ce jeu de données de test, 9 600 données microphoniques multicanales, correspondant à autant de positions de sources, sont présentées en entrée du réseau gelé, après entraînement et convergence. Les erreurs de localisation sont ensuite calculées pour chacune de ces 9 600 positions, et représentées sur un histogramme permettant de quantifier la proportion des erreurs. Les erreurs de localisation en fonction de la position réelle de la source sont regroupées par quantiles correspondant à des tranches de 1 ° d'erreur angulaire. Ainsi, lorsque les erreurs d'estimation angulaire sont élevées, les barres de l'histogramme polaire s'éloignent du cercle à 0°, que l'on considère donc comme le cercle de référence représentant une localisation angulaire *parfaite*, avec moins d'un degré d'erreur. Lorsque les erreurs sont positives, les barres de l'histogramme polaires sont orientées vers l'extérieur, tandis qu'elles sont orientées vers l'intérieur du diagramme lorsque les erreurs sont négatives. La proportion statistique des erreurs dans chaque quantile angulaire de 1° est représentée en niveau de couleurs dans chaque barre de l'histogramme polaire. À titre d'exemple, une proportion de 10% de sources sonores positionnées à 90° qui seraient localisées par le réseau de neurones entre 88° et 89° serait représentée par une *case* de l'histogramme positionnée en $(90, -2)$, dont la couleur correspondrait à la valeur de 10%.

Cette représentation en histogramme polaire présentée sur la figure 2.16(a) est très complète et porteuse d'une information exhaustive et synthétique. En revanche, elle peut être complétée avantageusement - ou remplacée - par une représentation encore plus synthétique, en faisant abstraction des quantiles angulaires (voir figure 2.16(b)). Sur ce type de représentation simplifiée, plutôt que de représenter la proportion statistique dans chaque quantile angulaire, seuls les deux premiers moments statistiques sont conservés. Malgré cette simplification, comme illustré sur la figure 2.16(b), le diagramme polaire permet d'analyser finement les performances de localisation du réseau. En premier lieu, l'erreur moyenne sur 360°, tracée en noire, permet de vérifier que notre estimateur n'est pas biaisé, puisque pour les 9 600 sources du jeu de test, l'erreur angulaire moyenne est nulle. C'est en soi l'objectif de la phase d'entraînement du réseau, qui vise à minimiser statistiquement les erreurs pour un maximum de sources, et non pour quelques positions bien déterminées. Les coefficients appris par le

6. De plus amples informations à propos des bases de données sont fournies dans les chapitres suivants (chapitres 3 et 4)

réseau dans ses couches atteignent donc leur objectif fixé : généraliser au mieux la tâche de localisation à un ensemble de sources, y compris pour des données non présentées au cours de l'entraînement. Par ailleurs, la courbe en bleu représente l'erreur moyenne pour chaque position angulaire. En effet, le jeu de données de test présenté pour analyser les performances du réseau étant constitué de 9 600 sources, pour chaque degré angulaire d'estimation, cela offre une statistique sur 26 sources. Autour de cette courbe, la zone violette représente l'écart type de la distribution. Ce qui permet de mieux mettre en évidence les zones où des erreurs de localisation sont commises.

Sur cet exemple illustratif, les performances sont donc très bonnes, malgré les quelques écarts de la moyenne locale par rapport à la moyenne globale. En effet dans l'exemple de la figure 2.16(b), l'écart constaté à 135° reste inférieur à 2° d'erreur absolue, avec un écart type n'excédant pas, lui non plus, les 2°. De plus ces écarts ne résultent pas d'une mauvaise représentativité de la base de données d'apprentissage, puisque la position de ces écarts locaux varient si un nouvel entrainement est réalisé, avec un nouveau tirage aléatoire lors de l'initialisation des variables d'apprentissage.

2.2.3 Généralisation de l'approche de régression pour une localisation angulaire à 3 dimensions

La généralisation de ces approches de localisation de sources par apprentissage supervisé pour un problème à 3D est tout à fait possible, que ce soit pour le problème de classification [63, 79], ou pour le problème de régression [68, 79]. Les remarques précédentes sur la diminution de la pertinence d'une approche de classification lorsque le nombre de classes augmente sont d'autant plus vraies en 3 dimensions. C'est la raison pour laquelle nous avons favorisé une approche par régression du problème de localisation angulaire à 3 dimensions dans le cadre de cette thèse de doctorat.

De manière analogue au cas à deux dimensions, plutôt que de travailler avec le couple d'angle (θ, ϕ) des sources, les informations exploitées pour le calcul de la fonction de coût sont les suivantes :

— le label de la source est un vecteur correspondant aux projections $\mathbf{P}_{norm} = (x_{norm}; y_{norm}; z_{norm})$ du vecteur coordonnées de la source sur la sphère unité

— l'estimation en sortie du réseau est un vecteur $\tilde{\mathbf{P}} = (\tilde{x}; \tilde{y}; \tilde{z})$ en 3 dimensions.

Tout comme pour le problème à 2 dimensions, les coordonnées du vecteur $\tilde{\mathbf{P}}$ ne sont pas contraintes à appartenir à la sphère unité. En réutilisant les notations sphériques introduites dans l'équation reliant les coordonnées sphériques à cartésiennes (Éq. 1.1), le label de chaque source utilisée pour constituer la base de données correspond donc à un vecteur $\mathbf{P}_{norm} = (x_{norm}; y_{norm}; z_{norm})$, paramétré par la position angulaire de la source :

$$\begin{cases} x_{norm} = x/r = \cos(\theta)\cos(\phi) \\ y_{norm} = y/r = \sin(\theta)\cos(\phi) \\ z_{norm} = z/r = \sin(\phi) \end{cases} \tag{2.10}$$

Dans le cas à 3 dimensions, la fonction de coût choisie représente également la distance euclidienne à 3 dimensions entre la position $(\tilde{x}; \tilde{y}; \tilde{z})$ et la position sur la sphère unité $(x_{norm}; y_{norm}; z_{norm})$, grâce à une simple norme $\mathcal{L}_2$. Par conséquent, cette fonction de coût $\mathcal{F}$ s'exprime de la manière suivante :

$$\mathcal{F} = \sqrt{(x_{norm} - \tilde{x})^2 + (y_{norm} - \tilde{y})^2 + (z_{norm} - \tilde{z})^2}$$

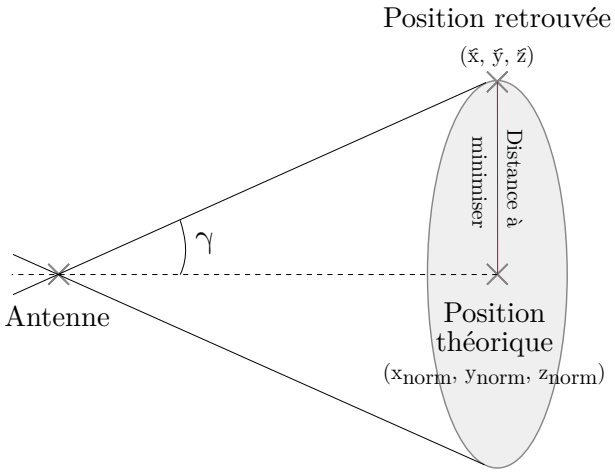


FIGURE 2.17 – Demi Cône permettant de définir l'erreur 3D angulaire de localisation

Même si la fonction de coût exploite des coordonnées 3D cartésiennes, il est évidemment possible d'analyser finement les performances de localisation en fonction des angles (θ, ϕ) séparément, afin de vérifier qu'il n'y a pas une direction privilégiée lors de l'apprentissage, et que les erreurs de localisation sont équivalentes en azimut et en élévation. Une autre grandeur pertinente pour l'analyse des performances de localisation en 3 dimensions consiste en l'erreur absolue angulaire, qui combine l'influence des erreurs en azimut et en élévation (voir figure 2.17). La zone ainsi définie est donc un demi-cône dont le sommet est l'antenne, la génératrice est une droite passant par l'antenne et la position retrouvée, et dont la courbe directrice est un cercle dont le centre est la position réelle de la source. L'angle γ entre la génératrice du cône et la droite passant par le sommet du cône et la position réelle, qui représente ainsi l'erreur angulaire 3D absolue de localisation, est déterminée à partir du vecteur position de la source retrouvée $\tilde{\mathbf{P}}$ et le vecteur position réel de la source $\mathbf{P}_{norm}$, grâce à la formule suivante :

$$\gamma = \arctan\left(\frac{\mathbf{P}_{norm} \wedge \tilde{\mathbf{P}}}{\mathbf{P}_{norm} \cdot \tilde{\mathbf{P}}}\right) \tag{2.11}$$

2.3 Analyse du réseau en profondeur

Le fil conducteur qui a guidé la conception du réseau associé à l'approche de localisation de sources par BeamLearning est l'interprétabilité physique des opérations réalisées par les différentes couches du réseau. À ce titre, la plupart des analyses seront détaillées sur des situations représentatives dans le chapitre 5 de cette thèse de doctorat. Toutefois, pour mieux comprendre l'intérêt de l'architecture proposée, nous proposons ici une analyse générale du comportement – après entraînement et convergence du réseau – des filtres constituant les premières couches du réseau (le sous-réseau "bancs de filtres"), situées avant le calcul de la pseudo-énergie.

Par analogie avec l'analyse de filtres spatiaux et temporels utilisés classiquement en acoustique, leur réponse sera analysée en fonction de la fréquence, en leur présentant en entrée des signaux monochromatiques émis par des sources dont la position est définie, afin d'offrir une analyse à la fois fréquentielle et angulaire.

L'objectif de l'analyse proposée dans les sous-sections qui suivent n'est pas de dresser un catalogue

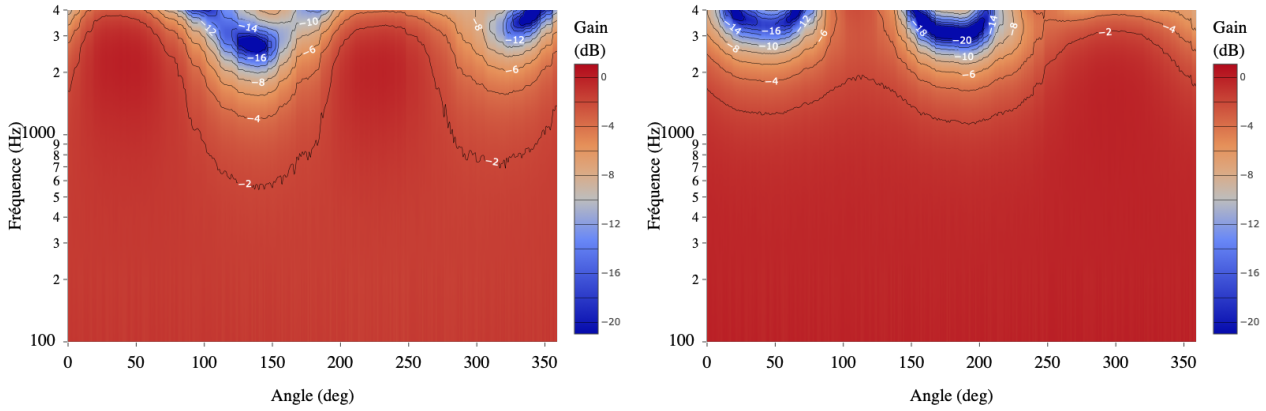
exhaustif des comportements observés, mais plutôt de mettre en avant les tendances pour plusieurs couches de filtres, et de mettre en exergue l'influence des non linéarités entre couches neuronales. L'approche proposée est très similaire à celle utilisée pour la caractérisation en fréquence et en réponse angulaire de capteurs et d'antennes microphoniques : afin de caractériser le système plutôt que son environnement, ce type de mesure est réalisée en environnement anéchoïque ou neutre, pour des signaux monochromatiques. Les filtres analysés dans la suite sont donc ceux obtenus pour un problème de régression, entraînés avec une base de donnée constituée de signaux monochromatiques, pour un problème de localisation de sources à 2 dimensions en champ libre. Des bases de données plus complexes seront présentées dans les chapitres suivants, et leurs résultats interprétés au chapitre 5 de ce document.

2.3.1 Analyse de la première couche de filtre

Dans un premier temps, on se propose d'analyser le comportement de la première couche de filtres du sous-réseau de bancs de filtres présenté en figure 2.3. Ces filtres sont tous multicanaux, puisqu'ils ont tous en entrée les N_{mic} canaux microphoniques. Compte tenu de la topologie du réseau, ces filtres sont donc un ensemble de $N_f = 128$ filtres à N_{mic} entrées et 1 sortie, possédant un noyau de longueur 3, avec un facteur de dilatation de 1 (voir section 2.1.3). Pour cela, l'interprétation portera sur le comportement de la sortie de chaque filtre de la couche, lorsque l'entrée du réseau correspond aux signaux captés sur une antenne microphonique résultant du champ de pression émis par des sources monochromatiques, sur le domaine [100 Hz ; 4 000 Hz], par pas de 100 Hz [7] . Cette analyse est menée pour 360 positions de sources par pas de 1 degré dans le plan de l'antenne, afin d'étudier le filtrage spatial offert par ces filtres construits par le réseau pendant sa phase d'entraînement.

Cette analyse offre trois degrés de liberté d'analyse : le comportement fréquentiel, angulaire, ainsi que le numéro du filtre de la couche considérée. Dans notre cas, chaque couche du sous-réseau de bancs de filtres étant constituée de N_f filtres ($N_f = 128$ dans le cas de cette thèse), chaque filtre possède ses propres caractéristiques fréquentielles et angulaires. L'objectif n'est pas ici de dresser un catalogue exhaustif de ces filtres, mais plutôt d'extraire les grandes tendances comportementales pour cette première couche, celle qui est au plus près des données microphoniques.

7. Ce domaine correspond au domaine de validité imposé par le dispositif de spatialisation présenté au chapitre 4

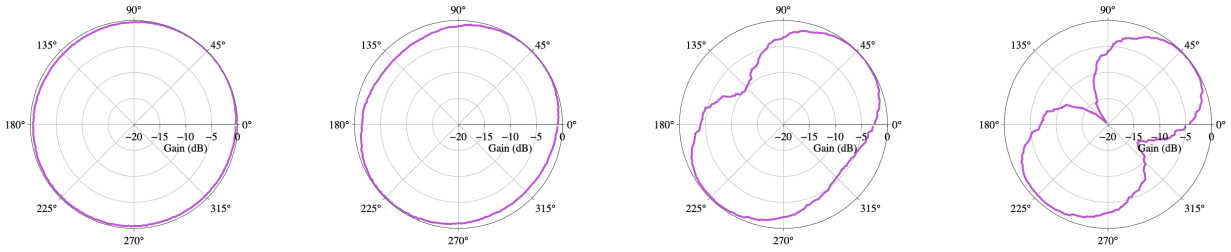


(a) (b)

FIGURE 2.18 – Exemple de la réponse en fréquence de deux différents filtres passe-bas de la première couche convolutive de BeamLearning

La première tendance est illustrée par la figure 2.18, qui représente, dans un diagramme angle-fréquence le gain offert en sortie de deux filtres de cette couche, qui possèdent un comportement globalement passe-bas. En effet, l'analyse de la figure 2.18 montre que pour ces deux filtres, les gains en sortie de forte amplitude se situent majoritairement dans le domaine fréquentiel couvrant la bande $[100Hz, 1000Hz]$, avec un comportement globalement omnidirectionnel dans cette bande.

Ce comportement de directivité en sortie peut également être observé en traçant la réponse angulaire du filtre pour quelques fréquences discrètes, comme présenté en figure 2.19, qui représente la directivité du filtre multicanal de la figure 2.18(a) pour les fréquences 500 Hz, 1 000 Hz, 2 000 Hz et 3 000 Hz. À partir de 1 000 Hz, on voit apparaître une direction privilégiée dans la réponse angulaire du filtre multicanal, à partir des données brutes captées par l'antenne microphonique, en entrée du réseau. Cette direction, correspondant à l'axe $(225°, 45°)$, se précise de plus en plus avec l'augmentation de la fréquence jusqu'à devenir un diagramme dipolaire à partir de 3 000 Hz. Cette direction privilégiée était en fait déjà visible dans la figure 2.18(a). De la même manière, le filtre présenté en figure 2.18(b), présente un comportement monopolaire en basse fréquence (jusqu'à 1 500 Hz environ), et évolue vers un comportement dipolaire en hautes fréquences, à partir d'une fréquence de 3 000 Hz. La différence réside dans la direction privilégiée en hautes fréquence, qui pour le second filtre, est quant à elle dirigée selon l'axe $(120°, 300°)$.



(a) 500 Hz (b) 1000 Hz (c) 2000 Hz (d) 3000 Hz

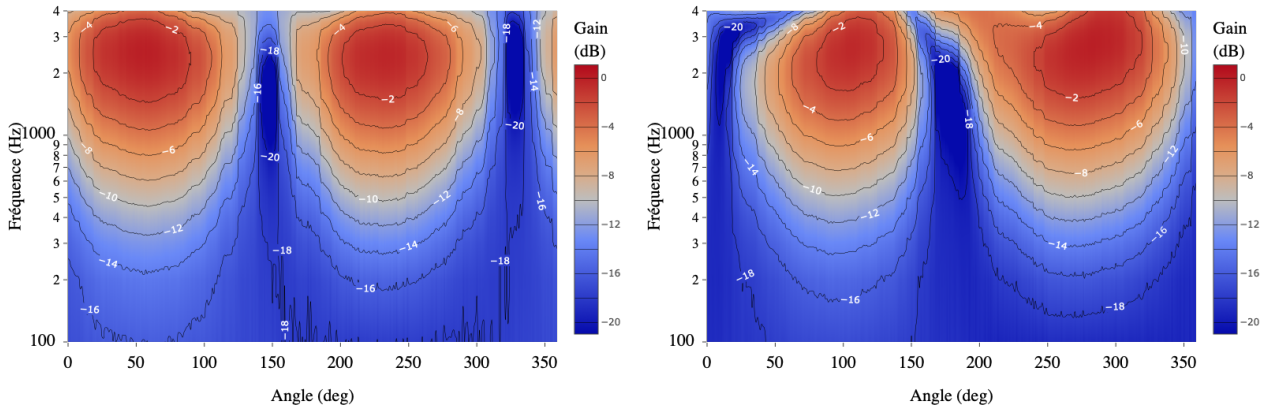
FIGURE 2.19 – Diagramme de directivité aux fréquences 500, 1 000, 2 000 et 3 000 Hz du filtre passe bas présenté en figure 2.18(a)

Bien entendu, ces deux filtres choisis pour illustration ne représentent qu'une portion des 128 filtres multicanaux de la première couche, qui offrent une diversité de directions privilégiées en haute fréquence. Cette observation permet de montrer que dès la première couche du réseau, une sélectivité spatiale est offerte au delà de 2 000 Hz. En revanche, en basses fréquence, cette couche de filtres est insuffisante pour offrir un filtrage spatial efficace, ni même une sélectivité fréquentielle importante.

La deuxième tendance observable dans la réponse des filtres de la première couche du sous-réseau de bancs de filtres par approche BeamLearning correspond à un comportement fréquentiel de type coupe-bas [8] . En effet, comme le montre la figure 2.20, contrairement aux filtres présentant des comportements passe-bas observés en figures 2.18 et 2.19, les gains maximums en sortie de ces filtres sont situés essentiellement en hautes fréquences, au delà de 1 000 Hz. Pourtant, les filtres ne peuvent pas être rigoureusement caractérisés de passe-haut, puisque leur réponse angulaire n'est pas omnidirectionnelle en hautes fréquences, ce qui signifie que la sélectivité angulaire impose des directions de coupure, y compris en hautes fréquences. Cette dépendance angulaire en haute fréquences correspond en revanche à celle observée aussi en hautes fréquences pour les filtres dits passe-bas illustrés sur les figures 2.18 et 2.19. On peut ainsi conclure de l'analyse de la première couche de filtre, que ceux-ci ont un comportement identique en haute fréquence, avec une directivité dipolaire marquée, mais différent

8. Cette appellation est privilégiée, puisqu'on ne peut parler rigoureusement de comportement passe-haut : les signaux en entrée sont tous filtrés au delà de 4000 Hz d'une part, et la directivité de ces filtres appris par le réseau impose des directions de coupure marquée, sur tout le domaine fréquentiel.

en basse fréquence.



(a) (b)

FIGURE 2.20 – Exemple de la réponse en fréquence de deux différents filtres "coupe-bas" de la première couche convolutive de BeamLearning

Le fait que les filtres multicanaux de la première couche ne présentent une directivité marquée qu'en hautes fréquences s'explique par le type de noyaux de convolutions utilisés pour ces couches. En effet, la première couche de chaque banc de filtres du sous-réseau est construite à l'aide de noyaux de convolution de longueur 3, avec un facteur de dilatation de 1. Il est donc normal que les filtres de la première couche ne puissent pas être très discriminants en basses fréquences, et correspondent globalement à des comportements fréquentiels de type passe-bas et coupe-bas. La discussion proposée en section 2.1.3.4 permet ainsi de comprendre le comportement observé de cette simple couche de filtres : pour ce type de noyau de longueur 3 et une fréquence d'échantillonnage de 44 100 Hz, pour avoir accès à au moins 25% de la longueur d'onde, il faut que la fréquence soit d'environ 3 500 Hz. Or, c'est bien aux alentours de cette fréquence qu'un comportement de directivité dipolaire apparaît pour les premières couches de chaque banc de filtre du sous-réseau.

2.3.2 Influence de la cascade de filtres multi-échelles par convolution à trous

Afin de mettre en évidence l'intérêt des filtres multi-échelles, dans cette section, nous proposons une analyse du comportement du premier banc de filtres parmi les M bancs de filtres constituant le sous-réseau. Une attention particulière est portée sur l'apport de la succession de couches convolutives à facteurs de dilatation croissants. Pour cette analyse, l'objectif est dans un premier temps de sépa-

rer l'apport des non-linéarités vis à vis des filtres. Aussi, dans le cadre de cette section uniquement, le premier banc de filtres est reconstruit en supprimant les fonctions d'activations et les couches de normalisation, mais en conservant les couches convolutives à trous, ainsi que les connexions résiduelles entre chaque couche convolutive (voir figure 2.21).

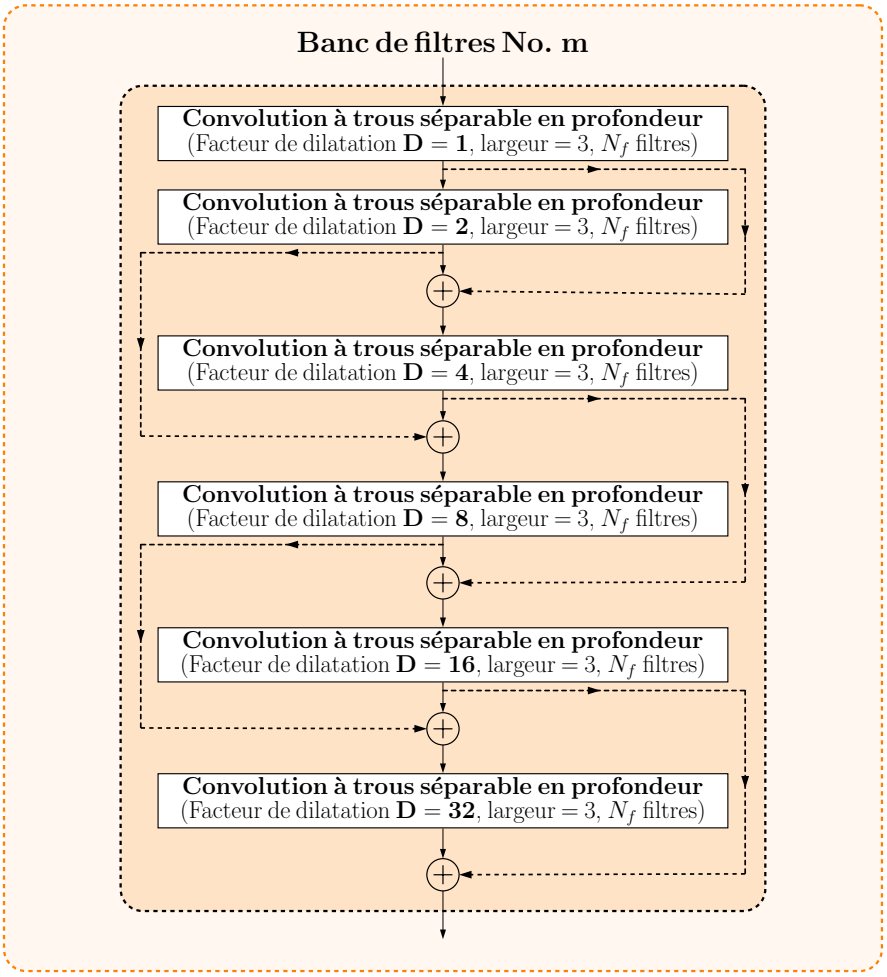


FIGURE 2.21 – Schéma de l'architecture du sous-réseau après suppression de toutes les non-linéarités présentes initialement pour la phase d'entrainement.

Bien entendu, ce réseau ne constitue pas exactement celui qui est utilisé pour la localisation de sources, puisque les couches non linéaires ont une importance primordiale dans le comportement du réseau proposé, mais cette analyse permet de mettre en exergue le comportement strictement linéaire du réseau et l'importance de l'analyse multi-résolution offerte par l'approche de bancs de filtres par

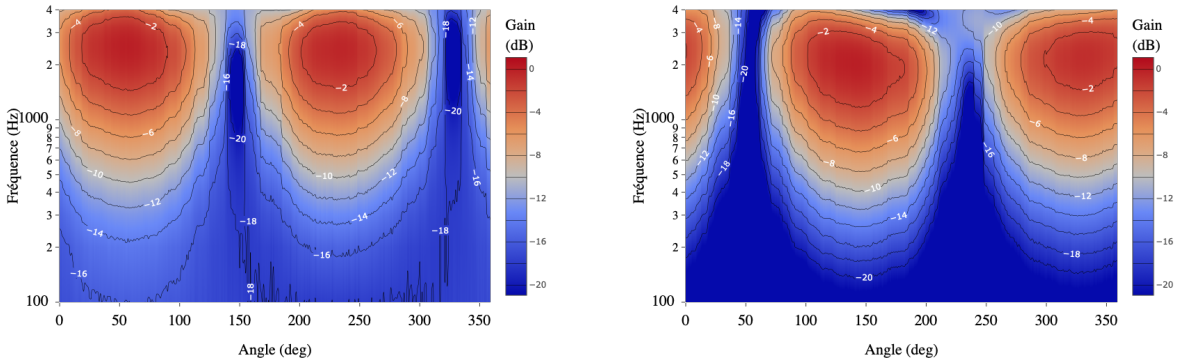
sous-réseaux de convolutions à trous que nous avons proposé.

Afin d'illustrer ce comportement, la figure 2.22 présente la réponse en termes de gain de sortie de chaque couche convolutive, dans un diagramme fréquence-angle, pour chaque couche du sous-réseau correspondant à un facteur de dilatation croissant. Il est par ailleurs essentiel de noter que pour construire ces représentations de sorties, pour chaque facteur de dilatation, la figure est produite en calculant la sortie après la cascade de couches de filtres en amont du réseau, et en conservant l'apport des connexions résiduelles. Tout comme dans la section précédente, l'objectif n'est pas ici de réaliser un catalogue des 128 filtres offerts par chaque couche convolutive à trous, mais plutôt d'en illustrer le comportement par des exemples bien choisis. Parmi ces 128 filtres pour chaque couche, tout comme pour la première couche, une partie des filtres se comportent comme des passe-bas, et les autres se spécialisent comme des filtres « coupe-bas ». Dans la figure 2.22, seuls des filtres coupe-bas sont représentés.

La comparaison des sorties de chaque couche à facteur de dilatation croissant sur la figure 2.22 permet de mettre clairement en évidence le fait que – même avec des filtres possédant très peu de coefficients non nuls (dans notre cas, 3 coefficients) – la cascade de filtres et l'utilisation de convolutions à trous permet de spécialiser les filtres en termes de sélectivité spatiale aux basses fréquences. Cette capacité est offerte grâce à l'augmentation de la longueur du filtre équivalent lorsque les signaux traversent le banc de filtre. On observe en effet clairement sur la figure 2.22 que la directivité devient de plus en plus marquée en basses fréquences lorsque le facteur de dilatation est grand et que le nombre de couches convolutives à trous traversées est grand.

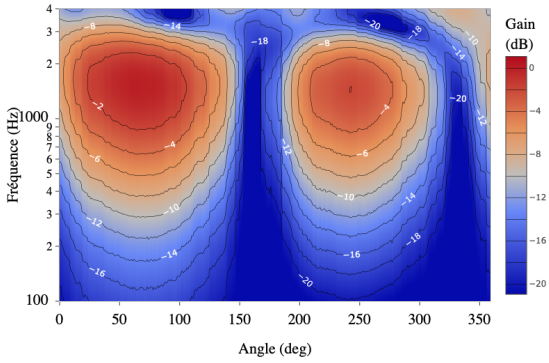
Par ailleurs, en plus d'avoir accès à une sélectivité angulaire accrue aux basses fréquences, les interconnexions résiduelles permettent de conserver une trace des directivités des couches précédentes. En effet, à partir du facteur de dilatation 8, correspondant à la figure 2.22(d), on peut observer, en plus des deux taches principales à 800 Hz, offertes par l'utilisation de noyaux de convolutions à fort facteur de dilatation, des taches de directivités à 3 000 Hz pour lesquels se sont spécialisés les noyaux de convolutions à faible facteur de dilatation des couches précédentes.

De plus, au delà de ce comportement de filtrage spatial, on peut également observer une augmentation de la sélectivité fréquentielle offerte par la cascade de filtres lorsque les signaux traversent les

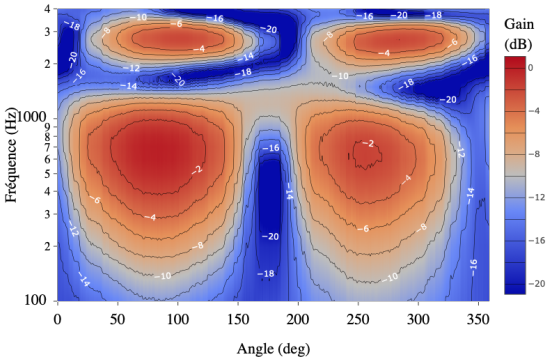


(a) facteur de dilatation : 1

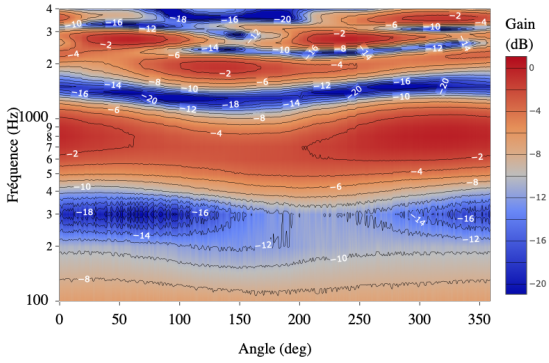
(b) facteur de dilatation : 2



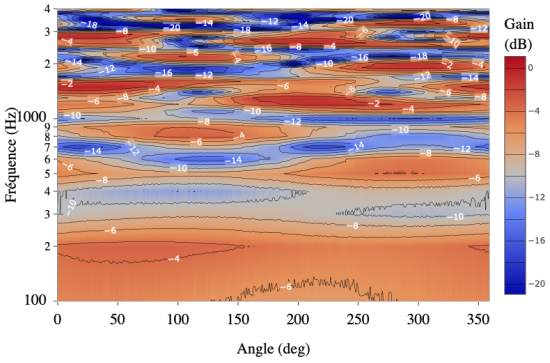
(c) facteur de dilatation : 4



(d) facteur de dilatation : 8



(e) facteur de dilatation : 16



(f) facteur de dilatation : 32

FIGURE 2.22 – Réponses en fréquence et angulaire des différentes couches correspondant à différents facteurs de dilatation du premier banc de filtre de BeamLearning. Chaque sous-figure présentée correspond à la sortie d'une des couches convolutives à trous pour un signal qui a traversé toutes les couches convolutives à trous, jusqu'au facteur de dilatation en question.

couches convolutives à facteurs de dilatation croissants. Par exemple, lorsque l'on compare les taches de directivité à 3 000 Hz entre le facteur de dilation 2 (2.22(b)) et 8 (2.22(d)), la bande passante fréquentielle à -4dB des tâches est de 2 000 Hz pour le facteur de dilatation 2, tandis que la même bande passante à -4dB vaut moins de 1 000 Hz pour le facteur de dilatation 8. D'un point de vue traitement du signal, cette sélectivité accrue correspond à une augmentation du facteur de qualité équivalent du filtre obtenu par la cascade de couches convolutives, ainsi qu'à une augmentation de l'ordre du filtre équivalent.

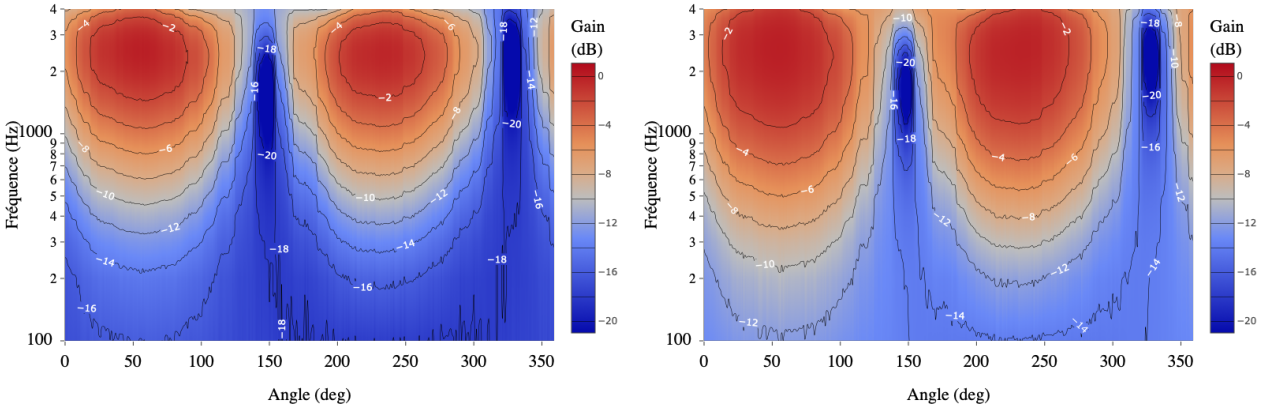
Pour finir, le comportement illustré sur la figure 2.22 démontre que la sélectivité angulaire accrue, couplée à la sélectivité fréquentielle offerte par l'approche de bancs de filtres proposée, permet à chacun des N_f filtres en cascade de présenter plusieurs directions privilégiées. Au cours de l'entraînement du réseau, chaque cascade de filtres se spécialise donc pour plusieurs bandes de fréquences et plusieurs positions angulaires. À ce titre, la figure 2.22(e) montre bien que les taches de directivités sont localisées dans des portions angulaires différentes suivant les domaines de fréquences.

L'intérêt des bancs de filtres multi-échelles est donc clairement établi. En outre, comme les filtres dont le facteur de dilatation vaut 32 (fig. 2.22(f)) permettent d'obtenir des zones d'intérêts jusqu'à la limite fréquentielle de notre modèle (100 Hz), rien ne sert d'augmenter le facteur de dilatation au delà de cette valeur. Par ailleurs, il est essentiel de noter que l'analyse présentée ici ne concerne que le premier banc de filtre parmi les M bancs de filtres successifs qui composent le sous-réseau. Cette approche permet ainsi d'offrir une sélectivité et une spécialisation de plus en plus importante lorsque les signaux traversent les M bancs de filtres.

2.3.3 Influence des opérations non linéaires dans le réseau

Au delà de leur rôle prédominant dans le mécanisme de rétropropagation des erreurs et de mise à jour des coefficients du réseau au cours de l'entraînement du réseau [39], les fonctions d'activation et les couches de normalisation présentées en section 2.1.3 permettent d'obtenir des filtres plus expressifs. La figure 2.23 permet par exemple d'illustrer le fait que ces non-linéarités permettent d'étendre le comportement de directivité des filtres sur des bandes fréquentielles plus larges. Sur cet exemple, l'étalement fréquentiel offert par l'usage des fonctions non linéaires du réseau se fait sans modification

apparente de l'allure de la fonction de directivité.



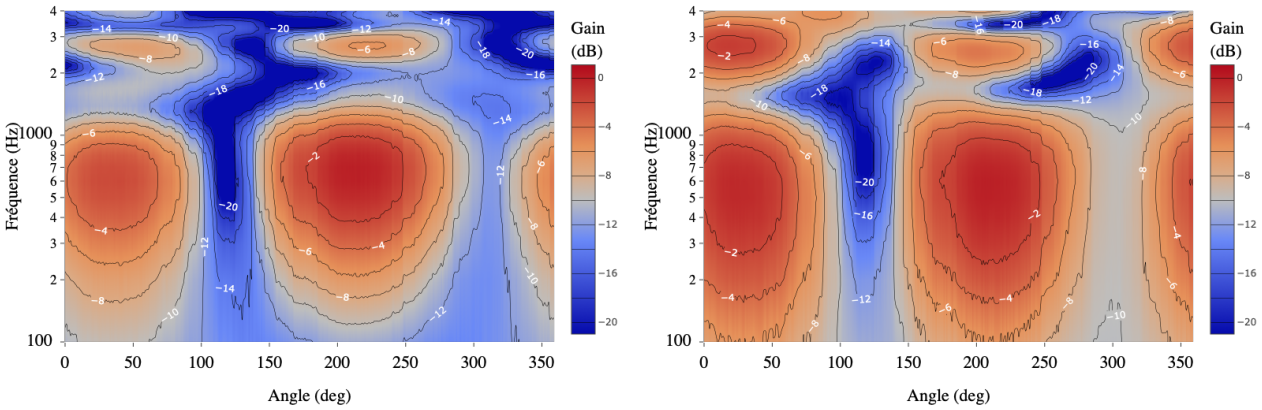
(a) Sans non-linéarités (b) Avec non-linéarités

FIGURE 2.23 – Comparaison de la réponse en fréquence d'un filtre coupe-bas de la première couche convolutive de BeamLearning sans (gauche) et avec (droite) non-linéarités appliquées

Ce constat étant réalisé pour la première couche du banc de filtres, dans la suite, on se propose de simplifier la comparaison entre les résultats des filtres obtenus avec et sans non-linéarité. Pour cela, on étudie la sortie de chaque couche convolutive à facteur de dilatation croissant, présentée avant application de la fonction d'activation, mais en conservant l'application des fonctions d'activations non linéaires, et des couches de normalisation de toutes les couches précédentes. Par conséquent, pour les figures 2.24 et 2.25 qui suivent, la sortie d'une couche m du sous-réseau est calculée à partir des données filtrées par toutes les couches précédentes $m-1, m-2, m-3...$, et de toutes les opérations non linéaires intermédiaires du réseau, jusqu'à la couche m.

Lorsque l'on compare sur la figure 2.24 la sortie des filtres au facteur de dilatation de 8 pour le premier banc de filtre, on observe un deuxième avantage que présente l'utilisation de non linéarités intermédiaires entre chaque couche convolutive : elles permettent une meilleure expressivité des filtre dans leurs zones d'intérêt. En d'autres termes, l'analyse des figures 2.24(a) et 2.24(b) révèle le fait qu'au delà de l'étalement fréquentiel illustré en figure 2.24, les zones angulaires de spécialisation de la couche neuronale présentent un gain plus élevé que si les non linéarités n'étaient pas utilisées. Cet effet est majoritairement observable pour la zone angulaire [0,100]°, avec une amélioration du gain dans la

zone d'intérêt allant de 2 à 12 dB.



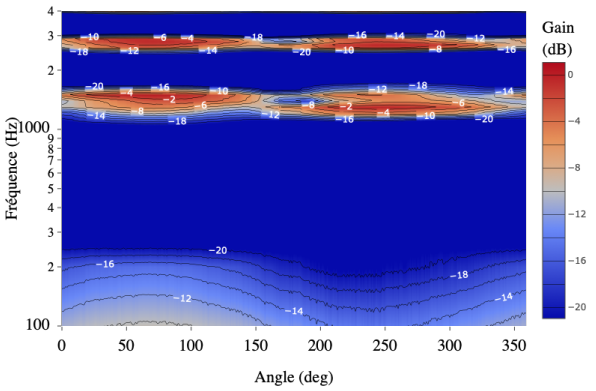
(a) facteur de dilatation : 8, sans non-linéarité

(b) facteur de dilatation : 8, avec non linéarité

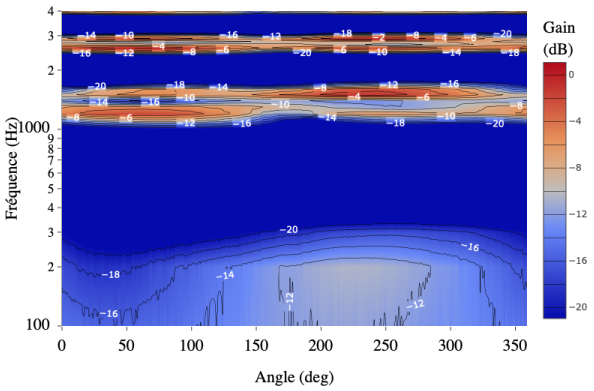
FIGURE 2.24 – Comparaison de la réponse en fréquence d'un filtres coupe-bas de la première couche convolutive de BeamLearning sans (gauche) et avec (droite) non-linéarités appliquées. Les non-linéarité de la couche à laquelle appartient le filtre ne sont toute fois pas utilisées pour une meilleur comparaison.

L'apport des opérations non linéaires au sein du réseau est d'autant plus visible que la couche étudiée est profonde. En effet, si aucune non-linéarité n'est utilisée dans le réseau, on observe une perte drastique d'informations dans les basses fréquences. Cette propriété est illustrée par la figure 2.25, qui représente le comportement en sortie de deux filtres différents parmi les 128 filtres de la dernière couche du 3e banc de filtre du réseau utilisé pour le BeamLearning, avec et sans utilisation de non linéarités du réseau. Lorsqu'aucune non linéarité n'est utilisée (figure 2.25(a) et 2.25(b)), le comportement en sortie de cette couche est certes très sélectif fréquentiellement, en ne sélectionnant que deux bandes fréquentielles étroites, à 1 500 Hz et 2 800 Hz, mais ne laissent passer que les hautes fréquences.

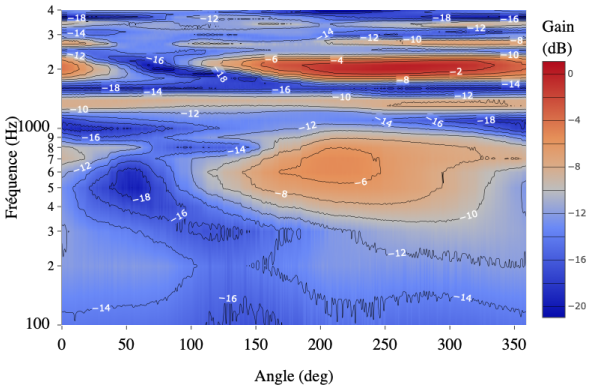
En revanche, lorsque les non-linéarités sont conservées tout au long du réseau (figure 2.25(c) et 2.25(d)), le comportement est plus complexe : plusieurs bandes de fréquences fines sont sélectionnées, et correspondent également à des zones angulaires plus sélectives que sans application de fonctions non linéaires au sein du réseau. Sans utilisation des fonctions non linéaires du réseau, les filtres ont des diagrammes de directivité majoritairement dipolaires, alors que lorsque le réseau possède ses fonctions d'activations non linéaires et ses couches de normalisation, on observe que chaque bande fréquentielle sélectionnée par cette branche du réseau se spécialise dans une direction différente.



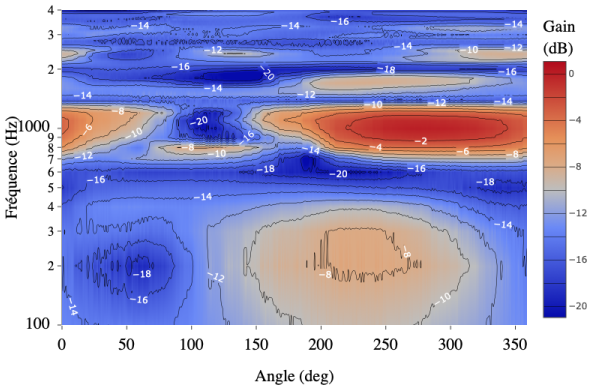
(a) Premier exemple de filtre appartenant au 3e banc de filtre, facteur de dilatation : 32, sans non-linéarité



(b) Second exemple de filtre appartenant au 3e banc de filtre, facteur de dilatation : 32, sans non linéarité



(c) Premier exemple de filtre appartenant au 3e banc de filtre, facteur de dilatation : 32, avec non-linéarité



(d) Second exemple de filtre appartenant au 3e banc de filtre, facteur de dilatation : 32, avec non linéarité

FIGURE 2.25 – Comparaison de la réponse en fréquence de deux filtres de la dernière couche convolutive de BeamLearning sans (haut) et avec (bas) non-linéarités appliquées. Les non-linéarités de la couche à laquelle appartient le filtre ne sont toute fois pas utilisées pour une meilleure comparaison.

Bien entendu, l'illustration proposée sur la figure 2.25 ne représente que deux filtres parmi les 128 filtres construits par le réseau au cours de son entraînement, ce qui permet de montrer le comportement complexe, mais relativement intuitif développé par ce sous-réseau de bancs de filtres. En traversant les couches des M bancs de filtres, 128 filtres sélectifs en fréquences et en angles permettent de construire 128 canaux dans un nouvel espace de représentation, pour lesquels l'énergie est maximisée

dans plusieurs *bins* espace-fréquence. Par combinaison de l'énergie portée par ces 128 canaux dans le sous-réseau suivant, cela permet donc d'ouvrir la voie à une localisation de sources variées, quel que soit le profil de leur spectre émis.

2.4 Synthèse de l'approche BeamLearning

L'approche BeamLearning présentée dans ce chapitre est constituée de quatre sous parties qui ont chacune leur utilité propre :

— Les données d'entrées sont des signaux microphoniques regroupées par tranche temporelles d'une longueur de 1 024 échantillons. Ils peuvent être regroupés par lots (mini-batch) lors de la phase d'apprentissage pour permettre de calculer une statistique sur les erreurs d'estimations et ainsi permettre d'optimiser les variables d'apprentissages au cours de l'entrainement.

— Les bancs de filtres sont une succession de convolutions à trous séparables en profondeur, dont les facteurs de dilatation varient pour obtenir de l'information à différentes échelles temporelles (ou fréquentielle, selon le point de vue) du signal. Ces convolutions à trois points sont entrecoupées de tangentes hyperboliques et de fonctions de normalisation, pour assurer une meilleure rétropropagation des gradients, lors de la phase d'optimisation.

— Une partie quasi-déterministe convertit les données filtrées, en données énergétiques pour se rapprocher des grandeurs recherchées par les algorithmes reposant sur des modèles.

— La sortie du réseau, qui peut être aussi bien une approche par classification, qu'une approche par régression, suivant l'application visée. Toutefois, si une précision importante est demandée pour la localisation de sources acoustiques, alors l'approche régression est à préférer, même si cette approche demande un période d'apprentissage plus gourmande en termes de temps de calcul (les durées caractéristiques des apprentissages sont présentées au chapitre 5)

Enfin, une analyse originale des variables d'apprentissage a été menée dans ce chapitre, qui justifie

l'architecture du réseau profond proposé, en particulier de sa partie bancs de filtres.

L'optimisation de ces variables d'apprentissage requiert un nombre important de données parfaitement étiquetées. En effet, c'est à partir de ces étiquettes qu'une fonction de coût est calculée pour caractériser l'erreur d'estimation de l'approche BeamLearning, et que les gradients d'erreurs sont rétropropagés à travers les couches du réseau. La méthodologie employée pour constituer ces jeux de données d'entrainement est donc proposée dans les deux chapitres suivants, tant pour des jeux de données simulées numériquement (chap. 3), que pour des jeux de données expérimentales (chap. 4).

Chapitre 3

Création de bases de données multicanales en environnement réverbérant obtenues par simulations numériques

Tant que les lois mathématiques renvoient à la réalité, elles ne sont pas absolues, et tant qu'elles sont absolues, elles ne renvoient pas à la réalité. (Discours à l'Académie Scientifique de Prusse, Einstein, Janvier 1921)

Contenu du chapitre

3.1 Modélisation numérique de réponses impulsionnelles de salles

La modélisation des champs de pression en acoustique des salles est à l'origine d'un grand nombre de travaux, notamment sur les approches numériques permettant de modéliser ces champs [120–122]. Les méthodes proposées dans la littérature scientifique reposent essentiellement sur trois types d'approches qui peuvent être considérées comme complémentaires : l'approche statistique, l'approche ondulatoire et l'approche géométrique. Pour chacune de ces approches de modélisation, des optimisations et une littérature riche continuent d'être proposées par la communauté scientifique. La classification de ces méthodes n'est toutefois pas aisée ; sans prétendre en faire ici une liste exhaustive, cette introduction présente les grands principes des méthodes les plus utilisées.

La catégorie des approches statistiques fait l'hypothèse que le champ de pression dans la pièce peut être vu comme résultant d'un grand nombre de réflexions sur les parois d'une onde émise par une (ou plusieurs) sources dans la salle. Lorsque la densité temporelle des réflexions est suffisante, le champ est caractérisé de champ diffus. Ce champ diffus apparaît après le « temps de mélange » caractéristique de la salle [123]. Cette hypothèse de champ diffus permet alors de traiter le champ d'un point de vue statistique, en supposant qu'il se présente sous la forme d'un signal aléatoire et statistiquement homogène dans la salle. La pression en un point est alors considérée comme une somme d'ondes planes décorrélées, provenant de toutes les directions de l'espace et ayant la même amplitude. Sous cette approche, aucune des réflexions n'est traitée en particulier. Au contraire, les caractéristiques du champ de pression sont estimées à partir de la théorie probabiliste. Cette approche a donné naissance en particulier aux méthodes de Sabine et d'Eyring qui seront développées plus loin dans ce manuscrit, en section 3.1.5.

Une autre approche consiste à exploiter la théorie ondulatoire, en utilisant un calcul basé sur la résolution de l'équation des ondes dans la salle en y adjoignant les conditions aux limites sur les parois délimitant le domaine de la salle étudiée. Ces méthodes modales, initialement introduites d'un point de vue analytique et théorique par Van Den Dungen, et reprises par Morse [124], ont par la suite été exploitées pour développer des méthodes numériques. Même si ces méthodes sont très performantes pour l'analyse modale des espaces clos, elles sont limitées aux premiers modes [125]. Celles-ci peuvent

être exploitées tant dans le domaine fréquentiel, avec la méthode des éléments finis [126] ou la méthode des éléments de frontières, que dans le domaine temporel, en exploitant des schémas aux différences finies temporels et spatiaux [127]. L'utilisation de codes de différences finies est néanmoins rare en acoustique à cause du coût computationnel en hautes fréquences, même si des améliorations ont été proposées il y a quelques années [128].

L'approche de l'acoustique géométrique en acoustique des salles, utilisée pour cette thèse de doctorat, exploite quant à elle en partie l'analogie entre acoustique et optique, en généralisant la notion de chemin d'un rayon lumineux, ou d'image d'une source sous l'effet d'une paroi réfléchissante. Cette méthode est la plupart du temps exploitée là où les approches modales ne représentent plus un bon compromis entre précision et temps de calcul. En basses fréquences, des raffinements ont également été proposés pour les méthodes exploitant l'approche d'acoustique géométrique, permettant ainsi d'obtenir des résultats satisfaisants [122]. Pour une partie des méthodes d'acoustique géométrique, par analogie avec les rayons lumineux, les rayons sonores sont des droites perpendiculaires aux fronts d'ondes se déplaçant à vitesse constante. Il en découle les concepts de puissance incidente et de coefficient d'absorption, qui seront présentés plus loin dans ce manuscrit de thèse (voir sec. 3.1.1).

Cette vision du problème de propagation donne en particulier naissance à deux classes de méthodes d'acoustique géométrique. La première est celle du tracé (ou tir) de rayons, consistant à simuler le trajet d'un grand nombre de rayons [1] , portant chacun une intensité propre, et émis dans toutes les directions depuis les sources présentes dans la salle. Le chemin emprunté par ces rayons jusqu'au récepteur est en particulier impacté par les réflexions multiples sur les parois, chacune de ces réflexions entraînant une diminution de l'intensité portée par le rayon d'un facteur correspondant au coefficient d'absorption des surfaces rencontrées. La seconde est appelée la méthode des sources images. Cette méthode consiste à considérer que chaque réflexion sur une paroi d'une onde acoustique émise par une source correspond à la contribution d'une source dite *image*, dont la position est le symétrique de la source primaire par rapport à la paroi. Ainsi, dans une salle fermée, un ensemble de sources images est calculé par symétries successives par rapport à chaque paroi. Toutefois, cette méthode suppose que les réflexions sont spéculaires, et ignore donc tout effet de diffusion ou de diffraction, tout comme un

1. rayons, cônes, ou pyramides, suivant les implémentations numériques et raffinements apportés à cette approche

grand nombre d'approches numériques en acoustique des salles. Une discussion plus approfondie sur ce point est menée en section 3.1.7.

3.1.1 Réflexion d'une onde plane sur une paroi en incidence normale

Soit une onde plane se propageant dans la direction des x croissants, définie par sa pression $p(x,t)$ et sa vitesse particulaire orientée $v(x,t)$, portée par l'axe x. On a par définition [129] :

$$\begin{cases} p(x,t) = p_0 \ e^{j(\omega t - kx)} \\ v(x,t) = \frac{p_0}{\rho_0 c} \ e^{j(\omega t - kx)} \end{cases} \tag{3.1}$$

La présence d'une paroi verticale partiellement réfléchissante dans le plan yOz (fig. 3.1) génère donc un phénomène de réflexion. En notant R le coefficient de réflexion en pression, il vient pour la pression de l'onde réfléchie $p_r(x,t)$ et pour sa vitesse $v_r(x,t)$ [130] :

$$\begin{cases} p_r(x,t) = R \ p_0 \ e^{j(\omega t + kx)} \\ v_r(x,t) = -R \ \frac{p_0}{\rho_0 c} \ e^{j(\omega t + kx)} \end{cases} \tag{3.2}$$

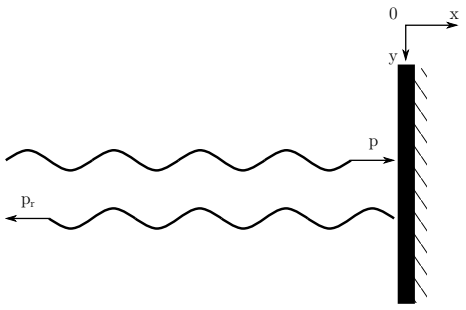


FIGURE 3.1 – Réflexion d'une onde plane sur une paroi

La pression et la vitesse particulaire résultantes à la paroi ($x = 0$) sont donc la somme des deux expressions associées aux ondes incidente et réfléchie. Soient $p_{tot}(x,t)$ et $v_{tot}(x,t)$ la pression et la

vitesse de l'onde résultante. On a, en combinant les équations 3.1 et 3.2 :

$$\begin{cases} p_{tot}(0,t) = (1+R)\ p_0\ e^{j(\omega t)} \\ v_{tot}(0,t) = (1-R)\ \frac{p_0}{\rho_0 c}\ e^{j(\omega t)} \end{cases} \tag{3.3}$$

Une manière courante de décrire le comportement d'une paroi en acoustique des salles est de définir son impédance Z. L'impédance de la paroi est de manière générale complexe, et peut également varier avec l'angle d'incidence de l'onde, dans le cas des parois à réaction non localisée. Sans perte de généralité, l'impédance de la paroi sera par la suite supposée comme celle d'une paroi à réaction localisée. Cette grandeur est définie comme le rapport entre la pression pariétale et la vitesse normale à la paroi v_n, orientée vers l'intérieur du volume de la salle :

$$Z = \frac{p_{tot}(0,t)}{v_n(0,t)} = \rho_0 c \frac{1+R}{1-R} \tag{3.4}$$

On a donc pour un mur parfaitement réfléchissant ($R = 1$) une impédance infinie ($Z = \infty$). Au contraire, pour une paroi parfaitement absorbante, $R = 0$ et l'impédance est égale à $\rho_0 c$. La notion qui découle naturellement de cette dernière remarque est l'impédance acoustique spécifique, définie comme le rapport entre Z et l'impédance caractéristique de l'air $\rho_0 c$:

$$\xi = \frac{Z}{\rho_0 c} = \frac{1+R}{1-R} \tag{3.5}$$

Enfin, il est possible de définir le coefficient d'absorption en énergie α comme le rapport entre l'énergie incidente et réfléchie :

$$\alpha = 1 - |R|^2 = \frac{4Re(\xi)}{|\xi|^2 + 2Re(\xi) + 1} \tag{3.6}$$

3.1.2 Réflexion sur une surface à réaction localisée en incidence oblique

Que ce soit pour des parois à réaction localisée ou non localisée, le coefficient de réflexion R est une fonction dépendant de l'angle d'incidence. D'après les lois de Snell-Descartes, et en supposant que les réflexions sont spéculaires, l'angle d'incidence θ_i est égal à l'angle réfléchi θ_r. On pose alors $\theta = \theta_i = \theta_r$. Le coefficient de réflexion vaut alors :

$$R(\theta) = \frac{\xi cos(\theta) - 1}{\xi cos(\theta) + 1} \tag{3.7}$$

Le coefficient d'absorption en énergie s'en trouve lui aussi modifié et vaut de ce fait :

$$\alpha(\theta) = \frac{4Re(\xi)cos(\theta)}{(|\xi|cos(\theta))^2 + 2Re(\xi)cos(\theta) + 1} \tag{3.8}$$

Dans le cas général, le coefficient d'absorption en énergie utilisé est plutôt un coefficient moyen α_{moy}, quantifiant l'effet statistique de α pour toutes les incidences. Pour calculer α_{moy}, on suppose que les ondes incidentes ont des amplitudes distribuées uniformément sur toutes les directions d'incidences possibles. Ainsi, chaque angle solide $d\Omega$ contient la même intensité acoustique I. Les phases étant supposées elles aussi distribuées de manière aléatoire, l'énergie des ondes se somme alors simplement. Soit $Idsd\Omega$ l'énergie sonore par seconde arrivant sur un élément infinitésimal de surface ds et provenant de l'angle solide $d\Omega$. L'énergie totale arrivant par seconde sur la surface ds vaut :

$$E_i = Ids \int_0^{2\pi} d\phi \int_0^{\frac{\pi}{2}} cos(\theta)sin(\theta)d\theta = \pi Ids \tag{3.9}$$

Le flux d'énergie qui se trouve absorbé par la paroi vaut de la même manière :

$$E_a = Ids \int_0^{2\pi} d\phi \int_0^{\frac{\pi}{2}} \alpha(\theta) cos(\theta) sin(\theta) d\theta \tag{3.10}$$

Le coefficient d'absorption moyen est alors le rapport entre les deux précédentes équations, soit :

$$\alpha_{moy} = \frac{E_a}{E_i} = \int_0^{\frac{\pi}{2}} \alpha(\theta) sin(2\theta) d\theta \tag{3.11}$$

Cette expression est aussi connue sous le nom de *formule de Paris*. La valeur de α_{moy} peut enfin être déterminée si nécessaire en fonction de ξ en utilisant l'équation 3.8.

3.1.3 Réflexion d'une source sur une paroi en incidence oblique : le concept de source image

Considérons un microphone placé au point M à une distance d_0 d'un mur réfléchissant, et une source sphérique de pulsation harmonique $\omega = 2\pi f$ se situant à une distance A du mur et d du microphone (voir schéma 3.2). Dans ce cas, l'approximation de la source image consiste à exprimer la pression mesurée au point M comme la somme de l'onde incidente ayant parcourue une distance d et de l'onde réfléchie ayant parcouru une distance d_r correspondant à la distance entre le symétrique de la source par rapport à la paroi et le microphone :

$$p_{tot}(M,t) = p_0 \left(\frac{e^{j\omega(t-\frac{d}{c_0}))}}{d} + \frac{R \cdot e^{j\omega(t-\frac{d_r}{c_0}))}}{d_r} \right) \tag{3.12}$$

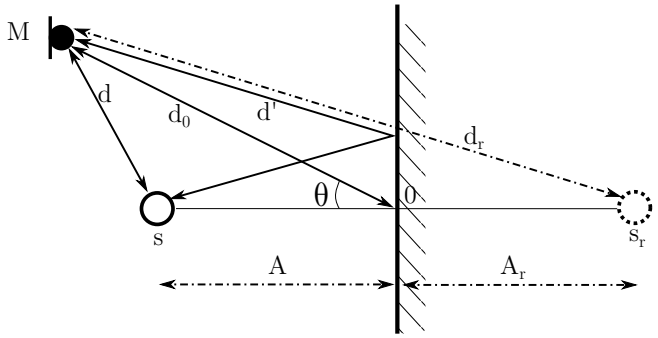


FIGURE 3.2 – Schéma de positionnement du microphone (M) de la source (s) et de la source image (s_r)

Il est essentiel de noter que l'expression 3.12 est parfaitement valide pour les parois infinies à impédance infinie ou nulles, mais qu'elle représente une approximation dans le cas des parois finies absorbantes [131]. Dans ces derniers cas, la solution exacte peut être calculée rigoureusement dans le formalisme de Green à l'aide d'une intégrale de contour, mais elle souffre d'une faible convergence. L'analyse proposée par Mechel dans [131] permet en revanche de démontrer que l'approximation de source image reste valide lorsque la source et le récepteur sont tous les deux positionnés à une distance supérieure à la longueur d'onde par rapport à la paroi.

Sous cette approximation, le champ mesuré est donc simplement calculé comme la superposition de signaux provenant de la source S primaire d'une part, et de la source image Sr d'autre part, sans le mur, mais dont l'amplitude serait multipliée directement par le coefficient de réflexion R (cf. fig 3.2) : c'est le principe de la théorie de la source image [122, 130, 132]. Les distances parcourues par les deux ondes sont facilement calculables :

$$\begin{cases} d = \sqrt{d_0^2 + A^2 - 2d_0 A\cos(\theta)} \\ d_r = \sqrt{d_0^2 + A^2 + 2d_0 A\cos(\theta)} = d' \end{cases} \tag{3.13}$$

On peut alors vérifier que cette modélisation permet de respecter en particulier la conditions de Dirichlet avec $R = 1$, puisque la pression est dans ce cas maximale sur la paroi, et que la composante

normale de la vitesse particulaire sur la paroi est au contraire nulle. Il est en outre possible de modéliser des conditions de Neuman, même si ce cas de figure est moins souvent utilisé en acoustique des salles, en utilisant un coefficient de réflexion à la paroi négatif, ce qui modélise une source en opposition de phase. La pression est donc minimale à la paroi et la vitesse maximale. Même si cette approche permet de modéliser un coefficient de réflexion complexe pour décrire les réactions localisées en incidence oblique de la paroi, le plus souvent, le coefficient utilisé n'est qu'un réel dont la valeur absolue est calculée à partir du coefficient d'absorption moyen défini à la section 3.1.2 :

$$|R| = \sqrt{1 - \alpha_{moy}} \tag{3.14}$$

Ce raisonnement peut être étendu au cas général d'un problème à multiples sources et/ou multiples surfaces, où chaque surface est remplacée par un ensemble de sources images.

3.1.4 Généralisation des sources images dans une pièce parallélépipédique

Dans le cas où il n'y a non plus une seule paroi réfléchissante mais plusieurs, l'onde acoustique subit de multiples réflexions sur les parois, chaque réflexion spéculaire donnant naissance à une nouvelle source image. Les sources images sont ensuite ordonnées en fonction du trajet du rayon acoustique dans la salle : s'il s'agit de la première réflexion sur une paroi, on parlera d'une *source de 1e ordre*. De même, si le rayon acoustique s'est déjà réfléchi une première fois sur une paroi, lors de la deuxième réflexion, la source image crée sera appelée *source de 2e ordre*, et ainsi de suite. Un ordre 2 correspond donc à deux réflexions successives, ce qui revient à symétriser une source image d'ordre 1 par rapport au mur correspondant. Cette symétrie permet alors d'accélérer le calcul des positions des sources images, puisque les chemins acoustiques n'ont plus à être tracés. Seuls les symétries sont calculées à partir de la géométrie de la pièce et de la position initiale de la source. Ce procédé est répété jusqu'à ce que l'intensité acoustique des sources images tombent en dessous d'un certain seuil, ou lorsque l'ordre de réflexion maximal voulu soit atteint. Ainsi, dans le cas de la salle parallélépipédique, du fait de la construction par symétries successives des sources images, toutes les sources sont visibles quelque soit la position du récepteur. La position des sources peut donc être calculée indépendamment de celle du récepteur dans la pièce, contrairement aux salles de géométries plus complexes, pour lesquelles des

stratégies de recherche de validité de chemins doivent être réalisées pour écarter les sources images qui ne contribueraient pas au champ mesuré au point de réception. Le schéma 3.4 présente la généralisation de la théorie des sources images à une pièce en deux dimensions.

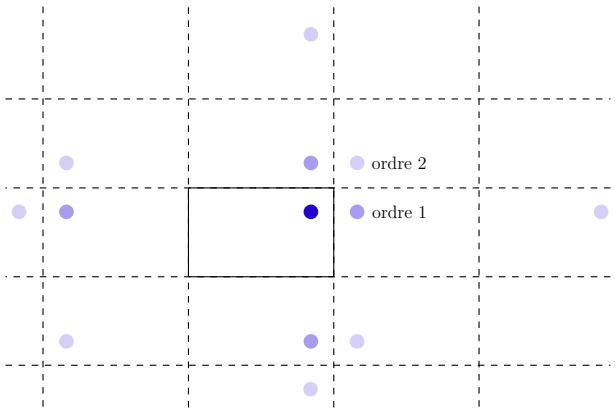


FIGURE 3.3 – Schéma des sources images d'une source primaire (bleu) dans une pièce. Seuls les ordres 1 (bleu pâle) et 2 (bleu très pâle) sont représentés

Pour chaque réflexion sur une paroi, l'onde acoustique perd une certaine quantité d'énergie, proportionnelle au coefficient d'absorption de la paroi. Comme expliqué précédemment en section 3.1.3 le coefficient de réflexion en pression est le plus souvent simplifié par un réel défini à partir du α_{moy}, ce qui signifie que les effets de déphasages sur la paroi et la dépendance angulaire sont négligés. De plus, il est possible de faire varier α en fonction de la fréquence, généralement en bande d'octave. Dans la suite, pour alléger les notations, l'indice $_{moy}$ sera omis pour les coefficient d'absorption en énergie, ainsi que la dépendance fréquentielle.

Dans le cas où différents matériaux sont utilisés sur le parois de la pièce, différents coefficients doivent être pris en compte. La pression acoustique est donc réduite à chaque réflexion d'un facteur correspondant à la surface en question. En utilisant l'indice i pour faire référence à la ième surface de la pièce, et n_i le nombre de fois où l'onde est réfléchie sur la surface i, on peut donc exprimer la contribution en pression de la source image positionnée en Sr, mesurée par le microphone positionné en M comme :

$$p_{s_r,M}(t) = \frac{\prod_i (\sqrt{1-\alpha_i})^{n_i}}{d_r} s(t - \frac{d_r}{c_0}) \tag{3.15}$$

Pour la méthode des sources images, comme discuté en section 3.1.3, chacun de ces champs est une solution exacte pour des parois à impédance nulle ou infinie, et représente une approximation satisfaisante dans les autres cas, à condition que la source soit à une distance d'au moins une longueur d'onde de la paroi. Le champ total au point M est alors approximé en calculant la somme de la contribution de toutes les sources images impliquées, jusqu'à un ordre de troncature qui sera discuté en section 3.1.6. Afin d'améliorer l'approximation y compris en basses fréquences, il est possible de remplacer rigoureusement le coefficient de réflexion de chaque paroi par un coefficient de réflexion fictif, permettant de minimiser l'erreur commise lorsque la source ou le récepteur sont proches des parois [131].

3.1.5 Détermination du coefficient d'absorption dans le cas général

Afin de déterminer expérimentalement le comportement d'absorption de matériaux comme la variable d'entrée des méthodes numériques reposant sur le principe des sources images, il existe plusieurs méthodes [130, 132, 133].

La première consiste à introduire un échantillon de matériau dans un tube d'impédance, généralement connu sous le nom de tube de Kundt. Dans ce tube fermé, qui se comporte comme un guide d'onde 1D, une onde incidente générée par un haut-parleur est partiellement réfléchie par un échantillon du matériau à tester. Les interférences entre l'onde incidente et réfléchie produisent une onde quasi stationnaire. En mesurant les caractéristiques de ces ondes quasi stationnaires, il est possible de déterminer le coefficient d'absorption du matériau testé. Cependant, il est essentiel de noter que cette approche n'exploite des réflexions qu'en incidence normale, ce qui signifie que le matériau n'a pas le même comportement que s'il était exposé à des incidences variées, comme c'est le cas en acoustique des salles.

La deuxième méthode repose elle aussi sur l'estimation du coefficient d'absorption en mesurant la

somme du champ de pression réfléchi et incident, mais cette fois ci en environnement 3D anéchoïque, ouvrant ainsi la voie à l'analyse de la dépendance angulaire du comportement de réflexion d'un matériau. Dans ce cas, plusieurs approches ont été proposées dans la littérature pour séparer le champ réfléchi du champ total mesuré par un doublet microphonique, afin d'en déduire l'impédance du matériau en incidence quelconque, ainsi que son coefficient de réflexion complexe [134, 135].

La troisième méthode exploite quant à elle les équations de l'acoustique statistique, valides en champ diffus. Pour cela, les formules de Sabine ou d'Eyring, qui relient le coefficient d'absorption à la durée de réverbération Tr de la pièce, sont inversées. Une simple comparaison des mesures par bandes fréquentielles de cette durée de réverbération dans une salle réverbérante lorsqu'un échantillon de matériau est présent ou non dans la salle permet ainsi de déterminer le coefficient d'absorption du matériau testé. Cette approche ne permet bien entendu pas de déterminer l'impédance complexe du matériau, ni même d'obtenir le coefficient de réflexion en fonction de l'angle d'incidence, mais elle est pourtant exploitée dans la norme européenne [136], puisqu'elle est plus proche du comportement *in situ* du matériau. Par ailleurs, cette estimation d'un coefficient d'absorption moyen permet d'utiliser une valeur de α ou de R ne dépendant pas de l'angle d'incidence, qui soit exploitable simplement par une méthode d'acoustique géométrique, même si les auteurs de [137] ont prouvé que la prise en compte de l'angle d'incidence dans le coefficient d'absorption a un effet non négligeable sur la durée de réverbération simulée de la pièce.

Dans le cas de la simulation de la réponse d'une salle avec la méthode des sources images, le choix du coefficient d'absorption de chaque mur est primordial. Il est en revanche important de noter que la méthode de sources images ne modélise pas l'influence des possibles meubles et personnes dans la pièce (voir section 3.1.7.3). De plus, l'approche classique ne permet d'obtenir qu'une approximation assez grossière du coefficient R à utiliser pour chaque paroi, ce qui ne garantit pas la minimisation de l'erreur introduite par l'approximation sources-images. Comme discuté par Mechel [131] avec des solutions analytiques, il est en revanche possible de remplacer rigoureusement le coefficient de réflexion des parois par un coefficient qui minimise les effets de cette approximation. Dans le cadre de cette thèse, nous proposons une approche permettant de déterminer itérativement une modification optimale des coefficients d'absorption des parois pour modéliser au mieux la décroissance énergétique du champ.

La section 3.5 de ce manuscrit sera dédiée à cette optimisation.

3.1.6 Troncature de la réponse impulsionnelle et sélection de sources images

Comme exposé précédemment dans le cadre de la méthode des sources images, la réponse impulsionnelle d'une salle (RIR) est, par construction calculée comme la somme des contributions de chaque source image. Elle dépend donc à la fois de la position de la source sonore et du microphone. D'un point de vue théorique, cette réponse impulsionnelle est infinie, mais tend vers 0 d'autant plus rapidement que l'absorption des parois est importante. Dans le cas d'une source ponctuelle placée dans une salle parallélépipédique, il est possible d'obtenir la RIR exacte par une superposition des contributions d'un nombre infini de sources. Cette RIR correspond alors à la solution exacte de l'équation des ondes obtenue par la théorie modale [1].

En pratique, compte tenu de cette décroissance énergétique naturelle, un critère d'arrêt est fixé pour déterminer le nombre de sources images à prendre en compte dans ce calcul. Généralement, ce critère est basé sur une approche énergétique [138] en considérant que seules les sources ayant une énergie suffisante contribuent à la RIR. Une seconde approche de sélection des sources consiste à exploiter le problème temporel, en ne sélectionnant que les sources images dont la distance au récepteur peut être parcourue en un temps donné. Or, le nombre de source images croît exponentiellement avec l'ordre de réflexion des sources images à calculer, ce qui peut impacter lourdement le temps de calcul de ces réponses impulsionnelles [130]. Aussi, la plupart des auteurs proposent de n'utiliser la méthode des sources images que pour la partie précoce de la RIR, et que la partie *tardive* caractérisant un champ *perceptivement diffus* soit calculé grâce à des méthodes statistiques [138]. Mais dans le cadre de cette thèse de doctorat, la méthode des sources images est utilisée pour l'ensemble de la RIR, comme dans les travaux de E. Lehmann [139].

Compte tenu du grand nombre de réponses impulsionnelles à calculer pour générer des jeux de données conséquents et adaptés à une approche d'apprentissage, nous proposons donc de réaliser ce calcul de sources images en exploitant la puissance de calcul offerte par les processeurs graphiques (GPU) et les librairies Tensorflow [2] . On lève ainsi la limitation du temps de calcul communément

2. ici utilisées pour du calcul scientifique déterministe par lots, et non pour du calcul d'apprentissage

reprochée à ce type de méthodes. Pour cela, un travail particulier a été fait sur la précision temporelle des calculs en section 3.3, afin de garantir une validité numérique des retards, que ce soit pour la partie précoce ou la partie tardive des réponses impulsionnelles.

Dans le cadre de cette thèse de doctorat, le critère utilisé pour tronquer la série de sources images contribuant au champ simulé est la durée de réverbération (T_R) cible de la salle. Ainsi la réponse impulsionnelle de la salle est :

$$h_{RIR}(t) = \sum_{s_r} p_{s_r,M}(t) \mid dist(s_r, M) < c_0 \cdot t_{max} \tag{3.16}$$

Dans des géométries plus complexes, comme les salles de concert, des techniques de test de « visibilité » de sources images sont utilisées pour minimiser le nombre de sources images à utiliser dans le calcul des réponses impulsionnelles [138]. Dans notre cas, dans l'implémentation du calcul, un très grand nombre de sources images sont calculées, puis les sources images sont triées par distance au microphone, et seules celles répondant au critère de distance sont conservées.

3.1.7 Limites de la méthode des sources images

La méthode des sources images est certes relativement simple à mettre en œuvre, mais il est essentiel de noter qu'elle repose sur des hypothèses qui peuvent montrer des limites dès lors que l'on veut une représentation précise et spatialement fidèle de la salle. En particulier, la complexité de la géométrie de la salle entraîne à la fois une augmentation du nombre de sources images à calculer, mais aussi demande une vérification à *posteriori* de la visibilité des sources images par le récepteur [138]. C'est la raison pour laquelle nous nous sommes limités dans ce manuscrit, à illustrer les capacités de la méthode BeamLearning dans des salles parallélépipédiques.

3.1.7.1 Domaine de validité fréquentielle

Comme expliqué précédemment, l'approche modale en acoustique des salles peut être extrêmement gourmande en terme de ressources de calcul, et perd grandement de son intérêt dès que la densité modale est trop importante. Lorsque celle-ci atteint une valeur telle qu'il n'est plus pertinent d'individualiser les modes propres d'une salle, on exploite généralement l'approche statistique.

La limite fréquentielle communément utilisée entre ces deux approches est appelée la fréquence de Schroeder, définie à partir la durée de réverbération de la pièce T_r, et de son volume V :

$$f \geq 2000\sqrt{\frac{Tr}{V}} \quad \text{(Hz)} \tag{3.17}$$

Dans le cadre de l'acoustique géométrique, en particulier pour la théorie des rayons [130], une hypothèse est posée, cette fois ci sur le caractère d'ondes planes. Cette hypothèse est elle aussi non valide en basses fréquences, mais pour des raisons différentes que l'hypothèse de champ diffus. Ici, l'approximation ondes planes repose surtout sur le fait que la longueur d'onde soit petite devant le trajet effectuée par le son. En revanche, dans le cadre de la théorie des sources images, cette limite est contournée, puisqu'une fois la position des sources images calculées, chacune est définie comme une source qui peut être ponctuelle et à rayonnement sphérique. En effet, l'une des différences fondamentales entre l'approche tir de rayons et l'approche sources images, est que l'hypothèse onde plane est levée dans le second cas. De plus, si le nombre de sources images est suffisamment grand, la réponse calculée pour la salle tend vers la solution exacte de la théorie ondulatoire, comme expliqué en section 3.1.6. En revanche, dans le domaine des basses fréquences, la convergence peut demander un nombre important de sources images et des raffinements sur le coefficient d'amplitude associé à chaque source image, notamment lorsque la source principale et le récepteur sont proches des parois. Dans ces travaux de thèse de doctorat, le nombre de sources choisies est très grand, donc aucune limitation basse fréquence n'est à prendre en compte à cause de cette approximation.

En revanche, comme discuté précédemment, le rapport entre la distance H de la source (ou du récepteur) au mur et la longueur d'onde λ est une grandeur limitante. Mais si la source (ou le récepteur) est suffisamment loin de la paroi ($H/\lambda > 2$) alors l'erreur liée à l'approximation est inférieure à 10% [131]. Dans le cadre des configurations proposées dans le cadre de cette thèse pour les bases de données simulées numériquement, cette condition est respectée, sauf dans quelques rares cas, pour la partie basse du domaine de définition fréquentiel, entre 100 Hz et 400 Hz environ.

De plus, l'utilisation d'un module plutôt qu'un nombre complexe pour le coefficient de réflexion en pression peut avoir une influence. En effet, en ignorant le déphasage introduit à la réflexion de la

paroi, on approxime le coefficient de réflexion en pression comme une grandeur réelle pure, ce qui peut avoir un impact sur la précision de la réponse impulsionnelle calculée, notamment en basses fréquence, comme montré dans [140].

Enfin, il est également nécessaire de préciser que la méthode des sources images ne permet pas de modéliser les phénomènes de diffraction et de diffusion, impactant ainsi potentiellement la modélisation en hautes fréquences. Une discussion spécifique sur ce point est proposée en section 3.1.7.3.

3.1.7.2 Dépendance fréquentielle et angulaire de l'absorption

Les salles simulées par la méthode des sources images sont définies à partir des données d'entrée simples que sont la géométrie des parois délimitant le volume de la salle, ainsi que leurs comportement d'absorption. Or, les coefficients d'absorption des matériaux de construction sont par essence très dépendants du domaine de fréquence des ondes interagissant avec les parois. La connaissance d'une telle information en bande fines est très rarement accessible, et les données exploitées en acoustique des salles ou en acoustique du bâtiment sont en général issues de mesures par bandes fréquentielles d'octaves, ou au mieux, en tiers d'octaves.

Pour cela, afin d'offrir une flexibilité d'utilisation du code de génération de réponses impulsionnelles sur GPU décrit en section 3.2, une dimension supplémentaire pour tous les tenseurs permettant le calcul des réponses impulsionnelles est réservée pour permettre un calcul des contributions par bandes fréquentielles pour les réponses impulsionnelles. Par ailleurs, pour les raisons évoquées en section 3.1.3, la dépendance angulaire de l'onde incidente est remplacée par un α moyen, de la même manière que cela est communément réalisé en acoustique géométrique ou statistique. Cette simplification peut donner des résultats légèrement différents de la réalité, mais elle a le mérite d'être cohérente avec les données d'entrées du problème de modélisation, qui ne permettent que très rarement un accès à cette dépendance angulaire, et sont la plupart du temps issues de mesures exploitant une approche statistique du champ. Par ailleurs, ces approximations sont compensées par un gain substantiel de la charge de calcul [122, 138].

3.1.7.3 Non prise en compte de la diffusion et de la diffraction

La théorie des sources images suppose que les phénomènes de diffusion et de diffraction soient négligées, mêmes si des auteurs ont proposé des adaptations pour intégrer ce type de contributions avec le cadre géométrique [141]. Or, en haute fréquences, ces phénomènes ne sont plus si négligeables si la géométrie de la salle présente des caractéristiques favorables à ces deux phénomènes.

Toutefois, dans le cadre de cette thèse de doctorat, seules des géométries parallélépipédiques ont été simulées, pour lesquelles les phénomènes de diffraction peuvent être rigoureusement ignorés lorsque les parois sont rigides, compte tenu de la présence d'angles droits uniquement [122]. En revanche, dans le cas de géométries plus complexes, en particulier dans le cas des géométrie non convexes, le fait de négliger la diffraction devient une approximation de plus en plus éloignée de la réalité. Malgré tout, la majorité des pièces étant parallélépipédiques, la théorie des sources images reste une excellente approximation de ce point de vue [122].

En ce qui concerne les phénomènes de diffusion, la méthode des sources images repose sur l'hypothèse que les parois délimitant le volume de la salle ne donnent naissance qu'à des réflexions purement spéculaires : toute diffusion lié à la présence d'irrégularités de la paroi est négligée, sauf si une méthode hybride est utilisée, comme proposé dans [138]. En revanche, l'utilisation de ces méthodes hybrides pose toujours le problème du choix du modèle comportemental de diffusion par la paroi, et de la jonction entre le modèle spéculaire et le modèle dominé par la diffusion.

Plutôt que de proposer des améliorations basées sur des hypothèses grossières sur la loi comportementale de diffusion – la plupart du temps modélisée arbitrairement par une loi de Lambert, faute de mieux – et de données non tabulées, l'approche proposée ici consiste à exploiter rigoureusement la méthode des sources images, en corrigeant les valeurs de coefficients d'absorption, de manière à compenser la non prise en compte de leur dépendance angulaire et la non prise en compte d'éventuels phénomènes de diffraction et de diffusion, comme proposé par exemple dans [139]. En effet, comme montré dans [142], la présence de parois diffusantes dans une salle, provoque une décroissance énergétique plus importante que si l'on considère des réflexions exclusivement spéculaires pour un même

coefficient d'absorption (voir figure 3.4).

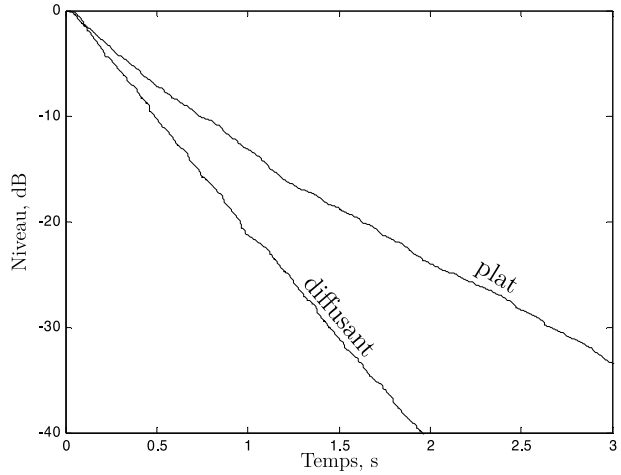


FIGURE 3.4 – Décroissance énergétique dans une salle de concert avec des murs lisses ou diffusants. Image tirée de l'article de M. Barron : *Non-linear decays in simple spaces and their possible exploitation* [142]

Aussi, lorsque la durée de réverbération T_r est connue *a priori* et constitue la grandeur d'entrée pour la simulation numérique de l'acoustique d'une salle, la valeur du coefficient d'absorption utilisé pour sa modélisation peut être différent de celui correspondant aux données des matériaux des parois de la salle. Cette approche est proposée par Mechel dans [131] d'un point de vue analytique pour limiter les effets de l'approximation de sources images par rapport à la solution exacte. La section 3.5 du présent document propose d'ailleurs un raisonnement similaire inspiré de [139], reposant cette fois-ci sur une recherche empirique rapide de cette modification du coefficient α des parois de la salle modélisée.

3.1.7.4 Précision temporelle du calcul

Le but de la simulation du comportement acoustique d'une salle dans le domaine temporel en exploitant le modèle des sources images est d'obtenir des jeux de réponses impulsionnelles de salles les plus complètes possibles, dans un temps de calcul raisonnable, tout en compensant les limites connues des simulations par sources images, notamment concernant le coefficient d'absorption effectif. Pour constituer les bases de données d'entraînement pour la localisation de sources en environnement réverbérant, l'objectif est ici d'exploiter ces jeux de nombreuses réponses impulsionnelles calculées pour un très grand nombre de positions de sources, les récepteurs étant positionnés sur des antennes de

petites dimensions, ce qui signifie que les déphasages relatifs entre les signaux sur les microphones de ces antennes doivent être connus avec une grande précision.

Dans le cadre de cette thèse, puisque l'approche de localisation proposée repose sur un apprentissage à partir de jeux de données conséquents, l'objectif est donc de calculer des réponses impulsionnelles de salles dans un grand nombre de situations. Ces réponses impulsionnelles sont exploitées pour générer des signaux variés afin d'entraîner le réseau de neurones profond proposé, par convolution avec des signaux sonores émis par les sources (voir section 3.3), sur le même principe que les techniques d'auralisation. Puisque les signaux d'entrée du réseau sont des données numériques (échantillonnés et quantifiés), ces calculs doivent être effectués pour des signaux eux aussi échantillonnés, ce qui est l'une des limites connues à l'implémentation naïve de la méthode des sources images, puisque les retards introduits par la distance source-image / récepteur n'a que très peu de chances d'être un multiple de la période d'échantillonnage des signaux. Il a été prouvé [139] qu'une simple approximation à l'échantillon entier des retards introduits par la propagation entre les sources images et un récepteur entraînait la création d'une réponse impulsionnelle à valeur moyenne non nulle, ce qui n'est pas valide d'un point de vue physique. Pour cette raison, puisque la précision et le réalisme des jeux de données est essentielle pour les techniques d'apprentissage, une attention particulière a été également portée sur la précision temporelle et l'implémentation d'une méthode de sources images précise et rapide (voir section 3.3).

3.2 Mise en place de la base de données simulées

Avant de détailler précisément l'implémentation numérique et l'approche de simulation sur GPU de réponses impulsionnelles de salles développée dans le cadre de cette thèse, il est nécessaire de fixer les objectifs liés aux bases données annotées constituées à partir de cet outil. Pour gérer de manière flexible la quantité importante de données obtenues par simulation et acquisition expérimentale, toutes les caractéristiques de sources, de salles, et d'antennes microphoniques sont rassemblées dans un fichier au format JSON [143] qui sera présenté en section 4.2.6.

3.2.1 Géométrie d'antennes pour les bases de données simulées numériquement

L'approche proposée de localisation de sources par Beamlearning se veut la plus indépendante possible de la topologie de l'antenne microphonique, ainsi que du nombre de capteurs la composant. Pour cela, plusieurs antennes ont été utilisées pour générer par simulation numérique des réponses impulsionnelles de salles multicanales. Cette section présente brièvement leurs caractéristiques dans le tableau 3.1.

Nom	Géométrie	Nombre de micros	Répartition des micros	Taille carac.
Circulaire 8 mic.	Circulaire	8	Répartis sur le périmètre de l'antenne	Rayon = 3,95 cm
Mini DSP	Circulaire	7	6 microphones répartis sur le périmètre de l'antenne, 1 au centre	Rayon = 4,3 cm
CMA Cube	Tétraédrique	7	4 microphones positionnées sur les sommets et 3 au centre des arrêtes supérieures	Coté = 10 cm

Tableau 3.1 – Résumé des antennes simulées

Les arguments concernant le choix de ces géométries d'antenne se fera ultérieurement, au chapitre 5. Cette discussion sera appuyée par des résultats de localisation obtenus après entraînement du réseau. Les informations synthétisées dans le tableau 3.1 sont explicitées ici, avec un schéma de la géométrie de chaque antenne. Parmi elles, deux géométries d'antennes ont été simulées explicitement pour comparer les performances de localisation à des jeux de données constitués expérimentalement sur des antennes présentant la même topologie (voir Chapitres 4 et 5). Dans le cadre de ces jeux de données expérimentales, qui seront donc traités dans les prochains chapitres, une antenne sphérique pleine a également été testée (voir section 4.2.2), mais n'a pas été simulée numériquement.

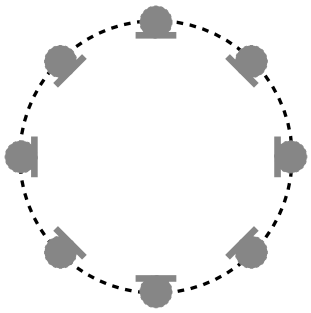


FIGURE 3.5 – Schéma de l'antenne plane à 8 microphones

La première antenne qui a été simulée est une antenne parfaitement circulaire, de 3,95 cm de rayon. Les huit microphones qui la composent sont répartis de manière homogène sur tout le pourtour de l'antenne. Elle est appelée antenne *circulaire 8 microphones* dans tout le document. Sa géométrie a été utilisée essentiellement pour des tâches de localisation angulaire à 2 dimensions. Sans être spécifiquement à 8 capteurs, ce type de répartition circulaire est celle qui équipait les premières générations d'assistants Google Home et Apple HomePod.

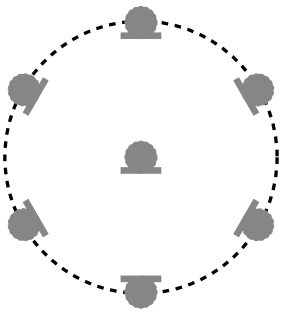


FIGURE 3.6 – Schéma de l'antenne circulaire du constructeur Mini DSP

La deuxième antenne qui a été simulée est également une antenne circulaire. Son rayon est légèrement plus grand et fait 4,3 cm. Six microphones sont sur le périmètre de l'antenne et un est au centre. Sa géométrie modélise l'antenne physique du constructeur Mini DSP qui sera utilisée pour acquérir des base de données expérimentales. Elle est appelée antenne *mini DSP* dans tout le document. Ce type de géométrie correspond aux premières générations d'antennes utilisés dans les assistants Amazon Alexa.

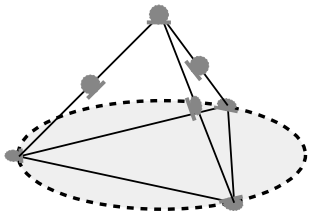


FIGURE 3.7 – Schéma de l'antenne tétraédrique CMA Cube développée avec des microphones double couche lors de la thèse d'Aro Ramamonjy [14]

La troisième antenne qui a été simulée est une antenne, qui contrairement aux deux premières, n'est pas à géométrie plane : les capteurs microphoniques de l'antenne sont ici répartis en trois dimensions. La structure globale est un tétraèdre régulier de 3,2 cm de coté. Quatre microphones sont situés aux sommets du tétraèdre, et les trois derniers sont au milieu des arrêtes supérieures. Sa géométrie modélise l'antenne *CMA Cube* conçue comme premier prototype lors de la thèse d'Aro Ramamonjy au laboratoire [14]. L'objectif initial était de comparer les performances obtenues en utilisant également cette géométrie pour constituer des bases de données mesurées expérimentalement (voir chapitre 4) [3] .

Pour chacune de ces antennes microphoniques, les capteurs microphoniques sont supposés parfaitement omnidirectionnels, avec une courbe de réponse en fréquence idéale. La diffraction par la structure de l'antenne est également négligée. Dans le domaine de l'imagerie acoustique exploitant des mesures sur une antenne microphonique, ces approximations sont généralement fortes, et peuvent entraîner des écarts non négligeables entre la modélisation et les résultats obtenus expérimentalement. En revanche, nous verrons plus loin dans le manuscrit que l'utilisation de la phase d'apprentissage sur des antennes réelles (voir chap.4) permet un étalonnage intrinsèque des microphones, ainsi que la prise en compte de la diffraction éventuelle par la structure de l'antenne, puisque ces caractéristiques sont incluses dans l'optimisation des coefficients du réseau au cours de la phase d'entraînement [144]. Ce point est l'une des spécificités des méthodes d'apprentissage pour la localisation, qui permettent ainsi d'obtenir des résultats équivalents en termes de performances en simulation et expérimentalement, sans pour

3. Cette antenne a le même nombre de capteurs que l'antenne Mini DSP utilisée pour la localisation dans le plan. Un des objectifs de la thèse était de vérifier expérimentalement sur différentes antennes les performances de l'approche BeamLearning. Malheureusement suite à une accumulation de problèmes indépendants de notre volonté (défaillance matériel, travaux de mise aux normes du laboratoire et pandémie internationales), ces comparaisons n'ont pas pu être menées pour l'antenne CMA Cube

autant nécessiter une phase de calibration explicite de tous les capteurs de l'antenne, qui peut être fastidieuse, voire non standardisée dans le cas d'antennes à base de microphones sur puces silicium. Une discussion plus approfondie sur ce sujet aura lieu plus loin (chap. 5.1.4).

3.2.2 Environnements acoustiques utilisés pour l'analyse du comportement du réseau sur des données simulées numériquement

Afin d'étudier la robustesse de localisation aux phénomènes de réverbération dans le cas de la localisation de sources, plusieurs environnements ont été simulés à l'aide du formalisme de l'acoustique géométrique optimisé d'un point de vue temps de calcul grâce à une parallélisation des opérations sur GPU (voir section 3.3). Pour cela, plusieurs salles aux géométries et coefficients d'absorptions différents ont été simulées. En début de thèse, au cours des développements préliminaires du réseau, un environnement de type champ libre a été utilisé, ainsi qu'un environnement de type semi-anéchoïque, où seul le sol est considéré comme réfléchissant. Dans le cadre de ce manuscrit, même si une grande variété de salles ont été modélisées, dans un souci de synthèse, une seule salle a été choisie pour illustrer les résultats de localisation, sans perte de généralité, puisque les tendances observées sur les performances de localisation ont été identiques avec d'autres géométries. Les caractéristiques de cette salle sont données dans la table 3.2.

Type d'environnement	Précisions
Champ libre	Vu comme une pièce aux parois parfaitement absorbantes
Champ semi-ouvert	Vu comme une pièce aux parois parfaitement absorbantes et un sol parfaitement réfléchissant
Pièce	Dimensions : 10x7x3,7 m^3 T_r : 0,5 s α constant : 0,312
Pièce	Dimensions : 10x7x3,7 m^3 T_r : 0,5 s α_{mur} : 0,312 α_{sol} : 0,212 $\alpha_{plafond}$: 0,412

Tableau 3.2 – Résumé des principaux environnements simulés

Même si les environnements libres ou semi-anéchoïques ne sont pas des salles, les réponses impulsionnelles sources-microphones, nécessaires à la constitution des bases de données pour ces environnements, ont également été calculées à l'aide du même code de calcul que pour les salles closes. Ceci permet de tirer profit des raffinements réalisés sur la précision temporelle offerte par les calculs de retards fractionnaires (voir section 3.4).

Comme indiqué dans le tableau 3.2, cette salle utilisée pour présenter les résultats dans le cadre de ce manuscrit est une pièce relativement grande, de dimensions caractéristiques d'une salle de classe type du Cnam (longueur 10 m, largeur 7 m et de hauteur 3,7 m). Les durées de réverbérations des salles de classes simulées sont de 0,5 s, ce qui correspond aussi en pratique aux durées observées dans les salles de classe. Enfin, pour ce qui est des coefficients d'absorption des murs, deux cas de figure sont utilisés : soit les matériaux des parois sont supposés identiques, soit l'absorption des parois est différenciée entre les murs, le plafond et le sol.

Toutes les salles qui ont été modélisées numériquement sont considérées comme vides de tout meuble, aucune diffraction n'est par ailleurs prise en compte, mais les expériences menées au cours de la thèse, notamment celles décrites au chapitre 5, démontrent qu'y compris en présence d'éléments diffractants (corps de l'antenne, table, structure présente dans la salle), l'approche proposée continue à fournir des résultats de localisation pertinents.

3.2.3 Paramétrisation des positions de sources pour l'apprentissage

Puisque la méthode de localisation proposée dans cette thèse repose sur un apprentissage supervisé, il est nécessaire d'entraîner le réseau de neurones profonds à l'aide d'un jeu de données représentatif et varié de positions de sources dans un environnement donné. Au fur et à mesure de l'avancement du projet, la localisation a évolué progressivement d'une tâche de détermination de DOA (*Direction Of Arrival*) purement 2D, dans le plan de l'antenne, à une tâche de localisation angulaire 3D complète. Pour cela, afin de constituer des jeux de données entièrement paramétrables et compatibles avec la géométrie des salles ainsi qu'avec une hypothèse de champ *lointain* et une variabilité de distance de sources par rapport à l'antenne pour une même direction d'arrivée, nous avons choisi de définir les positions des sources par un tirage aléatoire uniforme dans un volume entourant l'antenne, dont la

typologie est illustrée à la figure 3.8.

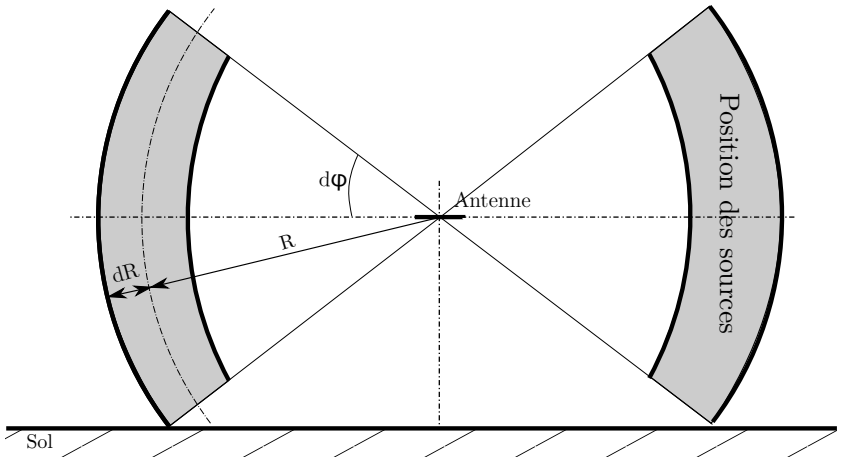


FIGURE 3.8 – Vue en coupe du volume dans lequel sont tirées aléatoirement les positions des sources acoustiques

Ce volume est construit ainsi : tout d'abord, un cercle est défini autour de l'antenne, pour que l'antenne et le cercle soient coplanaires. La vue de la figure 3.8 étant en réalité en coupe, ce cercle est donc dans le plan perpendiculaire à la vue. Le rayon R de ce cercle est défini comme le rayon nominal du volume. Ensuite, une variabilité de rayon dR est introduite. Ce qui définit un disque troué autour de l'antenne. Enfin, une variabilité d'élévation $d\phi$ est introduite pour créer un volume qui permette aux sources de rester dans la salle, en cas de présence de sol ou de plafond. Quand cette variabilité vaut $\pm 90°$, le volume est celui contenu entre deux sphère de rayon $R \pm dR$. La vue en coupe du volume ainsi constitué est schématisée en figure 3.8. L'intérêt premier de ce type de paramétrisation est d'offrir un même cadre de tirage aléatoire pour les positions de sources quelque soit la géométrie de la salle, et de permettre une évolution continue du problème de localisation 2D au problème de localisation 3D.

Lors de la constitution de ce volume, le rayon nominal R doit être suffisamment grand pour que les sources soient en champ lointain. Pour cela on doit vérifier que la distance source-antenne est supérieure à la distance de Fraunhofer [145]. Comme les sources sont supposées ponctuelles, et compte tenu du domaine de fréquence visé pour nos applications, on fixe cette distance à 1m50 environ. Ensuite, il faut évidemment vérifier que le volume soit entièrement contenu dans la salle. Par ailleurs, pour ne pas avoir des approximations trop importantes en basses fréquences avec la simulation de réponses

impulsionnelles par la méthode des sources images, les sources doivent également ne pas être situées trop proches des parois, comme déjà discuté en section 3.1.7.

Pour finir, afin d'améliorer la représentativité de l'ensemble des paramètres angulaires au cours du tirage de sources dans ce volume, l'échantillonnage aléatoire peut également être réalisé selon une loi uniforme sur des sous volumes. Ces sous-volumes peuvent être déterminés par la méthodes des hypercubes latins [146]. Dans le cas de la sphère, les sous-volumes peuvent être échantillonnés selon l'élévation Φ. Dans ce cas précis, les élévations de chaque sous-volume sont déterminés grâce à la formule de récurrence suivante (n étant le nombre de sous-volumes équivalents à obtenir) :

$$\begin{cases} \phi_0 & = -\frac{\pi}{2} \\ \phi_{i+1} & = \arcsin(\frac{2}{n} + \sin(\phi_i)) \end{cases} \tag{3.18}$$

3.2.4 Types de signaux émis par les sources

À ce stade, les hyperparamètres de la base de données – la géométrie de l'antenne, l'environnement acoustique des antennes et des sources, ainsi que les positions de sources – permettent de constituer des jeux de données de réponses impulsionnelles multicanales. Une fois ces RIR calculées, elles peuvent être convoluées avec n'importe quel signal pour simuler sa propagation dans la pièce choisie, depuis la position choisie jusqu'à l'antenne choisie. L'intérêt de cette approche est l'une des raisons qui a motivé le choix de la méthode des sources images, puisqu'elle permet ainsi une flexibilité : le calcul des jeux de données de réponses impulsionnelles multicanales étant mutualisé, cela représente un gain de temps de calcul indéniable. En effet, dans le processus de simulation, c'est le calcul des jeux de données de réponses impulsionnelles qui est le plus long, par rapport au calcul de *spatialisation* d'un signal particulier émis par les sources composant le jeu de données à construire (voir figure 3.9).

Conceptuellement, cette approche revient à considérer que les jeux de données sont essentiellement dimensionnés par le nombre de réponses impulsionnelles multicanales calculées, puisqu'elles correspondent chacune à une position de source dans un environnement donné ; le nombre de type de signaux émis par ces sources peut être alors vu comme un processus d'augmentation de données [147]

pour la tâche de localisation visée. En effet, ce qui caractérise la position de la source n'est pas le signal en lui même, mais uniquement les décalages de temps et les variations d'amplitude du signal inter-microphoniques au sein de l'antenne.

Sur le même principe que les méthodes d'auralisation, les réponses impulsionnelles étant calculées, la convolution avec n'importe quel type de signal peut être réalisée, à condition toutefois que les fréquences d'échantillonnage des signaux et réponses impulsionnelles soient concordantes. Dans le cadre de cette thèse, plusieurs types de signaux ont été choisis pour créer les jeux de données d'entraînement et de test du réseau. Leurs caractéristiques sont récapitulées dans le tableau 3.3.

Avant de réaliser la convolution de ces signaux avec les réponses impulsionnelles, tous les signaux sont filtrés, pour couper les très basses fréquences (en dessous de 100 Hz) et les hautes fréquences (au dessus de 4000 Hz). Ce filtrage en amont n'est pas strictement nécessaire, mais est réalisé pour être en cohérence avec le domaine de validité de la reconstruction ambisonique obtenu grâce à la sphère de spatialisation pour les bases de données expérimentales (voir chap. 4), dans la zone occupée par les antennes compactes étudiées.

Au cours de la constitution de la base de données, pour chaque position de source, une séquence de 1 024 échantillons est sélectionnée dans chaque typologie de signal, en s'assurant que la séquence ne corresponde pas à une période de *silence*. Le départ de cette séquence est tiré aléatoirement afin qu'une trame de signal différente soit utilisée pour chaque position, ce qui permet d'offrir une variabilité importante de signaux présentés en entrée du réseau, tout en contrôlant leur typologie. Ce processus diversifie la base de données et limite donc le phénomène de sur-apprentissage du réseau.

L'auralisation de ces signaux est alors effectuée en convoluant chaque portion de 1 024 échantillons temporels des signaux mono avec les jeux de réponses impulsionnelles multicanales précalculées (qui peuvent, elles, être beaucoup plus longues, puisque dépendantes du temps de réverbération de la salle).

L'ensemble de ces opérations étant réalisées sur les types de signaux d'entrées recensés dans le

tableau 3.3, les signaux sont ensuite convolués avec les jeux de données de réponses impulsionnelles multicanales de salles. Le nombre de ces convolutions pouvant aisément excéder des centaines de milliers de convolutions *longues* (voir figure 3.9), elles sont aussi calculées en exploitant la puissance de calcul sur GPU, et les optimisations offertes par la librairie *Tensorflow* et CUDA pour les opérations de convolutions par batchs. Il est par ailleurs tout à fait envisageable, au prix d'une légère augmentation des temps d'entraînement, d'économiser de l'espace de stockage de mémoire sur disque des bases de données auralisées. Pour cela, l'opération de convolution peut aisément être réalisée "à la volée", au cours de l'apprentissage. L'espace mémoire SSD offert par notre station de calcul Deep Learning étant suffisant pour stocker les bases de données auralisées utilisées dans le cadre de cette thèse, nous avons fait le choix de privilégier le temps de calcul d'apprentissage, et n'avons donc pas retenu cette solution. En revanche, pour des bases de données multi-salles, cette approche est déjà implémentée, et présenterait une solution élégante à l'explosion de la taille des jeux de données, qui pourrait dans ce cas facilement excéder la centaine de Tera-octets de données auralisées, contre seulement une dizaine de Tera-octets de données de réponses impulsionnelles stockées sous forme de tenseurs parcimonieux.

Le choix des différents types de signaux présentés dans le tableau 3.3 a été réalisé principalement afin d'étudier d'étudier la qualité d'apprentissage de localisation basée sur le BeamLearning face à différentes situations. Bien entendu, pour un entraînement plus complet, ces types de signaux pourraient être étoffés par d'autres types de signaux audio, mais nous verrons dans la suite du document que l'approche proposée offre une capacité de généralisation essentiellement basée sur les bandes fréquentielles contenues dans les signaux. Pour cela, les gabarits fréquentiels des signaux exploités dans le cadre du manuscrit sont très variables, depuis les simples signaux monochromatiques (testés pour les fréquences normalisées de bandes d'octaves), jusqu'aux pièces orchestrales symphoniques, en passant par des signaux vocaux féminins de type *cocktail party*. Ici, la raison pour laquelle nous avons exploité des signaux de type *cocktail party* repose essentiellement sur le fait que les contenus spectraux sont vocaux, et que ces signaux ont, par essence, une probabilité très faible de contenir des trames *silencieuses*, contrairement à un signal vocal d'un seul locuteur. La dynamique en amplitude des signaux étant elle aussi importante, toute l'échelle d'intensité des fichiers .wav est utilisée. En effet, une normalisation a été effectuée pour que le maximum en valeur absolue des amplitudes sur l'ensemble des signaux soit égal à 0,99, tout en gardant un rapport d'amplitude constant. Pour finir, au cours

Nom	Durée (s)	Description	Référence
Sinus	∞	Sinus pur à une fréquence donnée	
BackgroundSpeech	0 :25	Enregistrement anéchoïque de discussion de femmes danoises	[148]
Beethoven	3 :11	Enregistrement anéchoïque d'un orchestre symphonique	[149]
Brahms	1 :27	Enregistrement anéchoïque d'un orchestre symphonique	[149]
Mahler	2 :12	Enregistrement anéchoïque d'un orchestre symphonique	[149]
Mozart	3 :47	Enregistrement anéchoïque d'un orchestre symphonique et d'une chanteuse	[149]
Klaxon	0 :01	Un des exemples de klaxon de la base de données UrbanSound 8K. Seule la partie avec du signal est conservée	[34]

Tableau 3.3 – Résumé des caractéristiques des fichiers audios utilisés pour constituer la base de données

de l'apprentissage, ces signaux sont ensuite *augmentés* par l'ajout de bruit statistique décorrélé sur chacun des capteurs des antennes, avec un rapport signal à bruit variable.

3.3 Implémentation du calcul massif de réponses impulsionnelles multicanales de salles et d'auralisation, sur architecture GPU

L'objectif ici est de générer un très grand nombre de réponses impulsionnelles et de réaliser ensuite une tâche d'auralisation multicanale pour chacune de ces réponses impulsionnelles. Par conséquent, une part des travaux de développements au cours de cette thèse de doctorat a consisté à proposer une approche de calcul massif, en exploitant les architectures et les *frameworks* de calcul sur GPU. En effet, la plupart des outils développés par la communauté scientifique sont adaptés au calcul de réponses impulsionnelles, mais en faible nombre. Les temps de calcul de ces outils sont systématiquement incompatibles avec la taille des jeux de données dont nous avons besoin pour générer des jeux

de données d'apprentissage. Dans cette section, l'approche algorithmique que nous avons retenu est explicitée en détail.

Pour le calcul des jeux de réponses impulsionnelles, celui-ci est réalisé en plusieurs étapes. Dans un premier temps, les caractéristiques de la salle sont définies, en particulier les coefficients d'absorption des différentes parois délimitant le volume de la salle (voir section 3.5). Dans un second temps, les positions géométriques et facteurs d'atténuation correspondants pour l'ensemble des sources images correspondant à chaque position du volume choisi (voir section 3.2.3) sont calculées, grâce à la librairie Python Pyroomacoustics [150]. Cette librairie n'exploite pas un calcul sur GPU mais reste optimisée pour cette tâche purement géométrique. C'est le calcul des réponses impulsionnelles à partir de ces paramètres qui a, quant à lui, été développé spécifiquement dans le cadre de cette thèse pour être exécuté sur processeur graphique. À partir des positions et atténuations individuelles de sources images (en très grand nombre pour chacune des positions de sources du volume), une réponse impulsionnelle partielle et individuelle correspondant au retard non entier est construite efficacement sur GPU, grâce à une approche reposant sur l'utilisation des polynômes de Lagrange (voir Section 3.4) et l'utilisation de tenseurs parcimonieux. L'ensemble des contributions de ces sources images sont ensuite sommées et pondérées afin d'obtenir les réponses impulsionnelles multicanales pour chacune des positions du volume, toujours grâce à un calcul sur GPU en exploitant la librairie Tensorflow [116] de manière non conventionnelle, c'est à dire comme outil de calcul scientifique déterministe par lots. En effet, pour calculer des dizaines de milliers de réponses impulsionnelles en un temps de calcul raisonnable, l'utilisation d'une approche sur CPU est inenvisageable, même avec les optimisations offertes par l'approche de retards fractionnaires par interpolation de Lagrange. Pour finir, la librairie Tensorflow et les optimisations CUDA étant particulièrement adaptées au calcul de convolutions par lots, la tâche d'auralisation est, elle aussi, réalisée par calcul sur GPU.

3.3.1 Résumé des étapes de calcul

Comme expliqué dans les paragraphes précédents, l'approche algorithmique d'auralisation massive pour un grand nombre de sources dans une salle repose sur 3 étapes principales, correspondant à 3 codes de calculs modulaires. Les étapes algorithmiques et les codes sont schématisés sur la figure 3.9.

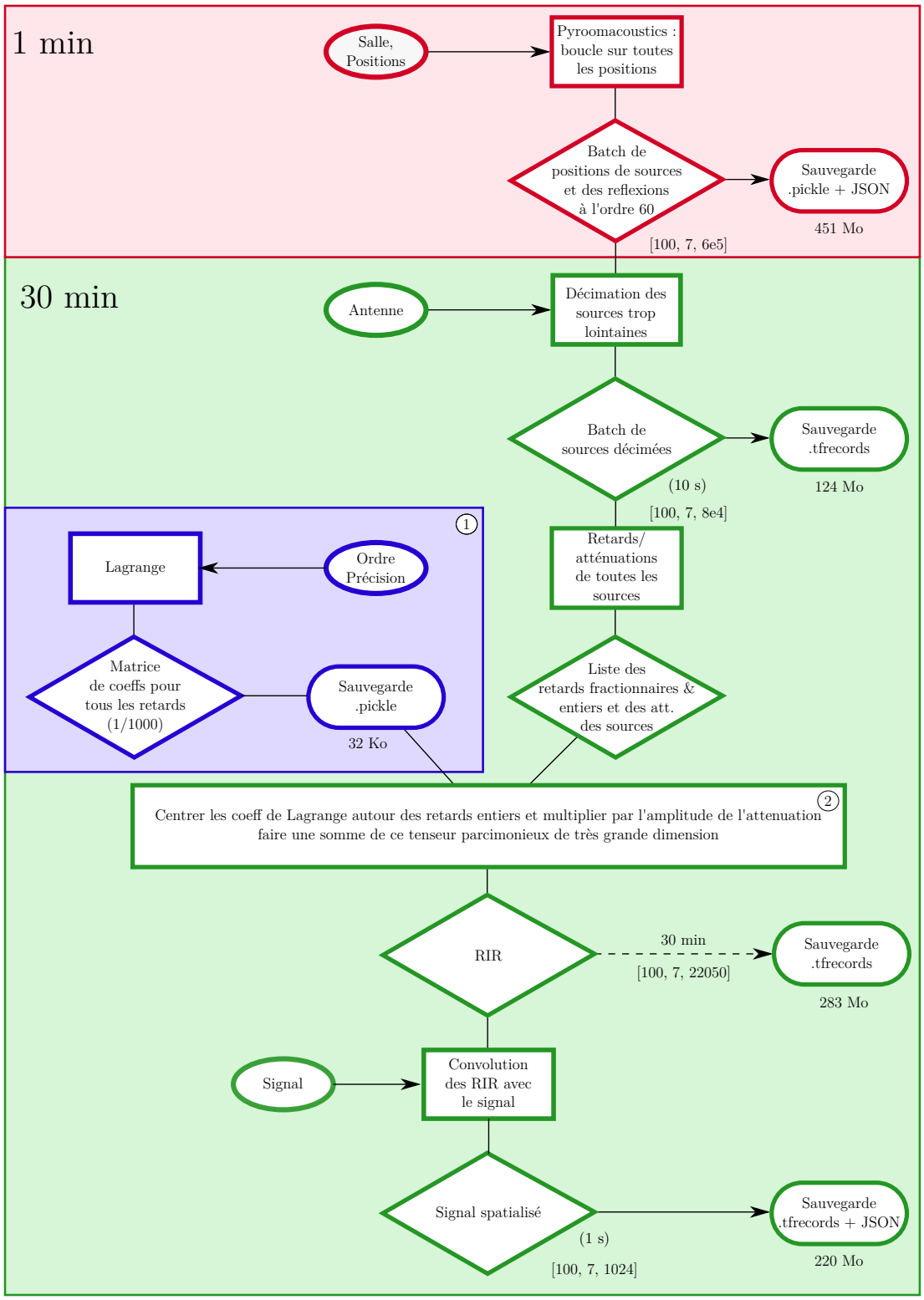


FIGURE 3.9 – Schéma bloc des différentes étapes de calculs du programme réalisé pour l'auralisation massive dans une salle (8000 positions de sources, 7 capteurs, environ 80000 sources images par position de sources).

La structure de ce schéma-bloc est la suivante : les éléments ovales représentent les entrées définissables par l'utilisateur ; les éléments rectangles symbolisent des actions réalisées, qui sont suivies de losanges, définissant les sorties remarquables de ces dernières. Enfin, lorsque ces sorties sont sauvegardées, elles pointent vers des rectangles aux bords arrondis précisant le format de sauvegarde. Pour fixer les idées, quelques ordres de grandeurs sont précisés sous chacun des blocs. Ces ordres de grandeurs sont issus d'un calcul d'auralisation massive pour une salle, avec une antenne de 7 microphones, et un ensemble de 8000 positions de sources dans le volume V décrit dans à la section 3.2.3.

Dans ce manuscrit, 2 portions de l'algorithme sont primordiales et représentent une approche originale. Elles sont signalées sur la schéma-bloc de la figure 3.9 par un chiffre entouré, et seront détaillées dans la suite du manuscrit. La figure 3.10 illustre ces deux tâches, qui permettent de calculer la portion de réponse impulsionnelle associée à un couple source image / microphone. Ces calculs étant réalisés sur GPU, toutes les grandeurs sont représentées par des tenseurs, offrant la possibilité de réaliser les calculs par lots (de sources, de microphones, et de sources images), afin d'accélérer ce traitement.

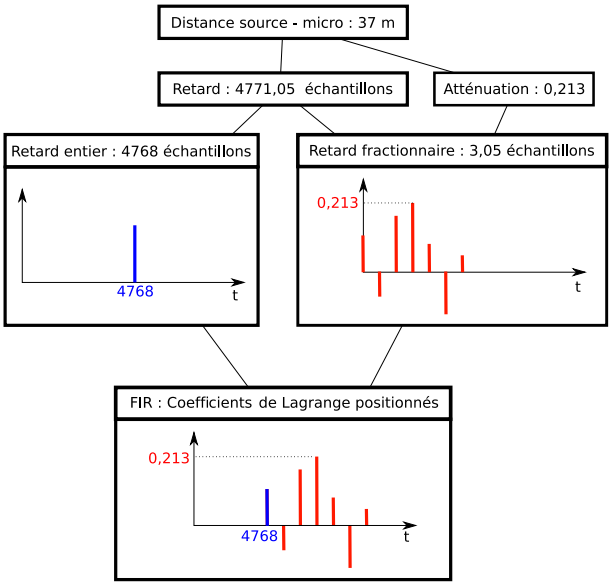


FIGURE 3.10 – Exemple schématique du calcul en deux temps (partie fractionnaire, partie entière) pour le calcul en lot de réponses impulsionnelles. Ici, l'illustration concerne le calcul de la portion de réponse impulsionnelle associée à une source image exclusivement, pour un microphone donné, et est stockée dans un tenseur parcimonieux. La réponse impulsionnelle multicanale pour chaque source du volume est ensuite calculée par sommation des tenseurs sur la dimension des sources images

① **Calcul des coefficients de Lagrange *a priori* :** La précision de la méthode des sources images pour des signaux échantillonnés reposant en grande partie sur la manière dont est réalisé le décalage temporel de signaux de moins d'un échantillon, nous avons décidé d'exploiter une approche reposant sur l'utilisation de polynômes de Lagrange, qui est beaucoup plus efficace algorithmiquement que la méthode - pourtant couramment utilisée - du sinus cardinal tronqué, tout en offrant une précision supplémentaire [93, 151–153]. La littérature scientifique a proposé d'exploiter un calcul *à la volée* de ces polynômes, ou de structures adaptatives [154–156]. Cependant, dans le cadre de nos applications, ceci nécessiterait d'utiliser un calcul à la volée pour toutes les sources images nécessaires à la constitution d'une base de données, ce qui impacterait trop lourdement le temps de calcul, pour un gain très discutable (pour l'exemple utilisé sur la figure 3.10, cela nécessiterait de créer 600 millions de réponses impulsionnelles individuelles). En effet, les optimisations proposées par les auteurs [154–156] sont essentiellement conçues pour créer des filtres à réponses variables, qui pourraient être efficaces pour modéliser des sources en mouvement, ce qui n'est pas notre cas ici. Par conséquent, nous nous sommes orientés vers une solution basée sur l'utilisation d'une table de correspondance [154]. Avec cette approche, 1 000 versions de réponses impulsionnelles de filtres à retard fractionnaire par interpolation de Lagrange sont pré-calculées, correspondant à 1 000 décalages possibles, multiples d'un millième d'échantillons. Ce pré-calcul étant réalisé une fois pour toutes, ces 1 000 réponses impulsionnelles de filtres sont stockés dans un tenseur statique, dans lesquels seront prélevées les réponses impulsionnelles pour créer le tenseur parcimonieux intermédiaire à la constitution de la réponse impulsionnelle de chaque source du volume (voir section 3.4)

② **Centrer les coefficients de Lagrange autour des retards entiers :** Contrairement au calcul de la partie fractionnaire détaillée au dessus, qui est basée sur une recherche dans une table de correspondance et un pré-calcul, les parties entières des retards correspondant à chaque source image sont quant à elles calculées à la volée, pour chaque salle et chaque position de source. En effet, l'approche choisie pour implémenter les retards fractionnaires avec les polynômes de Lagrange est la suivante : on définit une réponse impulsionnelle pour chaque couple (Source image, Micro), ces réponses impulsionnelles sont stockées dans un tenseur parcimonieux permettant d'obtenir une représentation la plus compacte possible. Pour créer la réponse impulsionnelle correspondant à chaque couple (Source image, Micro), on sépare tout d'abord les parties entières et fractionnaires. Les coefficients de Lagrange correspondant

à la partie fractionnaire du retard sont ensuite décallés du nombre entier d'échantillons calculé à la volée (voir figure 3.10).

3.4 Filtres à retards fractionnaires pour les signaux échantillonnés

Le fait de retarder un signal échantillonné est un problème qui se pose souvent lors de simulations numériques dans le domaine temporel, que ce soit pour la synthèse sonore d'instruments de musique, la simulation en acoustique des salles, ou l'imagerie acoustique. Lorsque la précision temporelle nécessaire au calcul est de l'ordre de grandeur de la période d'échantillonnage, le fait d'arrondir le retard à l'échantillon près est la méthode la plus rapide et la plus simple, mais peut mener à une approximation beaucoup trop grossière pour les applications visées. En revanche, dans le cas de la localisation acoustique sur antennes microphoniques compactes, il est primordial que la simulation numérique permette une précision temporelle élevée.

Compte tenu des fréquences d'échantillonnage communément utilisées par les systèmes d'acquisition, il semble logique de conserver ce type d'échantillonnage temporel pour les simulations numériques, même si certains auteurs ont proposé un sur-échantillonnage pour compenser les faiblesses de l'utilisation de retards à l'échantillon entier. Pour des bases de données déjà conséquentes, ce type de contournement n'a pas de sens, puisqu'il impacterait énormément l'empreinte mémoire des signaux stockés pour les jeux de données. Pour fixer un ordre de grandeur, une base de données de réponses impulsionnelles multicanale pour une seule salle, avec une antenne à 7 microphones et 10000 positions de sources, qui occuperait environ 60 Go de mémoire en temps normal, pourrait occuper 60 To de mémoire si le suréchantillonnage était identique aux pas de retards fractionnaires utilisés (un pas tous les millièmes d'échantillons). Cette solution n'est donc absolument pas viable, surtout dans l'objectif de stocker des réponses impulsionnelles d'un grand nombre de salles. Ainsi, nous avons opté pour une approche permettant de retarder le signal d'un nombre non entier d'échantillons, afin d'être en mesure de représenter le plus fidèlement la réponse impulsionnelle de la salle et ses variations sur la faible extension spatiale des antennes microphoniques compactes utilisées.

Cette approche, connue sous le nom de filtrage à retard fractionnaire, consiste à convoluer le signal

échantillonné $s[n]$ à retarder, avec une réponse impulsionnelle $h[n]$ d'un filtre numérique, permettant de modéliser ce retard d'un nombre d'échantillons non entier :

$$y[n] = \sum_{k=0}^{n} h[k] \cdot s[n-k] \tag{3.19}$$

Sous réserve d'être en mesure de faire cette opération de manière très rapide, cette approche permet de conserver une taille de fichiers acceptable pour le travail proposé. Le choix de ce cette fonction de retard est donc primordial, et nous proposons de présenter les choix et implémentations réalisées dans le cadre de cette thèse.

3.4.1 Cas général d'un signal numérique échantillonné

Pour un signal temporel $s(t)$, soit la suite $s[n]$ le signal correspondant à $s(t)$ échantillonné avec une période T_e. Par définition du théorème d'échantillonnage, on a :

$$s(t) = \sum_{n=-\infty}^{+\infty} s[n] \cdot \mathrm{sinc}\left(\frac{\pi}{T_e}(t - nT_e)\right) \tag{3.20}$$

Soit le signal $s_R(t)$, un signal temporel correspondant au signal $s(t)$ retardé de τ secondes. On a, par construction :

$$s_R(t) = s(t - \tau)$$

On cherche à présent la réponse impulsionnelle du filtre temporel continu $h(t)$ qui permettrait de retarder notre signal. Par construction du filtre on a alors :

$$s_R(t) = s(t) \circledast h(t) = \int_{-\infty}^{+\infty} s(u) \cdot h(t-u)\ du \tag{3.21}$$

et la distribution de Dirac retardée $\delta(t - \tau)$ est la solution naturelle à notre problème.

Afin de trouver son équivalent pour des signaux échantillonnés, nous cherchons maintenant le filtre numérique permettant de retarder un signal numérique d'un même temps τ. En réinjectant la

distribution de Dirac retardée dans l'équation 3.21, et en utilisant le théorème d'échantillonnage 3.20, on obtient :

$$\begin{aligned} s_R(t) &= \int_{-\infty}^{+\infty} s(u) \cdot \delta(t-\tau-u)\ du \\ &= \sum_{n=-\infty}^{+\infty} \int_{-\infty}^{+\infty} s[n] \cdot \mathrm{sinc}\left(\frac{\pi}{T_e}(u-nT_e)\right) \cdot \delta(t-\tau-u)\ du \\ s_R(t) &= \sum_{n=-\infty}^{+\infty} s[n] \cdot \mathrm{sinc}\left(\frac{\pi}{T_e}(t-\tau-nT_e)\right) \end{aligned} \tag{3.22}$$

On peut donc *théoriquement* interpoler n'importe quel signal numérique par un sinus cardinal échantillonné et retardé, pour obtenir un signal temporel décalé dans le temps par rapport au signal numérique original. Cette formule permet donc de comprendre l'origine de l'utilisation de la fonction sinus cardinal dans la plupart des implémentations numériques de retards non entiers pour des signaux échantillonnés. En effet, comme la formule 3.22 est valable pour toutes les valeurs de t, elle est en particulier valable pour $t = kT_e$, ce qui mène à l'expression suivante pour le signal retardé échantillonné :

$$s_R[k] = \sum_{n=-\infty}^{+\infty} s[n] \cdot \mathrm{sinc}\left(\pi(k-\frac{\tau}{T_e}-n)\right) \tag{3.23}$$

Nous sommes ici en présence d'un filtre non causal infini, qui pose le problème de la réalisation du filtrage numérique associé, et justifie l'utilisation de la méthode du *sinus cardinal tronqué*, obtenu par fenêtrage de la réponse impulsionnelle $\mathrm{sinc}\left(\pi(k-\frac{\tau}{T_e}-n)\right)$. Dans le cas où le retard à effectuer est un nombre entier d'échantillons $\tau = mT_e$, le filtre devient $\mathrm{sinc}\left(\pi(k-(m-n))\right)$. Comme la fréquence d'échantillonnage du signal est $\frac{1}{T_e}$, ce sinus cardinal est équivalent à l'impulsion unité, notée par analogie avec la distribution de Dirac $\delta[n]$. Dans ce cas simple d'un retard multiple de la période d'échantillonnage on retrouve donc simplement l'équivalent numérique de la solution analogique. En revanche, dès que le retard n'est plus entier, la réponse impulsionnelle idéale du filtre numérique (non implémentable en pratique) est une réponse impulsionnelle infinie et non causale, puisque les valeurs de $h[k]$ ne sont plus confondus avec les zéros de la fonction sinus cardinal (voir figure 3.11). :

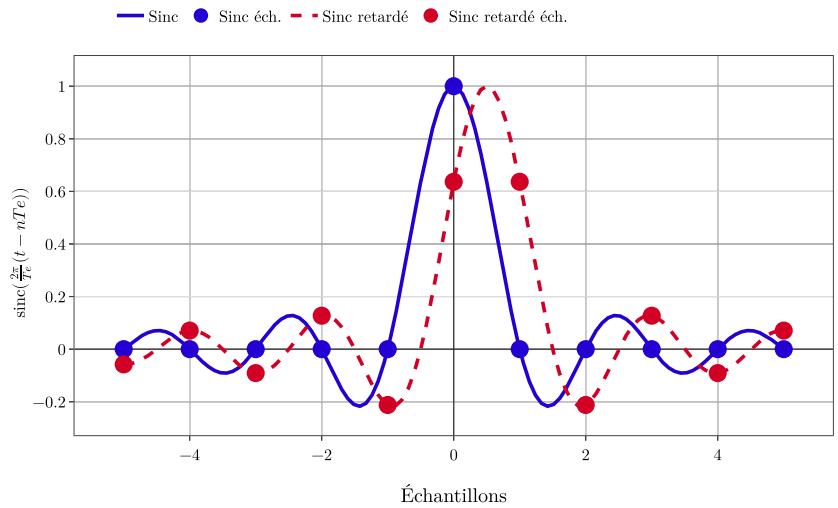


FIGURE 3.11 – Comparaison de deux sinus cardinaux et de leurs échantillons temporels : cas particulier où la période du sinus cardinal est égale à la période d'échantillonnage. Le retard présenté est d'un demi échantillon.

Afin de rendre implémentable numériquement cette convolution par une fonction sinus cardinal, il est donc nécessaire de fenêtrer la réponse impulsionnelle du filtre par une fonction de préférence symétrique [93] (les fenêtres de Hamming, Blackman ou encore Kaiser sont communément utilisées). Cependant, ces approches nécessitent tout de même un grand nombre de coefficients pour que l'approximation ne génère pas des erreurs de troncatures trop importantes. La communauté scientifique s'est penchée sur d'autres approches moins courantes pour retarder un signal [93, 157–159]. L'objectif ici n'est pas de réaliser une revue bibliographique complète de ces méthodes, mais le tableau 3.4 résume différentes méthodes mettant en avant leurs caractéristiques essentielles et le nombre de coefficients à utiliser.

Pour optimiser la vitesse de calcul, le choix s'est porté sur l'approche de filtre de retard fractionnaires par interpolation de Lagrange, reconnue pour sa grande précision et l'efficacité algorithmique offerte, pour un faible nombre de coefficients [154] [4] .

4. Dans le cas d'un échantillonnage uniformément espacé, l'interpolation tend vers un sinus cardinal quand l'ordre du polynôme tend vers l'infini [93]

Nb. points	1	N	N grand	∞
FIR	Linéaire	Lagrange	Sinc fenêtré	
IIR	Passe tout	Thiran		Sinc complet

Tableau 3.4 – Tableau résumant les différents filtres utilisables pour retarder un signal échantillonné d'après [93]

3.4.2 Filtrage à retard fractionnaire par interpolation de Lagrange

L'approche de filtrage par interpolation de Lagrange peut être justifiée et explicitée soit dans le domaine temporel, soit dans le domaine fréquentiel. Ces deux justifications sont présentées dans la suite du document, car elles sont complémentaires et apportent un éclairage sur des avantages différents de la même méthode.

3.4.2.1 Justification de l'approche dans le domaine temporel

Cette approche est mathématiquement la plus simple à formuler. Elle consiste à chercher une fonction polynomiale interpolant exactement un nombre fini de points. Soit $s_R(t) = s(t-\tau)$ la fonction que l'on cherche à approximer. Posons $s_K[t_k]$ la suite des N+1 échantillons de $s_R(t)$ disponibles. Le polynôme $P(t)$ d'ordre N qui passe exactement par les points $s_K[n]$ est :

$$P(t) = \sum_{k=0}^{N} l_k(t) \cdot s_K[t_k]$$
$$\text{avec } l_k(t) = \prod_{\substack{i=0 \\ i \neq t_j}}^{N} \frac{k-i}{t-i}, \text{ et } l_k(t_j) = \begin{cases} 1 & \text{si } j = k \\ 0 & \text{si } j \neq k \end{cases}$$

Cette interpolation polynomiale peut être vue comme le filtrage du signal $s_K[n]$ par le filtre dont les coefficients sont les polynômes $l_k(t)$ appliqués au retard voulu. En posant $h_\tau[n]$ le filtre d'ordre N permettant d'obtenir un retard fractionnaire de τ échantillons, ont peut alors l'exprimer selon la formule 3.24, qui correspond à l'expression analytique des valeurs de la réponse impulsionnelle du filtre

par interpolation de Lagrange d'ordre N, permettant de réaliser un retard d'un nombre non entier d'échantillons :

$$h_\tau[n] = \prod_{\substack{i=0 \\ i \neq n}}^{N} \frac{\tau - i}{n - i}, \qquad n = 0, ..., N \tag{3.24}$$

3.4.2.2 Approche par minimisation d'erreur fréquentielle

La deuxième approche consiste à montrer que les filtres de retards fractionnaires par interpolation de Lagrange reviennent à la minimisation d'une erreur entre la réponse fréquentielle d'un filtre FIR et le filtre idéal continu correspondant au retard. On rappelle que la convention utilisée pour passer d'une fonction $f(t)$ à une fonction $F(\omega)$ par transformée de Fourier (TF) est :

$$F(\omega) = TF\{f\} = \int_{-\infty}^{+\infty} f(t) \cdot e^{-j\omega t} dt \tag{3.25}$$

Dans le domaine de Fourier, retarder un signal temporel revient à déphaser toutes les fréquences par le même coefficient τ, le retard à introduire.

$$s_R(t) = s(t) \circledast \delta(t - \tau) \tag{3.26}$$

$$\Leftrightarrow S_R(\omega) = S(w) \cdot e^{-j\omega\tau} \tag{3.27}$$

Comme nous travaillons dans le cas de filtres numériques, la transformée de Fourier à temps discret de notre filtre $h[n]$ échantillonné avec un pas T_e est :

$$H(e^{j\omega T_e}) = \sum_{n=-\infty}^{\infty} h[n] e^{-jn\omega T_e} \tag{3.28}$$

En faisant l'hypothèse d'un filtre causal ($h[n < 0] = 0$) et d'ordre N, défini par sa FIR ($h[n > N] = 0$), il vient que le filtre est défini par :

$$H(e^{j\omega T_e}) = \sum_{n=0}^{N} h[n] e^{-jn\omega T_e} \tag{3.29}$$

On peut alors définir l'erreur $E(\omega)$ générée, comme la différence entre ce filtre et le filtre théorique :

$$\begin{aligned} E(\omega) &= e^{-j\omega\tau} - H(e^{j\omega T_e}) \\ &= e^{-j\omega\tau} - \sum_{n=0}^{N} h[n] e^{-jn\omega T_e} \end{aligned} \tag{3.30}$$

Afin de minimiser l'erreur fréquentielle introduite par le filtre, on se propose alors d'utiliser la méthode des fonctions n-plates (*maximally flat design error* en anglais [93, 160]). L'idée est de fixer à 0 la valeur des N premières dérivées de la fonction d'erreur au point $\omega = 0$. Ainsi, on force la fonction d'erreur à être la plus plate possible autour du point choisi. Évidemment, les dérivées peuvent s'annuler à un autre point, mais le choix de la pulsation nulle se fait sur deux critères. Premièrement, la minimisation des erreurs de phase en basses fréquences est un choix standard et physiquement valide en acoustique. La seconde raison de ce choix est plus pragmatique : la résolution des équations aux dérivées partielles à un autre point entraînerait une solution de filtre dont les coefficients sont complexes, donc plus difficiles à mettre en place [160]. Nous nous bornerons donc à trouver les coefficients du filtre minimisant les erreurs de phase autour du point $\omega = 0$. Mathématiquement, le problème se pose donc sous la forme suivante :

$$\left.\frac{d^k E(e^{j\omega})}{d\omega^k}\right|_{\omega=0} = 0, \qquad k = 0, ..., N \tag{3.31}$$

$$\Leftrightarrow \sum_{n=0}^{N} n^k h[n] = \tau^k, \qquad k = 0, ..., N \tag{3.32}$$

En posant $\mathbf{V}$ la matrice de Vandermonde telle que $v_{i,j} = j^i$, $\mathbf{h}$ le vecteur rassemblant les coefficients du filtre $h[n]$, et $\boldsymbol{v_\tau}$ le vecteur avec les puissance de τ de 0 à N : $0, \tau, ..., \tau^N$, le problème peut s'écrire sous la forme matricielle :

$$\begin{bmatrix} 0^0 & ... & N^0 \\ \vdots & \ddots & \vdots \\ 0^N & ... & N^N \end{bmatrix} \cdot \begin{bmatrix} h[0] \\ \vdots \\ h[N] \end{bmatrix} = \begin{bmatrix} \tau^0 \\ \vdots \\ \tau^N \end{bmatrix} \tag{3.33}$$

$$\Leftrightarrow \mathbf{V} \cdot \mathbf{h}^T = \boldsymbol{v_\tau}^T \tag{3.34}$$

La matrice de Vandermonde étant connue pour être inversible, on peut donc utiliser la méthode de Cramer pour trouver cette matrice inverse $\mathbf{V}^{-1}$: les coefficients du filtre peuvent donc être calculés simplement grâce à l'équation suivante :

$$\mathbf{h}^T = \mathbf{V}^{-1} \cdot \boldsymbol{v_\tau}^T \tag{3.35}$$

La résolution de l'équation 3.35 permet alors de montrer que le filtre obtenu par cette approche de minimisation se trouve être composé des coefficients des polynômes de Lagrange appliqués au retard fractionnaire τ, ce qui correspond strictement à la même expression que celle obtenue grâce à l'approche dans le domaine temporel :

$$h_\tau[n] = \prod_{\substack{i=0 \\ i \neq n}}^{N} \frac{\tau - i}{n - i}, \qquad n = 0, ..., N \tag{3.36}$$

3.4.3 Analyse de l'interpolation de Lagrange

Pour pouvoir implémenter cette méthode de filtres à retard fractionnaire par interpolation de Lagrange, il faut déterminer jusqu'à quel ordre N les polynômes devront être calculés, c'est à dire combien de points seront utilisés pour l'interpolation. On se propose donc de caractériser l'erreur maximale admissible du signal retardé, puis d'examiner quantitativement la méthode utilisée. Pour toute l'analyse qui suit, on introduit ω la pulsation nomalisée telle que :

$$\omega \in [-\pi; \pi[$$

3.4.3.1 Propriété des retards

Quelque soit l'ordre du polynôme de Lagrange choisi, l'erreur introduite par l'interpolation est dépendante du retard à introduire. Tout d'abord, lorsque le retard τ à introduire est entier, les coefficients du filtre se trouvent être exactement égaux à ceux de l'impulsion unité (cf. eq. 3.24). Ainsi pour un retard entier, l'erreur est rigoureusement nulle, tout comme pour la méthode du sinus cardinal tronqué. Par ailleurs, l'erreur introduite par le filtre peut être analysée du point de vue fréquentiel, à l'aide d'une analyse de type diagramme de Bode [161], ou en analysant l'erreur quadratique intégrale

entre les coefficients du polynôme de Lagrange calculé, et ceux du retard fractionnaire théorique et idéal, le sinus cardinal non tronqué et non fenêtré (eq. 3.23).

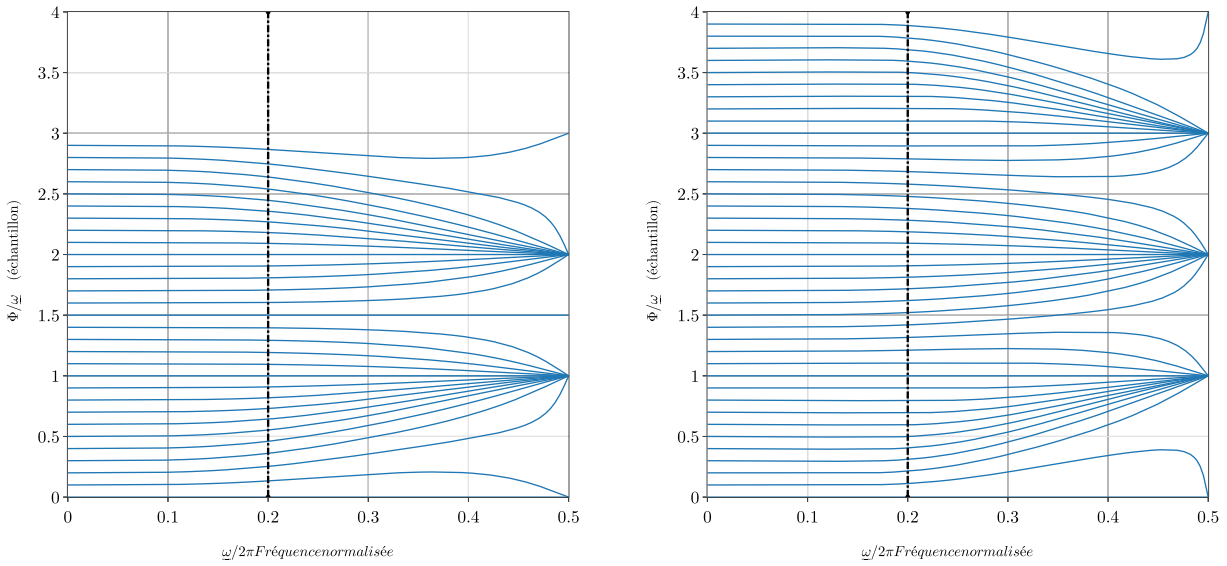
Dans la suite du document, on se propose d'analyser conjointement les deux types d'erreurs introduites par l'approximation, afin de déterminer dans quel intervalle de retards fractionnaires seront utilisés les filtres. Pour cette analyse menée dans les sections 3.4.3.2 et 3.4.3.3, l'erreur sera analysée exclusivement sur un intervalle d'un échantillon, puisque les erreurs observées sont rigoureusement nulles pour un nombre entier d'échantillons, et que le comportement est soit symétrique, soit antisymétrique par rapport au nombre entier d'échantillons de retard, en fonction de la parité de l'ordre du filtre [93].

3.4.3.2 Analyse du retard de phase des filtres

L'approche de retard fractionnaire par filtrage basé sur l'interpolation de Lagrange restant une approximation de la solution exacte analytique (non implémentable en pratique, pour les raisons évoquées plus haut), il est nécessaire de caractériser l'erreur introduite par cette approximation. Pour cela, on se propose tout d'abord d'étudier finement les retards de phase. Dans une situation idéale, le rapport $\phi/\underline{\omega}$ tracé sur la figure 3.12 doit rester constant sur toute la bande fréquentielle, jusqu'à la fréquence de Nyquist.

À titre d'exemple, ces retards de phases sont tracés la figure 3.12 pour un ensemble de retards, par pas de 0,1 échantillons, pour des filtres reposant sur l'utilisation de polynômes de Lagrange d'ordre 3 et 4.

L'analyse de ces deux graphiques permet de montrer que dans les deux cas, les retards de phases introduits sont symétriques par rapport à $\frac{N}{2}$, N étant l'ordre du polynôme de Lagrange utilisé pour construire le filtre. Même s'il est illustré ici exclusivement pour $N = 3$ ou $N = 4$, ce phénomène est vérifié pour tous les ordres des filtres à retards fractionnaires de Lagrange. Par ailleurs, on peut également remarquer que les retards introduits restent très proches des retards visés pour des pulsations normalisées faibles par rapport à la fréquence de Nyquist. En particulier, les retards de phase introduits pour $\underline{\omega}/2\pi < 0,2$ sont très proches d'une situation idéale, y compris pour ces ordres relativement



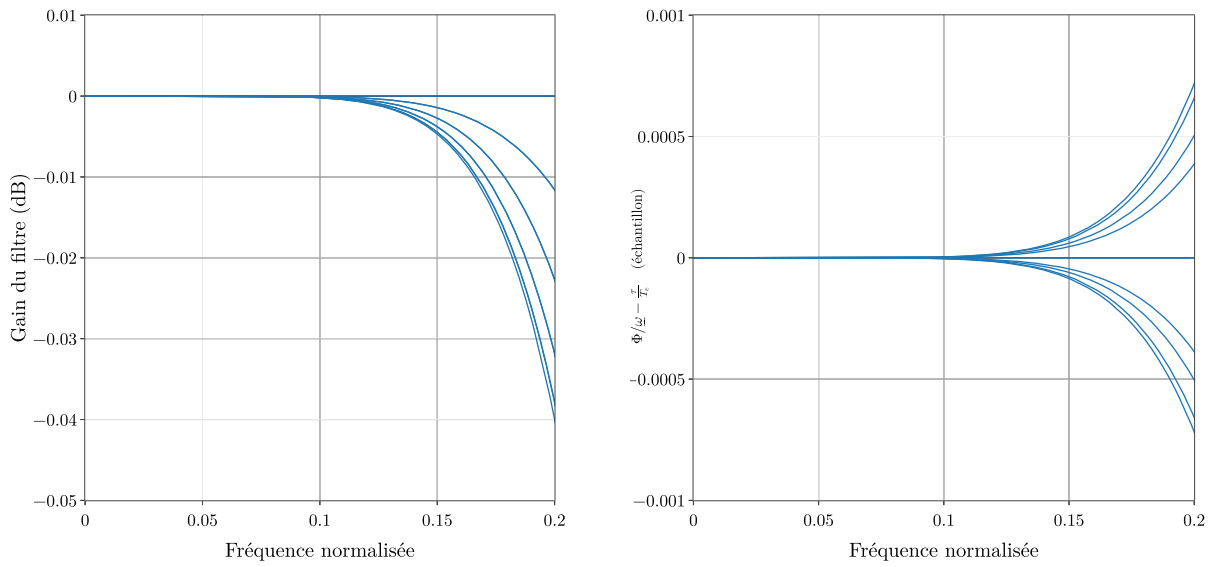
(a) Ordre 3 (b) Ordre 4

FIGURE 3.12 – Retard de phase des filtres d'ordre 3 (resp. 4) pour des retards allant de 0 à 3 (resp. 4) échantillons

faibles. Ce point est essentiel pour l'application qui nous intéresse, puisque dans notre cas, les signaux à déphaser possèdent un contenu fréquentiel exclusivement en dessous de 4 000 Hz, ce qui correspond à une pulsation normalisée de 0,18 environ pour une fréquence d'échantillonnage de 44 100 Hz. Au delà de cette pulsation normalisée, les erreurs d'approximation augmentent avec la fréquence, jusqu'à la fréquence de Nyquist, où les filtres de retards fractionnaires de Lagrange ne peuvent que déphaser d'un nombre d'échantillons entier. Ce point n'est pas limitant pour notre application, mais reste une des caractéristiques de ce type d'approche, et ce, quelque soit l'ordre N des filtres.

Par ailleurs, pour des considérations de continuité des retards de phase, il est préférable d'exploiter ce type de filtres pour un intervalle de retards de longueur 1 exclusivement, qui soit centré autour d'un retard entier. En effet, deux retards très proches doivent entraîner deux distorsions elles aussi très proches. Les discontinuitées sont donc à exclure. Il est donc préférable d'exploiter ces filtres sur l'intervalle $[\frac{N-1}{2}; \frac{N-1}{2} + 1[$ dans le cas des ordres pairs et sur l'intervalle $[\frac{N}{2}; \frac{N}{2} + 1[$ dans le cas des ordres impairs.

Sur ce principe, la figure 3.13 présente sur le domaine $\underline{\omega} = [0; 0.2]$ le diagramme de Bode pour un filtre d'ordre 7, sur l'intervalle $[3, 4[$. Cette figure permet d'observer qu'à cet ordre, on obtient une excellente approximation des retards fractionnaires dans le domaine de fréquence visé pour notre application : l'écart en amplitude au filtre idéal parfait est de moins de 0,04 dB sur cette gamme fréquentielle, et l'erreur sur la phase est inférieure à $7/10000^{ème}$ d'échantillons (ce qui justifie notre approche de pré-calcul par $1000^{èmes}$ d'échantillons). Toutefois, pour caractériser l'erreur commise pour un signal quelconque, il est plus visuel d'analyser l'erreur des moindres carrés entre le signal retardé par convolution et le signal théorique retardé.



(a) Module pour un ordre 7

(b) Erreur de retard de phases pour un ordre 7

FIGURE 3.13 – Diagramme de Bode pour des filtres d'ordre 7 avec des retards allant de 3 à 3,9 échantillons. Domaine fréquentiel restreint aux pulsations normalisées inférieures à 0.2

3.4.3.3 Analyse de l'erreur intégrale des moindres carrés

L'analyse en termes de précision sur le retard de phase introduit ayant été menée dans la section 3.4.3.2, il est également nécessaire d'observer l'effet sur le gain. Plutôt que de tracer un ensemble de modules calculés comme un diagramme de Bode, il est possible de caractériser l'erreur en amplitude introduite par l'approximation des filtres à retards fractionnaires par interpolation de Lagrange, en

calculant l'erreur au sens des moindres carrés entre le filtre de retard fractionnaire théorique idéal, et le filtre fréquentiel équivalent au filtrage interpolation de Lagrange, pour un ensemble de retards visés (voir figure 3.14).

Cette erreur correspond en fait à l'intégrale sur le domaine fréquentiel de l'erreur en amplitude obtenue dans la représentation de Bode, [158]. Soit $H_{Lag}(e^{j\underline{\omega}})$ la réponse du filtre constitué des coefficients du polynôme de Lagrange et $H_{sinc}(e^{j\underline{\omega}})$ la réponse du filtre idéal et théorique déterminé à partir de la fonction sinus cardinal non tronquée. L'erreur à la pulsation normalisée $\underline{\omega}$ est donc :

$$\zeta(e^{j\underline{\omega}}) = H_{sinc}(e^{j\underline{\omega}}) - H_{Lag}(e^{j\underline{\omega}}) \tag{3.37}$$

On peut donc calculer l'erreur intégrale des moindres carrés [158] :

$$\begin{aligned} E_{MC} &= \frac{1}{\pi}\int_0^{\pi} |\zeta(e^{j\underline{\omega}})|^2 d\underline{\omega} \qquad (3.38)\\ &= \sum_{k=0}^{\infty} |h_{Lag}(k) - h_{sinc}(k)|^2 \quad \text{(d'après le théorème de Parseval)} \\ &= \sum_{k=0}^{\infty} h_{sinc}(k)^2 + h_{Lag}(k)^2 - 2h_{sinc}(k) \times h_{Lag}(k) \\ E_{MC} &= 1 + \sum_{k=0}^{N} h_{Lag}(k)^2 - 2sinc(k-\tau) \times h_{Lag}(k) \qquad (3.39) \end{aligned}$$

L'erreur peut être ainsi tracée pour un filtre d'ordre 7. La figure 3.14 montre, comme précédemment observé pour les phases, une symétrie des erreurs par rapport au retard $\tau = \frac{N}{2}$. De plus, en choisissant un intervalle d'un seul échantillon, centré autour de la valeur $\tau = \frac{N}{2}$, on s'assure d'une erreur minimale au sens des moindres carrés sur tout l'intervalle.

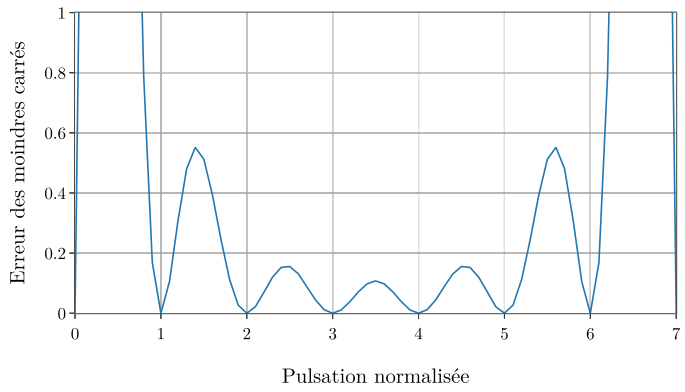


FIGURE 3.14 – Erreur intégrale des moindres carrés pour un filtre d'ordre 7

3.4.3.4 Comparaison entre les approches de retards fractionnaires par interpolation de Lagrange et par troncature du sinus cardinal

Comme exposé dans la section 3.4.2.2, le filtrage de retards fractionnaires par interpolation de Lagrange revient à opérer une minimisation des erreurs fréquentielles sur le domaine, avec des fonctions *n-plates*. C'est la raison pour laquelle nous avons choisi dans la section précédente d'exploiter un critère de minimisation de l'erreur quadratique entre un signal filtré et sa version décallée idéalement, afin de déterminer l'ordre des polynômes de Lagrange utilisés dans le cadre de nos applications.

Ce même critère est également utilisé dans cette section pour comparer les performances entre une approche basée sur l'interpolation de Lagrange et une approche plus classique reposant sur la troncature d'un sinus cardinal échantillonné. Pour cette analyse, un signal de référence $s(t)$ est échantillonné à deux fréquences différentes. À une fréquence d'échantillonnage normale $f_e = \frac{1}{T_e}$, qu'on appellera signal *sous-échantillonné* : $s[n]$, et à une autre fréquence plus élevée $f_{se} = \frac{1}{T_{se}}$, qu'on appellera signal *sur-échantillonné* : $s_{se}[m]$. On se propose d'utiliser $f_e = 44100$ Hz et $f_{se} = 10 f_e$.

$$\left.\begin{array}{ll} s[n] & = s(nT_e)\ ,\ n \in [\![0, f_e]\!] \\ s_{se}[m] & = s(mT_{se})\ ,\ m \in [\![0, f_{se}]\!] \end{array}\right\} \quad f_{se} = 10 f_e$$

Le signal sur-échantillonné est décalé d'un nombre Δ entier d'échantillons (ce décalage est donc exact). Le signal sous-échantillonné est quant à lui décalé d'un nombre non entier δ d'échantillons par rapport à T_e, soit en utilisant la méthode d'interpolation de Lagrange, soit en utilisant la méthode du sinus cardinal tronqué échantillonné (dans les deux cas, ces méthodes ne représentent donc qu'une approximation, pour lesquelles on cherche à déterminer laquelle est la plus précise). Pour comparer les solutions approchées – à partir des signaux échantillonnés à f_e – aux solutions idéales – à partir des signaux sur-échantillonnés à f_{se} – les décalages δ et Δ choisis pour l'analyse respectent la relation suivante : $\Delta \cdot T_{se} = \delta \cdot T_e$

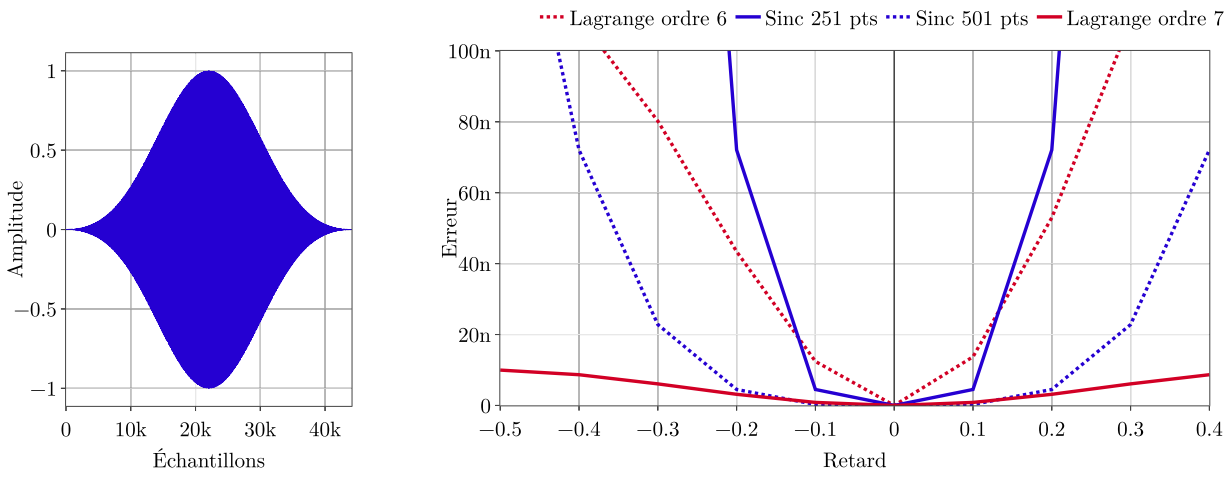
$$\begin{cases} s'_{\delta}[n] & = s((n-\delta)T_e)\ ,\ n \in [\![0, f_e]\!]\ ,\ \delta = -0.5, ..., 0.5 \\ s'_{\Delta,se}[m] & = s((m-\Delta)T_{se})\ ,\ m \in [\![0, f_{se}]\!]\ ,\ \Delta = -5...5 \end{cases}$$

En outre, compte tenu de la discussion effectuée en section 3.4.3.1, afin de pouvoir comparer différents ordres de filtres, les retards fractionnaires sont placés dans leurs intervalles optimaux respectifs : ceux-ci sont centrés autour de $[\frac{N-1}{2}; \frac{N-1}{2}+1[$ dans le cas des ordres pairs et de $[\frac{N}{2}; \frac{N}{2}+1[$ dans le cas des ordres impairs. Ainsi, un retard indiqué comme valant 0.3 vaudra en réalité 2.3 échantillons pour les ordres 3 et 4, mais 3.3 pour les ordres 5 et 6.

Comme exposé précédemment, pour quantifier la qualité de l'approximation, une erreur quadratique moyenne est ensuite calculée entre le signal idéalement déphasé $s'_{\Delta,se}[nT_e/T_{se}]$ et le signal déphasé par filtrage $s'_{\delta}[n]$. Pour déterminer l'ordre N qui sera retenu dans le cadre de notre projet, nous avons choisi de prendre le premier ordre qui permet d'obtenir une erreur quadratique moyenne inférieure à 10^{-7}, sur tout le domaine de retards possibles.

Sur la figure 3.15, le tracé de l'erreur quadratique moyenne pour différents types de filtres est présenté, pour un signal d'entrée correspondant à un sinus à 4 100 Hz, apodisé par une fenêtre de Blackman d'un dixième de seconde (fig. 3.15(a)). La fréquence du sinus a volontairement été choisie de l'ordre de grandeur de la fréquence maximale utilisée pour les signaux d'entrée du réseau, puisque l'on a précédemment constaté que les erreurs d'approximations sont plus élevées aux hautes fréquences,

tant en phase qu'en amplitude (voir section 3.4.3.2). En analysant les résultats obtenus pour différents ordres de polynômes de Lagrange présentés sur la figure 3.15(b), il résulte que l'ordre 7 satisfait pleinement au critère fixé, et ce pour tous les retards possibles :



(a) Signal de référence

(b) Maximum de l'erreur pour différentes fonction de retard

FIGURE 3.15 – Erreur maximale obtenue pour différents retards en convoluant le signal de référence 3.15(a) avec les polynômes de Lagrange d'ordre 6 et 7 (en rouge, respectivement en pointillés et en continu) et avec un sinus cardinal apodisé par une fenêtre de Blackman, avec 251 points et 751 points (en bleu, respectivement en pointillés et en continu).

Par ailleurs, en comparant l'erreur quadratique moyenne obtenue en utilisant la méthode d'interpolation de Lagrange et la méthode du sinus cardinal tronqué échantillonné, on peut mettre en évidence un avantage fondamental de l'approche reposant sur le filtrage par interpolation de Lagrange. Avec une convolution de seulement 8 points pour l'interpolation temporelle (à l'ordre 7), les résultats sont meilleurs qu'en utilisant des sinus cardinaux échantillonnés et tronqués de plus de 500 points. La réduction du nombre d'opérations et de taille de stockage de réponses impulsionnelles dans des tenseurs parcimonieux pour la convolution temporelle avec un signal d'une seconde échantillonné à 44 100 Hz est de l'ordre de 98%[5] grâce à cette forte réduction de la longueur du filtre à réponse impulsionnelle finie, tout en offrant une précision plus importante sur tout le domaine. Dans le cadre d'un calcul massif d'auralisation et de stockage de réponses impulsionnelles pour un nombre extrêmement conséquent de sources images, cette réduction offerte est primordiale, justifiant ainsi l'effort de développement

5. Cette énorme réduction du nombre de coefficient par rapport au sinus cardinal est uniquement valable car la fréquence du signal est très inférieure à la fréquence d'échantillonnage. On est donc dans la partie optimale du filtre.

pour implémenter cette méthode pour le calcul sur GPU de réponses impulsionnelles.

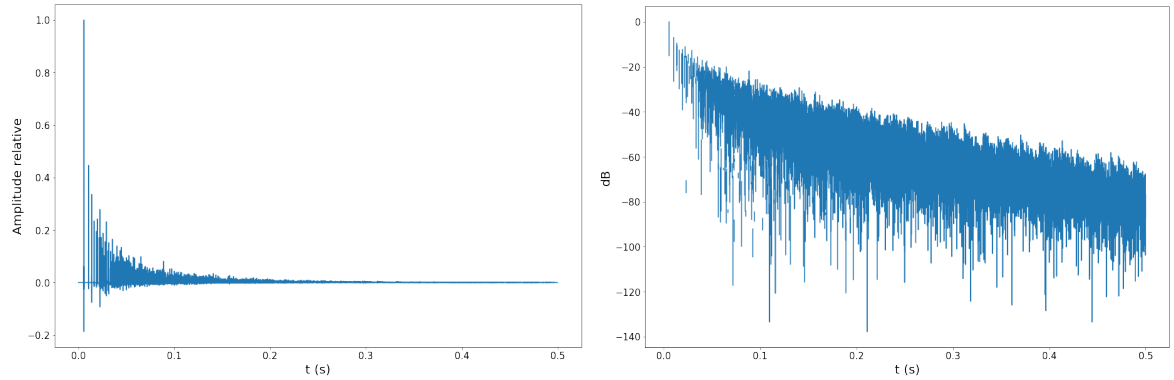
3.4.4 Résultats

En combinant la méthode des sources images avec une évaluation fine des retards grâce au calculs des retards fractionnaires, les réponses impulsionnelles de salles présentent la précision temporelle nécessaire pour le calcul numérique rapide et précis d'auralisation massive sur des antennes compactes.

De plus, le calcul optimisé sur GPU de ces réponses impulsionnelles permet de lever l'un des reproches communément adressé à l'encontre de la méthode des sources images : son temps de calcul pour les ordres élevés de réflexion, c'est à dire pour la partie tardive de la réverbération. En effet, le calcul de plusieurs dizaines de milliers de réponses impulsionnelles multicanales ne prend que quelques heures sur une carte graphique Nvidia 1080Ti du serveur de calcul que nous utilisons pour y exécuter nos programmes Tensorflow.

La figure 3.16(a) permet de vérifier que les réponses impulsionnelles obtenues ne présentent aucune dérive à basse fréquence de la composante *continue*. Ce problème de dérive est connu, et apparaît lorsque les retards temporels sont tronqués trop grossièrement. Pour corriger cet artefact, un filtre passe haut est communément utilisé pour recentrer les valeurs de la réponse impulsionnelle autour de 0 [139, 162]. D'autres auteurs ont quant à eux proposé, sur des arguments ayant une validité physique plutôt discutable, d'affecter à chaque coefficient de réflexion un changement de signe alternatif pour compenser ce défaut. Avec l'approche proposée, il est inutile de réaliser ce type de compensation, puisque la réponse impulsionnelle possède déjà un profil réaliste grâce au soin apporté pour la construire.

La réponse impulsionnelle présentée en échelle linéaire à la figure 3.16(a) (respectivement logarithmique à la figure 3.16(b)) correspond à la salle de cours présentée plus haut en section 3.2, avec un coefficient d'absorption constant sur toutes les parois valant 0,312. Ce coefficient est obtenu en inversant la formule d'Eyring [136] avec un temps de réverbération de 0,5 s. On peut aisément observer que la décroissance logarithmique observée en figure 3.16(b) ne correspond pas exactement à une décroissance de 60 dB en 0,5 s comme le suppose la théorie statistique d'Eyring. Une lecture graphique



(a) Représentation linéaire (b) Représentation logarithmique

FIGURE 3.16 – Réponse impulsionnelle simulée pour une durée de réverbération de 0,5 s

de la décroissance de la pente suggère que la durée de réverbération est plutôt de 0,6 s. Cet écart n'est pas un problème en soit dans le cadre de la simulation d'un environnement réverbérant **quelconque**, et ne fait que révéler que les approches statistiques et géométriques en acoustique des salles diffèrent. En revanche, il rend impossible l'utilisation de cette méthode pour la constitution d'un modèle d'une salle réverbérante **particulière** en exploitant un jeu de données d'entrées de coefficients d'absorptions issus de la théorie statistique. Pour y remédier, il est possible d'utiliser un coefficient d'absorption modifié, soit à partir d'une expression analytique [131], soit à partir d'une correction itérative des coefficients d'absorption [139]. Cette deuxième approche est développée en section 3.5.

3.5 Optimisation du paramètre de coefficient d'absorption de parois pour la modélisation par sources images

3.5.1 Démarche usuelle

Lorsque l'on veut modéliser une salle en particulier, les grandeurs physiques communément accessibles sont les dimensions géométriques de la pièce, et la durée de réverbération par bandes fréquentielles. La connaissance fine des matériaux composant les parois et de leur impédance, ou de la dépendance angulaire de leur coefficient d'absorption, est en revanche très rarement connue, même si c'est l'une des clés de l'amélioration de la modélisation en acoustique des salles [131]. À partir des seules dimensions géométriques et de la durée de réverbération, le formalisme de l'acoustique statistique peut être utilisé pour déterminer un coefficient d'absorption moyen (moyenne angulaire et moyenne sur toutes les parois de la salle), en inversant la formule de Sabine comme le propose la

norme [136]. L'exploitation de la formule d'Eyring permet d'étendre le domaine de validité pour des salles aux parois plus absorbantes (d'autres extensions existent, mais nous nous restreindrons dans le cadre de ce manuscrit à ces deux approches) :

$$T_r = \frac{0,16 \cdot V}{4 \cdot m \cdot V - S \cdot ln(1-\alpha)} \tag{3.40}$$

Si on néglige l'amortissement de l'air [6] m, on peut en déduire le coefficient d'absorption α en fonction du temps de réverbération T_r, du volume V et de la surface S des murs de la salle. Si la durée de réverbération est connue par bandes fréquentielles, cela permet d'obtenir la dépendance fréquentielle de ce coefficient d'absorption moyen :

$$\alpha = 1 - \exp\left(\frac{0,16 \cdot V}{-S \cdot Tr}\right) \tag{3.41}$$

L'expérience et la littérature [139] prouvent que lorsque le coefficient est déterminé de cette manière, et utilisé comme paramètre d'entrée pour le modèle de sources images, la durée de réverbération T_r de la salle simulée par la méthode des sources images présente une valeur biaisée par rapport à la valeur de la durée de réverbération qui a permis d'obtenir le coefficient α à l'aide de la formule 3.41. Cette erreur d'approximation peut être améliorée afin d'atteindre une simulation plus réaliste, notamment en utilisant un coefficient de réflexion modifié, issu d'un calcul analytique, comme proposé dans [131]. Cependant, cette approche analytique nécessite la connaissance de paramètres intrinsèques aux matériaux - la plupart du temps inaccessibles - et même avec cette connaissance, le calcul n'est pas trivial et peut impacter lourdement le temps de calcul. C'est la raison pour laquelle nous proposons ici une approche numérique rapide et itérative, inspirée des travaux de Lehman [139].

Dans leurs travaux publiés en 2007 et 2008 [139, 164], Lehmann et Johansson ont proposé une méthode itérative pour déterminer le coefficient d'absorption nécessaire à l'obtention d'un T_r voulu dans le cadre de la méthode de sources images. L'idée est ici de dénombrer le nombre de réflexions sur chaque paroi délimitant la salle simulée, plutôt que de faire une approche statistique comme pour obtenir les formules de Sabine ou d'Eyring [130]. Les approximations proposées dans l'article initial

6. ce qui reste valide si on cherche à modéliser des salles de taille raisonnable. Pour des salles de concert, l'influence de ce paramètre devient non négligeable, notamment en hautes fréquences [163]

pour résoudre ce problème, parfois trop importantes, nous ont amené à proposer un complément à cette démarche dans le cadre de cette thèse.

3.5.2 Présentation de la démarche proposée

La démarche proposée dans le cadre de cette thèse est similaire à celle de l'article cité précédemment [139], mais en y apportant des corrections sur certaines approximations, et en proposant une extension numérique à 3 dimensions différente de celle proposée initialement par les auteurs. L'approche itérative proposée implique de réaliser un calcul rapide de la décroissance énergétique dans la salle, en présupposant le coefficient α de la salle, puis d'en déduire une approximation du T_r que l'on obtiendrait avec la méthode des sources images, sans pour autant réaliser le calcul de sources images complet. Cette estimation rapide permet ensuite d'en déduire, par comparaison au T_r cible de la salle à modéliser, les modifications à réaliser sur le coefficient d'absorption utilisé pour la simulation numérique pour s'approcher au mieux du T_r cible. La méthode est itérative, puisque le coefficient α est modifié autant de fois que nécessaire pour obtenir un estimateur rapide satisfaisant. Une fois les valeurs satisfaisantes des coefficients d'absorption α des parois obtenues, la simulation par sources images en tant que telle peut être réalisée. L'intérêt de cette approche est qu'on peut économiser un temps de calcul précieux grâce à l'estimation rapide, puisqu'on évite ainsi de réaliser N simulations par sources images pour modifier le α d'entrée afin d'approcher au mieux la durée de réverbération T_r de la salle modélisée.

L'estimation rapide du Tr qu'on obtiendrait par une simulation de réponses impulsionnelles de la salle par la méthode des sources images repose sur un calcul approché et rapide $\tilde{h}_p(t)$ de la réponse impulsionnelle de la salle basé sur la puissance et non sur la pression acoustique. La durée de réverbération de la salle est ensuite simplement estimée à l'aide de cette estimateur de réponse impulsionnelle en puissance $\tilde{h}_p(t)$, approchée comme la durée nécessaire à la dissipation de 99,9% de l'énergie (*i.e.* une diminution de 60 dB après l'interruption de la source [130, 132, 136]). Cette décroissance énergétique est classiquement calculée grâce à la formule de Schroeder du décrément énergétique [165] :

$$E(t) = 10 \cdot \log \left(\frac{\int_t^\infty h^2(\xi) d\xi}{\int_0^\infty h^2(\xi) d\xi} \right) \tag{3.42}$$

Dans le cas de l'estimateur de réponse impulsionnelle en puissance, un estimateur de cette décroissance énergétique est quant à elle obtenue grâce à la formule suivante :

$$\tilde{E}(t) = 10 \cdot \log \left(\frac{\int_t^\infty \tilde{h_p}(\xi) d\xi}{\int_0^\infty \tilde{h_p}(\xi) d\xi} \right) \tag{3.43}$$

3.5.3 Cas en 2D

La réponse impulsionnelle approchée $\tilde{h_p}(t)$ correspond à la somme des contributions des sources images arrivant à l'instant t au microphone. Plutôt que de calculer l'intégralité de la réponse impulsionnelle par la méthode des sources images, ici, on suppose que les sources images concernées sont celles qui sont *suffisamment proches* du cercle de rayon $\rho(t) = c \cdot t$ (noté ρ dans la suite dans un soucis d'allègement des notations). Un exemple est donné dans en figure 3.17 pour le cadran supérieur du plan. On suppose ici que ce rayon soit très supérieur aux dimensions de la pièce : $\rho >> \max\{L_x; L_y\}$.

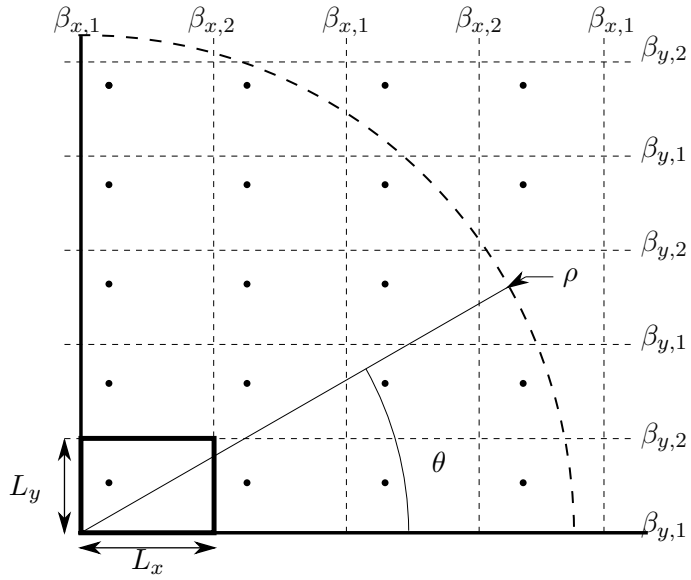


FIGURE 3.17 – Représentation schématique 2D d'une salle contenant une source et ses images fictives.

Soit une source image appartenant au cercle de rayon ρ, située à un angle θ : en partant de l'équation 3.15, la puissance acoustique mesurée par le microphone de la salle initiale [139, 166] sera liée à la distance ρ, au coefficient de réflexion en amplitude de chaque mur β et au nombre de parois

impliquées dans les réflexions W_x et W_y de l'onde émise par la source, suivant les directions x et y :

$$P(\rho,\theta) = \frac{(\beta_{x,1}^2)^{\frac{W_x}{2}}(\beta_{x,2}^2)^{\frac{W_x}{2}}(\beta_{y,1}^2)^{\frac{W_y}{2}}(\beta_{y,2}^2)^{\frac{W_y}{2}}}{(4\pi\rho)^2} \tag{3.44}$$

Pour estimer le nombre de réflexions dans les directions x et y impliquées dans ce quart de plan, il suffit d'utiliser les formules de trigonométrie suivantes (voir figure 3.17) : [7]

$$\begin{cases} W_x(\rho,\theta) = \frac{\rho\cos(\theta)}{L_x} \\ \\ W_y(\rho,\theta) = \frac{\rho\sin(\theta)}{L_y} \end{cases} \tag{3.45}$$

On peut alors en déduire que la réponse impulsionnelle approchée vue d'un point de vue énergétique, notée $\tilde{h}_p(t)$, est la somme des puissances venant de toutes les sources suffisamment proches du cercle, en posant I_c l'ensemble des indices des sources comptabilisées autour du cercle de rayon $\rho = c \times t$:

$$\tilde{h}_p(t) = \sum_{i\in I_c} P(\rho,\theta_i) \tag{3.46}$$

Si on voit le résultat de l'équation 3.46 comme une somme de Riemann, la somme discrète peut devenir une intégrale sur un domaine continu paramétré par l'angle θ [139] :

$$\begin{aligned} \tilde{h}_p(t)\cdot\Delta\theta &= \sum_{i\in I_c} P(\rho,\theta_i)\cdot\Delta\theta \\ \tilde{h}_p(t)\cdot\Delta\theta &\approx \int_0^{2\pi} P(\rho,\theta_i)d\theta \\ \tilde{h}_p(t) &\approx \frac{4}{\Delta\theta}\int_0^{\frac{\pi}{2}} P(\rho,\theta)d\theta \end{aligned} \tag{3.47}$$

Soit N_{2D} le nombre de sources comptabilisées sur tout le cercle, le petit écart angulaire entre deux sources est :

$$\Delta\theta = \frac{2\pi}{N_{2D}} \tag{3.48}$$

7. Dans l'article ayant inspiré notre approche, les valeurs W_x et W_y étaient approximées comme des fonctions linéaires de θ, ce qui est une approximation trop grossière du problème posé ici.

De plus, en utilisant l'hypothèse $\rho >> \max\{L_x; L_y\}$, on peut approximer la taille moyenne des murs par $\tilde{r} = \frac{L_x+L_y}{2}$. Ainsi on peut définir le nombre total de sources comme le rapport entre la taille du cercle et la taille caractéristique de la salle :

$$N_{2D} = \frac{2\pi\rho}{\tilde{r}} \tag{3.49}$$

En combinant ces trois dernières équations, on obtient l'énergie de la réponse impulsionnelle dans le cas d'une salle à deux dimensions :

$$\boxed{\tilde{h}_p(t) = \frac{4\rho}{\tilde{r}} \int_0^{\frac{\pi}{2}} P(\rho, \theta_i) d\theta} \tag{3.50}$$

En exploitant cet estimateur, on peut alors déterminer la décroissance énergétique et obtenir rapidement un estimateur du T_r obtenu par la méthode des sources images, en utilisant les formules 3.43 et 3.44.

3.5.4 Extension à 3 dimensions

Dans l'article de Lehman et Johansson ayant inspiré nos développements [139], les auteurs ont proposé une formule analytique approchée pour déterminer $\tilde{h_p}(t)$, en utilisant une approximation linéaire grossière de W_x, W_y, et W_z, sur le même principe qu'en 2D. Même si cette approximation présente l'avantage de pouvoir déterminer analytiquement l'intégrale 3.42, nous proposons ici un développement numérique, avec une approximation géométrique moins grossière sur les nombres de réflexions impliquées dans le calcul de $\tilde{h}_p$.

Par analogie aux développements proposés en 2 dimensions, nous proposons toujours ici de déterminer un estimateur $\tilde{h}_p(t)$ de la réponse impulsionnelle en puissance, en sommant à un instant t les sources images suffisamment proche de la sphère de rayon $\rho(t)$. Dans le cas 3D, le petit écart angulaire est une fraction d'angle solide $\Delta\Omega$, et l'équation 3.47 devient alors :

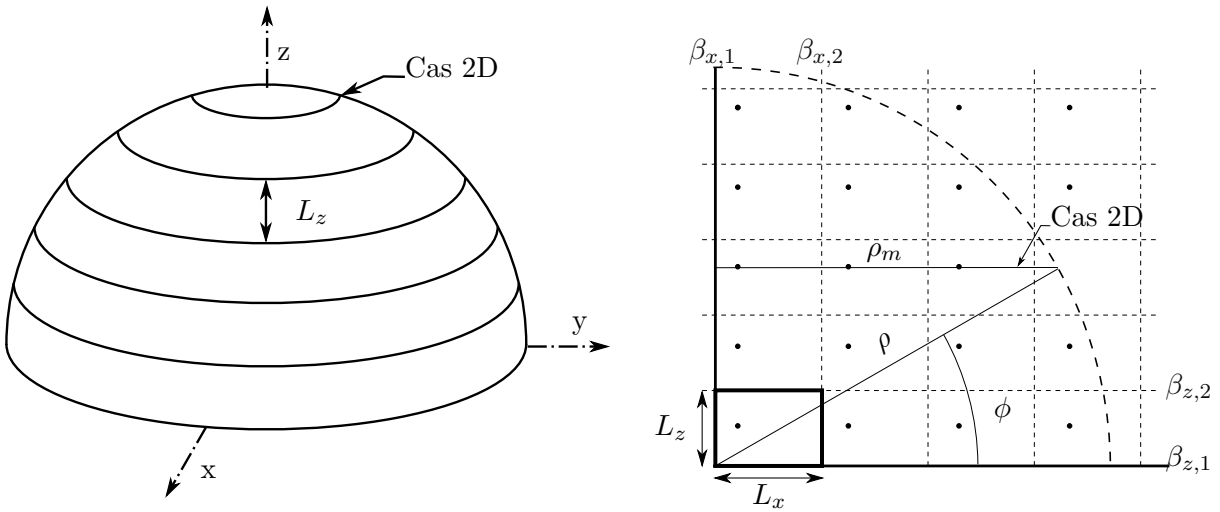
$$\tilde{h}_p(t) = \frac{8}{\Delta\Omega} \int_{\theta=0}^{\frac{\pi}{2}} \int_{\phi=0}^{\frac{\pi}{2}} P(\rho, \theta, \phi) sin(\phi) d\theta d\phi \tag{3.51}$$

En notant N_{3D} le nombre de sources comptabilisées sur l'ensemble de la sphère, la fraction d'angle solide $\Delta\Omega$ vaut :

$$\Delta\Omega = \frac{4\pi}{N_{3D}} \tag{3.52}$$

Pour dénombrer les sources présentes sur la sphère de rayon $\rho(t)$, on se propose de voir le cas en trois dimensions comme une superposition de cas en deux dimensions, où le rayon de chaque cercle varie en fonction de la hauteur (cf. figure 3.18(a)). En notant $\rho_m = \sqrt{\rho^2 - (m \cdot L_z)^2}$ le rayon du m-ième cas en deux dimensions à l'altitude mL_z , le nombre de sources pour cette hauteur précise vaut :

$$N_{2D,m} = \frac{2\pi}{\tilde{r}}\sqrt{\rho^2 - (m \cdot L_z)^2} \tag{3.53}$$



(a) Cas 3D vu comme superposition de cas 2D.

(b) Coupe selon le plan (x0z) du cas 3D

FIGURE 3.18 – Représentations schématiques d'une salle contenant une source et de ses images fictives, dans le cas 3D.

Pour compter le nombre de sources présentes sur la surface de la sphère, il suffit de sommer toutes les couches de deux dimensions. En posant la hauteur de la salle comme L_z, et en utilisant les propriétés de symétrie du problème, on obtient :

$$\begin{aligned} N_{3D} &= 2\sum_{m=0}^{\rho/L_z} N_{2D,m} \\ &= 2\sum_{m=0}^{\rho/L_z} \frac{2\pi}{\tilde{r}}\sqrt{\rho^2-(m\cdot L_z)^2} \end{aligned} \tag{3.54}$$

Cette approche intuitive peut être complétée en passant d'une somme finie à une intégrale. En posant de plus le changement de variable $dm = \frac{dz}{L_z} = \frac{\rho\cos(\phi)d\phi}{L_z}$, on obtient une formule générale :

$$\begin{aligned} N_{3D} &= 2\int_0^{\frac{\pi}{2}} \frac{2\pi}{\tilde{r}\cdot L_z}\sqrt{\rho^2\cdot(1-\sin^2(\phi))}\cdot\rho\cos(\phi)d\phi \\ &= \frac{\rho^2 4\pi}{\tilde{r}\cdot L_z}\int_0^{\frac{\pi}{2}}\cos^2(\phi)d\phi \\ N_{3D} &= \frac{\rho^2\pi^2}{\tilde{r}\cdot L_z} \end{aligned} \tag{3.55}$$

L'équation 3.51 se généralise en trois dimensions en posant :

$$\begin{cases} W_x(\rho,\theta) = \frac{\rho\cos(\theta)\cos(\phi)}{L_x} \\ W_y(\rho,\theta) = \frac{\rho\sin(\theta)\cos(\phi)}{L_y} \\ W_z(\rho,\theta) = \frac{\rho\sin(\phi)}{L_z} \end{cases} \tag{3.56}$$

En injectant ce résultat dans l'équation 3.51 extrapolée à trois dimensions et à l'aide de la formule 3.52, on peut obtenir la valeur approchée de la RIR énergétique. On remarquera que la diminution énergétique de la puissance en ρ^2 est compensée par le nombre de sources qui croît en ρ^2 :

$$\tilde{h}_p(t) = \frac{1}{2L_z\tilde{r}}\int_{\theta=0}^{\frac{\pi}{2}}\int_{\phi=0}^{\frac{\pi}{2}}\prod_{\substack{i=x,y,z\\j=1,2}}(\beta_{i,j})^{W_i} sin(\phi)d\theta d\phi \tag{3.57}$$

3.5.5 Détermination en 3 dimensions de la durée de réverbération à l'aide de cette estimation rapide

Sur le même principe que ce qui a été exposé en 2 dimensions, l'estimateur $\tilde{h}_p(t)$ de la réponse impulsionnelle de la salle basé sur la puissance étant déterminé à l'aide de l'équation 3.57, la décroissance énergétique se calcule directement en utilisant la méthode usuelle du décrément énergétique [139,165]. Cette formule doit toutefois être modifiée pour prendre en compte l'approximation faite pour comptabiliser les sources : $\rho >> \max\{L_x; L_y; L_z\}$. On se propose donc de poser que cette condition est vérifiée $\forall \rho > 3 \times r_{moy}$ avec $r_{moy} = \frac{L_x+L_y+L_z}{3}$, ce qui entraîne que l'approximation n'est valable qu'à partir d'un certain temps $t_0 = 3\frac{r_{moy}}{c}$. Sans être fondamentale, la valeur de $3 \cdot r_{moy}$ doit toutefois être un compromis :

- elle doit être suffisamment grande pour pouvoir faire raisonnablement l'approximation que la distance parcourue par la source image jusqu'au microphone est grande par rapport aux tailles caractéristiques de la pièce
- elle doit rester suffisamment faible pour ne pas trop retarder l'instant t_0 à partir duquel l'évaluation de la décroissance énergétique peut se faire.

Sous ces conditions, l'équation 3.42 devient donc :

$$\tilde{E}(t > t_0) = 10 \cdot \log\left(\frac{\int_t^\infty \tilde{h}_p(\xi)d\xi}{\int_{t_0}^\infty \tilde{h}_p(\xi)d\xi}\right) \tag{3.58}$$

La valeur du T_r est finalement obtenue grâce à la méthode proposée dans la norme [136] : une régression linéaire est calculée à partir des valeurs comprises entre -5 dB et -30 dB [8] , ce qui permet de calculer un T_{r25}, qui est ensuite converti en T_{r60}.

8. La valeur de -30 dB est préférée à celle de la norme de -25 dB, car dans le cas de simulation, aucun bruit de fond n'empêche de prendre un écart en dB plus grand, ce qui permet une meilleure estimation de la régression linéaire.

3.5.6 Validation de la méthode : estimation de la durée de réverbération d'une salle simulée avec la méthode des sources images

L'estimation de la décroissance de l'énergie de la réponse impulsionnelle d'une salle grâce à la méthode proposée permet ainsi de trouver, par itérations successives, le coefficient d'absorption optimal pour obtenir une durée de réverbération cible avec la simulation de réponses impulsionnelles de salles par sources images. En effet, comme illustré en section 3.4.4, le fait d'utiliser un coefficient d'absorption obtenu à partir de l'inversion d'une formule issue de l'acoustique statistique (comme la formule de Sabine ou d'Eyring), ne permet pas d'obtenir la durée de réverbération voulue dans une salle simulée à partir de la méthode des sources images, issue d'une approche d'acoustique géométrique.

Lors d'une conférence, E. Lehmann a détaillé sur un grand nombre de géométries différentes la qualité de l'estimation du temps de réverbération d'une salle simulée par la méthode des sources images grâce à son approche [164], en la comparant à un ensemble de formules d'estimations de Tr issues de l'acoustique statistique. L'objectif n'est pas ici de reprendre tous ces exemples, mais simplement de mettre en évidence que, grâce au dénombrement des sources images proposé et aux corrections apportées pour l'estimation de $\tilde{h}_p$, la méthode introduite par E. Lehmann se trouve améliorée.

Pour ce faire, deux expériences complémentaires sont menées. Tout d'abord, un coefficient d'absorption arbitraire est choisi pour simuler la réponse impulsionnelle d'une salle. Ce coefficient d'absorption est constant sur toutes les surfaces de la pièce, sans perte de généralité pour ce calcul. La durée de réverbération de la salle est ensuite calculée grâce à plusieurs approches :

— grâce à des approches statistiques en utilisant les formules de Sabine et d'Eyring,
— en utilisant la méthode proposée par É. Lehmann, dans sa version originale,
— en utilisant la méthode itérative inspirée des travaux de Lehmann, modifiée selon les éléments décrits dans les sections précédentes.

Les résultats obtenus sont ensuite comparés au calcul *direct* du temps de réverbération de la salle grâce à la méthode des sources images. Le tableau 3.5 résume les résultats obtenus grâce à ces différentes

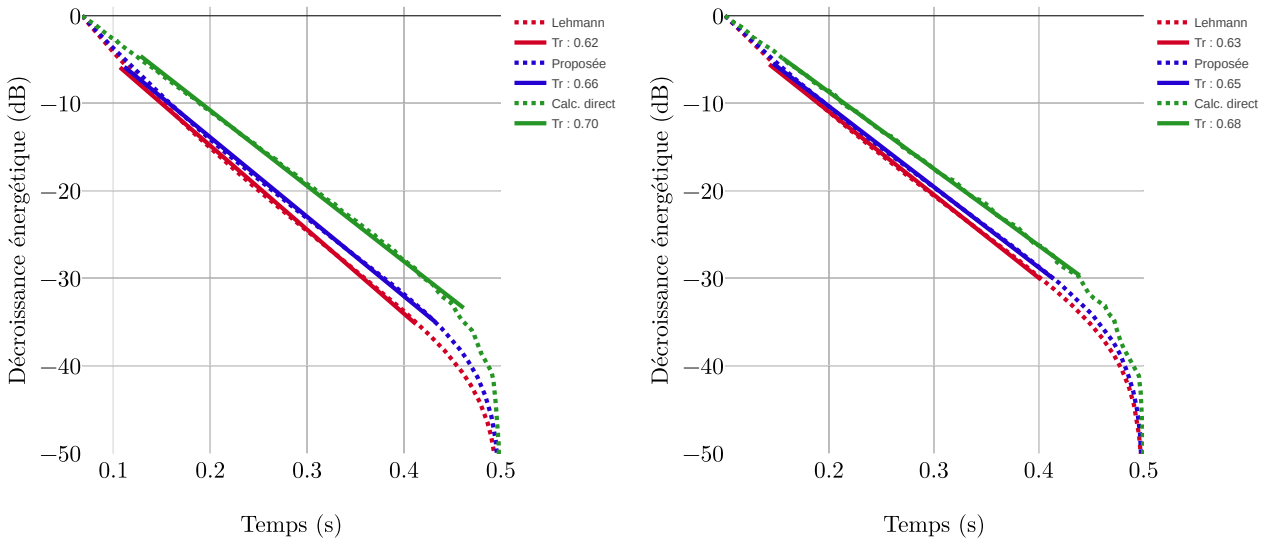
méthodes. Il apparaît que dans ce cas, le T_r est toujours sous-estimé avec les méthode d'acoustique statistique par rapport au résultat obtenu par simulations de sources images. Au contraire, les T_r prédits grâce à la méthode de Lehmann originale et celle proposée dans ce travail de thèse de doctorat, ne s'écartent que de respectivement 10% et 5% environ. Ce résultat illustre ainsi l'intérêt de la méthode proposée par É. Lehmann, mais il valide ainsi le fait que les optimisations proposées dans les sections précédentes permettent une meilleure estimation du T_r obtenue grâce au calcul des sources images.

Salle	1	2
Dimensions (m)	7 ; 4 ; 2,9	10 ; 7 ; 3.7
α	0,22	0,31
Tr sources images (s)	0,7	0,68
Tr Sabine (s)	0,5 (29%)	0,5 (26%)
Tr Eyring (s)	0,44 (37%)	0,42 (38%)
Tr Lehmann (s)	0,62 (11%)	0,63 (7%)
Tr proposé (s))	0,66 (6%)	0,65 (4%)

Tableau 3.5 – Récapitulatif des salles testées, ainsi que de leur T_r évaluée par différentes méthodes

Pour compléter ces résultats, la figure 3.19 présente les décroissances énergétiques (en pointillés) ainsi que l'estimation de cette décroissance par régression linéaire (en traits pleins), obtenues à partir d'un calcul énergétique depuis une réponse impulsionnelle de salle modélisée directement avec la méthode des sources images (vert), ou à partir de la méthode proposée par Lehmann, dans sa version originale (en rouge), ou avec les modification suggérées dans les sections précédentes (en bleu).

Connaissant le coefficient d'absorption des parois, il est donc possible grâce à la méthode proposée d'estimer la durée de réverbération de la salle simulée par la méthode des sources images. Mais la connaissance *a priori* du T_r de la salle à partir d'un coefficient d'absorption donné, n'a pas d'intérêt en soit pour notre application. En revanche, le fait de pouvoir déterminer le coefficient d'absorption modifié à utiliser comme paramètre d'entrée d'une simulation par sources images, pour obtenir une



(a) Salle 2 (b) Salle 3

FIGURE 3.19 – Comparaison des décroissances énergétiques (traits pointillés) et leurs régressions linéaires (trait plein), dans deux salles, à partir d'un calcul de sources images (vert), ou d'une estimation du nombre de sources images d'après la méthode de Lehmann [139] (rouge) et d'après la méthode proposée (bleu)

durée de réverbération cible dans une salle de géométrie donnée, permet de simuler au plus près les caractéristiques acoustiques d'une salle. C'est donc dans ce sens que nous proposons d'exploiter la méthode proposée.

3.5.7 Détermination des coefficients d'absorption pour l'obtention d'une durée de réverbération cible

Dans le cas de la constitution d'une base de données simulées en vue d'un apprentissage supervisé pour une tâche de localisation de sources acoustiques, le fait de pouvoir simuler fidèlement les caractéristiques d'une salle particulière apparaît comme primordial. Cette section présente et évalue une méthodologie pour atteindre cet objectif à partir de l'estimation de la décroissance énergétique de la réponse impulsionnelle de puissance (voir équation 3.58).

Dans cette section, on suppose que la géométrie de la salle est connue, ainsi que sa durée de

réverbération cible, notée $T_{r,c}$. Ces deux paramètres étant mesurables simplement, il apparaît donc raisonnable de n'exploiter que ces grandeurs pour simuler une salle. L'objectif est de trouver les coefficients d'absorption des parois, pour obtenir une réponse impulsionnelle simulée grâce â la méthode des sources images, qui soit au plus proche de la durée de réverbération cible. Pour démontrer que cette méthode est utile y compris lorsque les parois sont hétérogènes, on suppose également que les coefficients d'absorption ne sont plus identiques pour toutes les parois. Chaque paire de parois en vis à vis, a un coefficient d'absorption différent. On indice avec x (respectivement y et z) les grandeurs relatives aux parois normales à l'axe x (respectivement y et z). La seule condition imposée pour pouvoir résoudre ce problème, est le rapport entre un coefficient d'absorption global $\tilde{\alpha}$ et ceux des parois : $\alpha_x = w_x \times \tilde{\alpha}$, $\alpha_y = w_y \times \tilde{\alpha}$ et $\alpha_z = w_z \times \tilde{\alpha}$, avec (w_x, w_y, w_z) des coefficients prédéfinis. La grandeur recherchée est donc $\tilde{\alpha}$, le coefficient d'absorption global de la salle qui permet d'obtenir les coefficients d'absorption de chaque paroi.

Ce coefficient d'absorption global est recherché de deux manières différentes :

— en inversant les formules de Sabine et d'Eyring, provenant de l'acoustique statistique,
— en utilisant la méthode proposée par É. Lehmann, dans sa version originale [139, 164],
— en utilisant la méthode itérative inspirée des travaux de Lehmann, modifiée selon les éléments décrits dans les sections précédentes.

Ces deux dernières approches permettent d'optimiser itérativement $\tilde{\alpha}$. Pour l'étape d'initialisation, $\tilde{\alpha}$ utilise la valeur fournie par l'inversion de la formule d'Eyring. Un $T_{r,e}$ est alors estimé grâce à la méthode d'estimation rapide basée sur le calcul de $\tilde{h}_p$, et la méthode d'estimation rapide du T_r est alors réutilisée pour mettre à jour la valeur de $T_{r,e}$. Si $T_{r,e} > T_{r,c}$, alors $\tilde{\alpha}$ doit être augmenté, et inversement, si $T_{r,e} < T_{r,c}$, alors $\tilde{\alpha}$ doit être diminué. Cette étape est alors itérée N fois, permettant ainsi de déterminer par dichotomie la valeur de $\tilde{\alpha}$ pour atteindre $T_{r,e} = T_{r,c}$, avec une tolérance de 10^{-4} secondes. Il est important de noter que cette recherche itérative converge en quelques itérations seulement et est assez rapide, car l'estimation de la décroissance énergétique peut n'être faite qu'en quelques points, tout en restant valide [139].

Une fois le coefficient d'absorption global $\tilde{\alpha}$ obtenu à partir d'une de ces 4 méthodes, 10 réponses impulsionnelles de chaque salles sont calculées avec la méthode des sources images. Ces réponses impulsionnelles correspondent à 10 couples source-récepteur, où les sources sont positionnées aléatoirement à une distance de 1 à 2,5 m du récepteur dans la salle. Pour chacune des méthodes, on obtient donc 10 valeurs de $T_{r,eff}$ obtenues par la méthodes de sources images.

Afin d'estimer la cohérence entre le $T_{r,eff}$ obtenu par la méthode de sources images pour les 4 méthodes de détermination du paramètre d'entrée $\tilde{\alpha}$, on peut alors calculer une erreur quadratique moyenne ϵ sur le $T_{r,eff}$ par rapport au $T_{r,c}$:

$$\epsilon = \sqrt{\sum_i (T_{r,eff,i} - T_{r,c})^2}$$

Les erreurs quadratiques moyennes ϵ obtenues pour les 4 méthodes sont récapitulées dans le tableau 3.6, pour deux géométries de salles différentes :

	Méthode	Sabine	Eyring	Lehmann	Proposée
Salle 2 ($T_{r,c} = 0,5s$) (10m ; 7m ; 3,7m)	$\tilde{\alpha}$	0,382	0,328	0,568	0,608
	ϵ (s)	0,17 (34%)	0,24 (48%)	0,03 (6%)	0,017 (3,4%)
Salle 3 ($T_{r,c} = 0,6s$) (5m ; 4m ; 2,9m)	$\tilde{\alpha}$	0,213	0,196	0,265	0,287
	ϵ (s)	0,14 (23%)	0,19 (32%)	0,03 (5%)	0,012 (2%)

Tableau 3.6 – Récapitulatif des salles testées ainsi que leur Tr évalué par différentes méthodes dans le cas où l'on cherche à simuler un T_r particulier

L'analyse de ces résultats démontre que la méthode proposée dans le cadre de cette thèse est la plus précise, et permet une grande cohérence de résultats entre le T_r cible et le T_r effectif obtenu par méthodes de sources images, grâce à l'utilisation d'un coefficient $\tilde{alpha}$ plus adapté. Comme évoqué par Lehmann, ces résultats démontrent que l'approche communément utilisée par un grand nombre d'auteurs, de réaliser une simulation de réponse impulsionnelle en utilisant des coefficients $\tilde{\alpha}$ tirés de la théorie statistique, mène inévitablement à une simulation erronée du comportement de la salle.

En revanche, lorsqu'un soin particulier est apporté aux paramètres d'entrées de la simulation, avec une approche cohérente avec le modèle d'acoustique géométrique sous-jacent, la simulation numérique obtenue permet de modéliser beaucoup plus précisément le comportement de la salle. Les résultats de la section 3.5.6, sont donc confirmés dans le cas de pièces aux coefficients d'absorption différents selon les parois. Enfin, la méthode proposée pour estimer le coefficient d'absorption modifié à utiliser dans le cas de la simulation des sources images, permet d'obtenir une réponse impulsionnelle de salle, dont la décroissance énergétique ne diffère que de quelques pourcents de la décroissance cible.

3.6 Synthèse des apports principaux liés au calcul de jeux de données simulées

Ce chapitre a permis d'exposer les approches numériques et les optimisations apportées afin de générer des jeux de données par simulation, suffisamment réalistes pour éprouver les performances d'une approche de localisation de sources acoustiques par BeamLearning. La qualité de l'apprentissage étant fortement dépendante de celle des données présentées lors de l'optimisation des variables du réseau de neurones, un soin particulier a été porté au calcul de ces données.

La méthode choisie pour simuler la propagation des sources est la méthode des sources images. Cette méthode a été optimisée spécifiquement au cours de cette thèse pour le calcul parallèle sur processeur graphique, afin de simuler une grande quantité d'exemples réalistes en un temps minimal. Pour exemple, calculer la propagation de 8 000 sources vers 7 microphones dans une salle réverbérante impliquant l'utilisation de 80 000 sources images nécessite environ 30 minutes en exploitant la puissance de calcul d'un GPU NVidia 1080 Ti.

De plus, pour garantir une précision temporelle suffisante pour la localisation de sources sur antennes compactes, les retards introduits par la propagation des sources images peuvent être fractionnaires, grâce à l'approche de filtres basés sur l'interpolation de Lagrange, permettant, dans le domaine de fréquence visé, de simuler fidèlement des retards jusqu'au millième d'échantillon tout en restant économe en temps de calcul.

Enfin, une méthode originale, s'inspirant de celle proposée par É. Lehmann [139], est proposée pour estimer le coefficient d'absorption à utiliser dans le cadre de la simulation des sources images. En effet, l'utilisation d'un α obtenu à partir de l'inversion des formules de Sabine ou d'Eyring, entraîne une forte différence (30% environ dans les cas présentés) entre le T_r cible utilisé pour inverser la formule d'acoustique statistique et le T_r effectivement mesuré après le calcul des sources images. Au contraire, la méthode proposée aide au choix d'un coefficient d'absorption qui permet d'obtenir un T_r effectivement mesuré proche à 3 ou 5 % près d'un T_r cible. Grâce à cette méthode, il est envisageable de simuler la réponse d'une pièce particulière et donc d'optimiser l'apprentissage d'un réseau de neurone pour un environnement particulier.

Malgré le soin apporté aux simulations, les phénomènes tels que la diffraction du corps de l'antenne ou la réponse en fréquence propre de chaque microphone ne sont pas pris en compte. C'est pourquoi ces jeux de données simulées doivent être complétés par des jeux de données mesurées, qui intègrent par essence ces phénomènes.

Chapitre 4

Création de bases de données multicanales expérimentales grâce à la spatialisation 3D par synthèse ambisonique à ordres élevés

Tout ce qui est susceptible d'aller mal, ira mal (Loi de Murphy)

Contenu du chapitre

Les méthodes d'apprentissage supervisées nécessitent d'exploiter des jeux de données conséquents et réalistes. L'une des contributions originales de cette thèse repose sur l'utilisation d'un outil de spatialisation 3D expérimental, permettant de dépasser le stade de la validation numérique et de démontrer l'intérêt de l'approche de localisation par Deep Learning. Pour cela, les travaux exposés ici ont bénéficié d'un outil expérimental fonctionnel, développé précédemment au laboratoire : le spatialisateur par synthèse ambisonique 3D *SpherBedev* qui a été conçu pendant la thèse de doctorat de Pierre Lecomte au LMSSC [42]. Cet outil n'a pas nécessité de développement supplémentaire pendant ma thèse de doctorat, mais son exploitation a permis une approche originale et particulièrement efficace pour la constitution de jeux de données expérimentaux. En effet, la précision de synthèse du champ spatialisé à partir du formalisme ambisonique est très satisfaisant dans une sphère de rayon de 10 cm à 20 cm environ, ce qui correspond tout à fait au volume des antennes microphoniques compactes qui nous intéressent dans le cadre de cette thèse.

Cette section présente donc plus spécifiquement, outre le spatialisateur 3D, la manière dont les jeux de données expérimentales ont été constitués, depuis la géométrie des antennes, jusqu'au processus automatisé de synthèse et d'acquisition de dizaines de milliers de signaux issus de sources dont la position est étiquetée de manière automatique.

4.1 Le dispositif de synthèse ambisonique 3D au Cnam

4.1.1 La méthode ambisonique d'ordres élevés

La synthèse de champs 3D spatialisés représente un enjeu scientifique et industriel, et il existe de nombreux systèmes et formats *concurrents* de restitution, de complexités variées. Tandis que certaines approches ne sont basées que sur un élargissement de l'image sonore et des critères d'illusion perceptive, comme la simple stéréophonie, le son Surround ou le Vector Based Amplitude Panning (VBAP), d'autres approches plus complexes ont également fait l'objet de nombreuses recherches académiques, visant notamment à reproduire un champ sonore dans une zone spatiale élargie. Parmi elles, la synthèse de fronts d'onde ou WFS (Wave Field Synthesis), introduite initialement par Berkhout *et al.* [167], est une technique de reproduction de champs sonores, qui suppose une propagation en champ libre. Pour des raisons pratiques (taille du réseau, nombre de canaux), un système WFS est généralement

limité à la reproduction dans le plan horizontal, par un nombre fini de sources discrètes, en utilisant des simplifications appropriées de la formulation intégrale sous-jacente, difficilement transposables à un réseau de sources pour une synthèse tridimensionnelle [42]. Une méthode alternative concurrente plus adaptée à la synthèse tridimensionnelle est la méthode ambisonique d'ordres élevés. Cette technique de captation et de synthèse de champs acoustiques, reposant sur la description du champ sur une base d'harmoniques sphériques, a été introduite dans les années 70 [168], pour des ordres faibles. L'augmentation de cet ordre de reproduction a motivé de nombreuses recherches, et constitue le cadre théorique de l'ambisonie à ordres élevés (HOA) [169–171]. Les techniques de HOA ont ainsi permis d'élargir la zone de reconstruction valide physiquement, au prix d'un nombre croissant de sources et de canaux de pilotage.

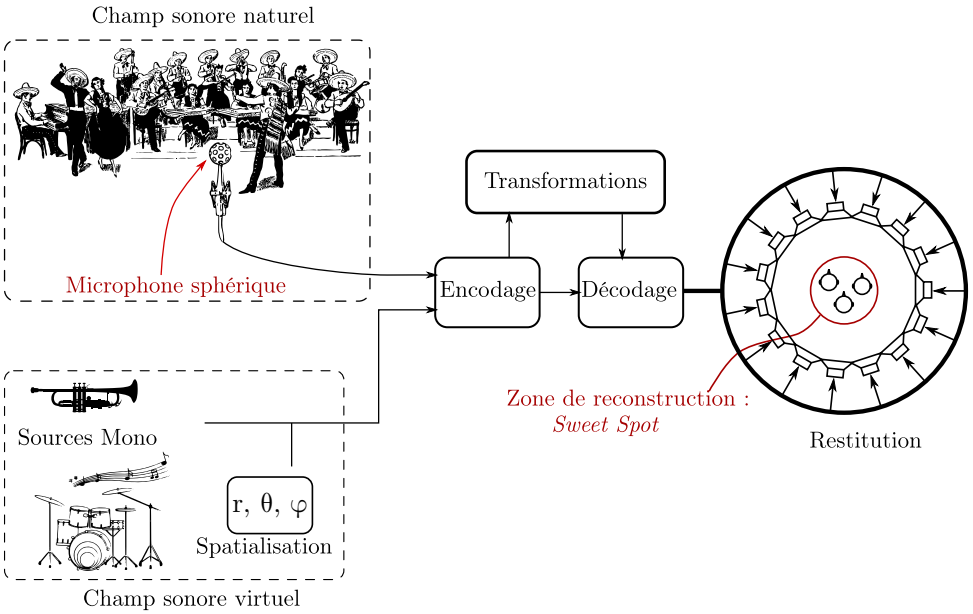


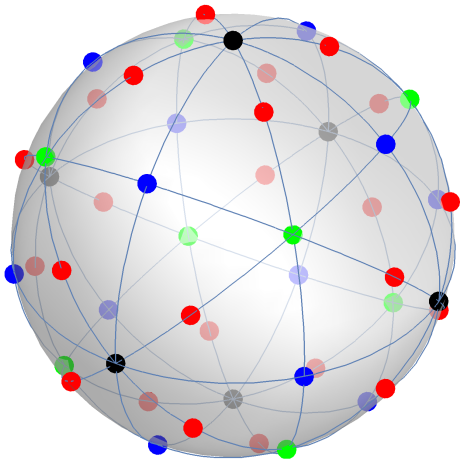
FIGURE 4.1 – Principe général de la méthode ambisonique. D'après la thèse de Pierre Lecomte [42]

Sur cette base, Pierre Lecomte a conçu au cours de sa thèse de doctorat au LMSSC un spatialisateur 3D et une suite logicielle de synthèse de champs en temps réel, offrant ainsi un outil parfaitement adapté à nos besoins de conception de bases de données expérimentales. Ces éléments étant parfaitement fonctionnels pendant le déroulement de ma thèse de doctorat, ils n'ont donc pas fait l'objet de développements spécifiques, mais il apparaît important de décrire succintement les outils utilisés et leurs bases physiques, ainsi que leur domaine de validité.

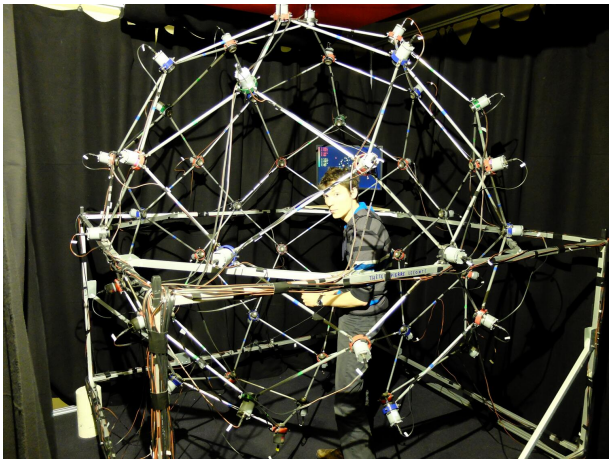
Comme illustré sur la figure 4.1, les techniques d'ambisonie à ordres élevés concernent essentiellement 3 grands domaines, qui ont tous étés traités au cours de la thèse de Pierre Lecomte : la captation, la transformation et la restitution de champs naturels ou la synthèse virtuelle de champs sonores. Dans tous les cas, les méthodes reposent sur l'encodage d'un champ de pression sur une base d'harmoniques sphériques. Dès qu'il s'agit de restituer ces champs encodés dans le domaine ambisonique, il est ensuite nécessaire de réaliser le décodage ambisonique, qui consiste à calculer les signaux d'entrée individuels de chaque haut-parleur du réseau utilisé par le spatialisateur.

La fidélité de la restitution dépend à la fois de la fréquence et de la zone de l'espace. Dans la gamme de fréquences utilisées pour les expériences, la zone de reconstruction (sweet spot) est une petite zone d'une dizaine de cm de rayon au centre de la sphère de haut-parleurs. Les caractéristiques précises du champ acoustique ainsi créé sont développées plus loin (4.1.7).

La spatialisateur 3D conçu au laboratoire est une sphère de haut-parleurs suivant un maillage de Lebedev à cinquante points, qui permet une reconstruction du champ sonore jusqu'à l'ordre 5 avec le formalisme ambisonique. La figure 4.2 présente ce maillage, ainsi que sa mise en place pour le spatialisateur 3D.



(a) Maillage de Lebedev à 50 points



(b) Spatialisateur 3D SpherBedev

FIGURE 4.2 – Maillage de Lebedev à 50 points et sa mise en œuvre pour le spatialisateur 3D

4.1.2 Définition des harmoniques sphériques

Soit l'espace Hilbertien $\mathbb{L}^2$. Cet espace est dans ce document l'espace des fonctions de carrés intégrables sur la sphère : $\Omega = (x, y, z) \in \mathbb{R}^3 | x^2 + y^2 + z^2 = 1$. En reprenant les notations définies précédemment pour les coordonnées sphériques, on peut définir le produit scalaire entre deux fonctions f et g de $\mathbb{L}^2$:

$$\langle f, g \rangle = \frac{1}{4\pi} \int_{\theta=0}^{2\pi} \int_{\phi=-\frac{\pi}{2}}^{+\frac{\pi}{2}} f(\theta, \phi) g(\theta, \phi) cos(\phi) d\phi d\theta$$

La famille d'harmoniques sphérique, dénotée $Y_{m,n}$ forme une base orthonormée de $\mathbb{L}^2$ [129, 172]. Pour pouvoir définir les harmoniques sphériques, les polynômes de Legendre doivent d'abord être présentés. On définit par récurrence le polynôme de Legendre de première espèce $P_m(x)$:

$$\begin{cases} P_0(x) & = 1 \\ P_1(x) & = x \\ (m+1)P_m(x) & = (2m+1)xP_m(x) - mP_{m-1}(x) \ ; \ m > 1 \end{cases} \tag{4.1}$$

On définit ensuite les polynômes de Legendre associés d'ordre m et de degré $|n|$, $(m, n) \in (\mathbb{R}, \mathbb{Z})$. Ces polynômes sont définis sur $[-1, 1]$ par :

$$P_{m,|n|}(x) = (1 - x^2)^{\frac{|n|}{2}} \frac{d^{|n|}}{dx^{|n|}} P_m(x)$$

Les harmoniques sphériques sont alors définis par [42, 173] [1] :

$$Y_{m,|n|}(\theta, \phi) = \sqrt{(2m+1)\epsilon_n \frac{(m-|n|)!}{(m+|n|)!}} P_{m,|n|}(\sin(\phi)) \cdot \begin{cases} cos(|n|\theta) \text{ si } n \geq 0 \\ sin(|n|\theta) \text{ si } n < 0 \end{cases} \tag{4.2}$$

avec $\epsilon_n = 1$ si $n = 0$ et $\epsilon_n = 2$ si $n > 0$. Dans cette équation, m désigne l'ordre de l'harmonique sphérique, et n son degré. Les premières harmoniques de la famille sont représentés en figure 4.3.

1. Il existe plusieurs définitions possibles des harmoniques sphériques. Pour ce document, les harmoniques sphériques à valeurs réelles sont utilisés.

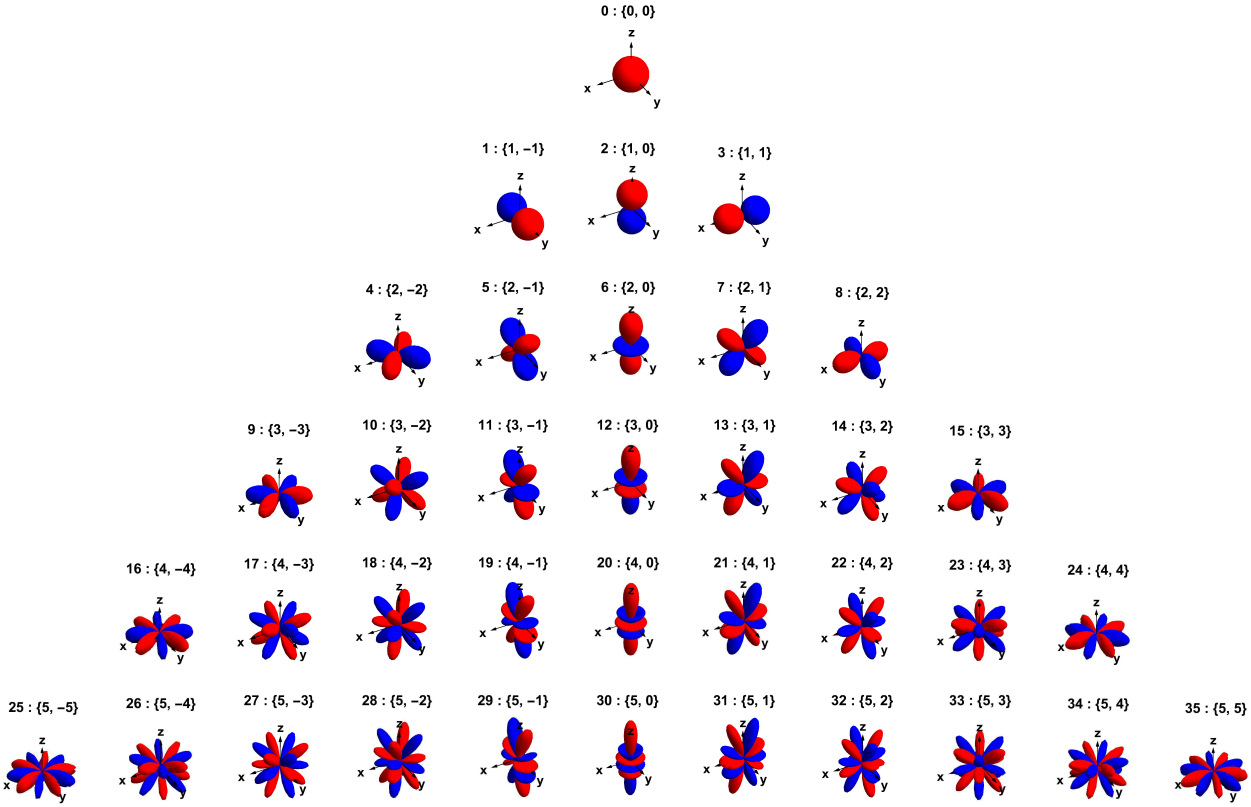


FIGURE 4.3 – Illustration des harmoniques sphériques jusqu'à l'ordre m = 5. Les couleurs rouge et bleue indiquent respectivement des valeurs positives et négatives. Le numéro ACN (*Ambisonic Channel Number*) [173] et les indices (m, n) correspondants sont notés au dessus de chaque figure. Figure tirée de [42]

4.1.3 Décomposition d'un champ de pression acoustique sur les harmoniques sphériques

Il est possible de décrire un champ de pression acoustique p en le projetant sur la base des harmoniques sphériques. Seules les principales équations seront présentées ici. On rappelle que l'équation de Helmholtz homogène sans terme source s'écrit :

$$\Delta p + k^2 p = 0 \tag{4.3}$$

Dans cette équation, Δp représente le laplacien de la pression acoustique, $k = \frac{\omega}{c}$ le nombre d'onde et c est la vitesse du son dans l'air. Les solutions du problème intérieur (source à l'extérieur d'une sphère) de l'équation d'Helmholtz homogène dans le système de coordonnées sphériques peuvent s'écrire comme

une série de Fourier-Bessel [174] :

$$p(kr, \theta, \phi) = \sum_{m=0}^{\infty} i^m j_m(kr) \sum_{n=-m}^{m} B_{mn} Y_{mn}(\theta, \phi) \tag{4.4}$$

Dans cette équation, $j_m(kr)$ représentent les fonctions de Bessel sphériques [175]. Les coefficients B_{mn} sont dénommés les composantes ambisoniques. Physiquement, ils correspondent aux dérivées spatiales successives de la pression acoustique calculées à l'origine du repère [176]. Ces coefficients peuvent être déterminés analytiquement dans des cas de modèles acoustiques simples, comme l'onde plane et l'onde sphérique. Dans les deux cas, la solution se trouve en remplaçant la pression par son expression, et en identifiant les termes avec la série de Fourier-Bessel de l'équation de Helmholtz (4.3). Dans le cas de l'onde plane monochromatique d'amplitude S et de direction de propagation (θ_p, ϕ_p), les composantes de la décomposition en harmonique sphérique sont :

$$B_{mn} = SY_{mn}(\theta_p, \phi_p) \tag{4.5}$$

Dans le cas de l'onde sphérique émise par une source monopolaire à la position (r_s, θ_s, ϕ_s), les composantes de la décomposition en harmonique sphérique sont [42] :

$$B_{mn} = SF_m(kr_s)Y_{mn}(\theta_s, \phi_s) \tag{4.6}$$

avec $F_m(kr_s) = i^{-(m+1)} k h_m^{(2)} \frac{(kr_s)}{4\pi}$. Cette fonction correspond à des filtres de champ proche qui modélisent la distance finie de la source à l'origine [40, 176].

4.1.4 Captation d'un champ de pression grâce au domaine ambisonique

L'objectif de cette section n'est pas de donner les éléments de calculs pour pouvoir déterminer un encodage ambisonique du champ de pression, mais seulement de présenter la méthode. Pour pouvoir encoder un champ de pression dans le domaine des harmoniques sphériques, il faut pouvoir extraire les composantes ambisonique du champ.

L'une des méthodes communément exploitée à partir de mesures du champs sur un antenne sphérique consiste à calculer la transformée de Fourier sphérique, qui permet de projeter une fonction angulaire f de $\mathbb{L}^2$ sur la base des harmoniques sphériques [177]. Ainsi, pour connaître les composantes

ambisoniques représentant un champ de pression, il suffit de connaître le champ de pression à la surface d'une sphère, ce qui est rendu possible par l'échantillonnage de ce champ par un ensemble de microphones [42, 171]. Le calcul de cette transformée de Fourier sphérique varie suivant que la sphère à la surface de laquelle la pression est mesurée est rigide ou modélisée comme une simple couche [42]. Dans tout le reste du document, on considérera comme acquis les principes de captation par une sphère rigide du champ de pression, et de sa décomposition en harmoniques sphériques.

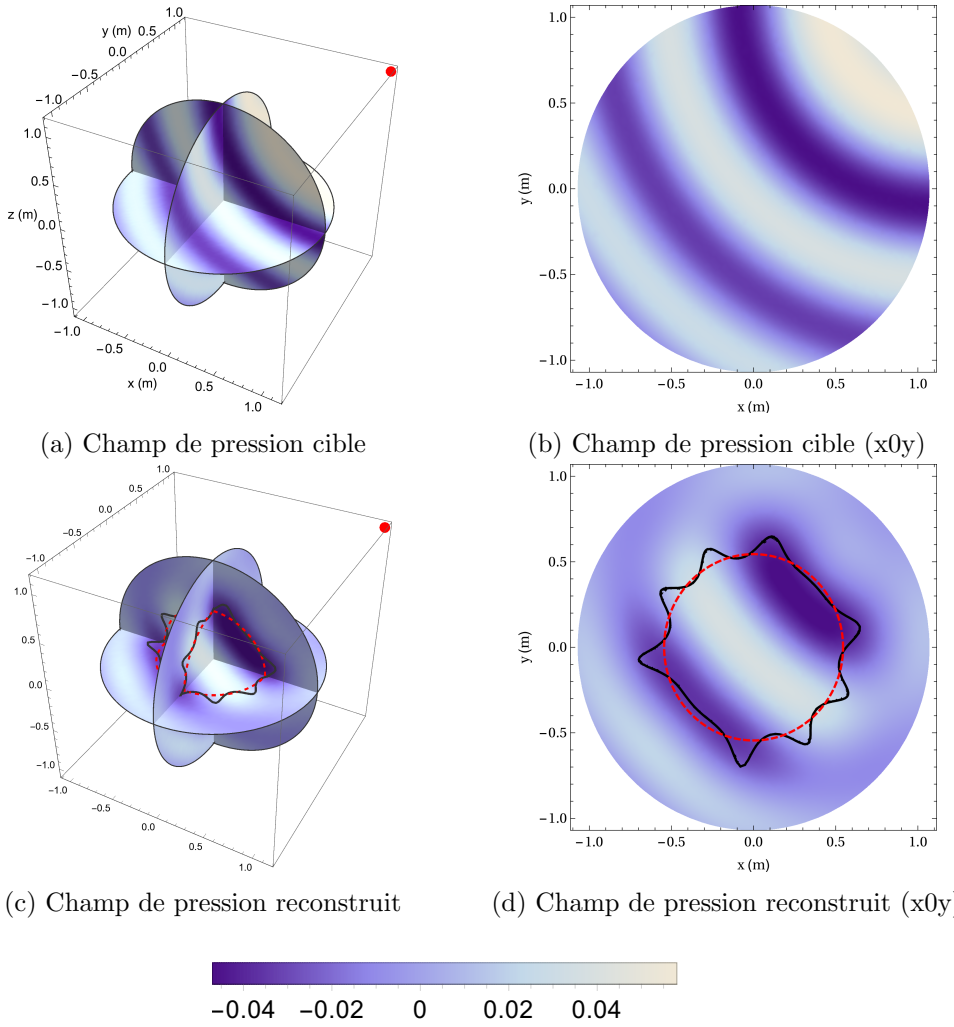
Une fois le champ projeté sur la base des harmoniques sphériques, (phase d'encodage ambisonique) il est possible de lui faire subir des transformations qui se prêtent bien au formalisme des fonctions de $\mathbb{L}^2$. En particulier les rotations ainsi que les symétries planes ou axiales peuvent s'obtenir facilement en inversant certains ordres des composantes ambisoniques [42].

4.1.5 Troncature

Comme exposé dans la sous-section précédente, pour estimer la valeur du champ de pression à la surface de la sphère, celui-ci est mesuré en un nombre fini de points grâce à des microphones. La discrétisation entraîne alors une fréquence de repliement, au delà de laquelle l'estimation des composantes ambisoniques n'est plus possible. Cette fréquence est approximativement $f_{alias} = \frac{Mc}{2\pi r_{mic}}$, où M est l'ordre maximal pour lequel la transformée de Fourier sphérique est identique à la transformée de Fourier sphérique discrète [178]. Ainsi, le champ de pression n'est représenté qu'à partir d'un nombre fini d'harmoniques. Cette troncature à l'ordre M entraîne une représentation du champ de pression grâce à $(M+1)^2$ composantes ambisoniques. Cette troncature entraîne nécessairement une erreur d'estimation du champ. On peut en particulier estimer ϵ_M, l'erreur de troncature à l'ordre M, normalisée et moyennée sur les angles (θ, ϕ), pour le cas de l'onde plane et de des ondes sphériques [179, 180]. Une règle générale d'approximation est établie [42] : pour une erreur ϵ_M de 4% (-14 dB), la zone de reconstruction est donnée par une sphère de rayon $r_{4\%}$:

$$\boxed{r_{4\%} = \frac{M}{k}} \tag{4.7}$$

Ainsi $r_{4\%}$ est proportionnelle à l'ordre de troncature, et inversement proportionnelle à la fréquence.



(a) Champ de pression cible

(b) Champ de pression cible (x0y)

(c) Champ de pression reconstruit

(d) Champ de pression reconstruit (x0y)

FIGURE 4.4 – Partie réelle du champ de pression émis par un monopole (en rouge). Cas du champ complet (4.4(a) et 4.4(b)) et tronqué à l'ordre $M = 5$ (4.4(c) et 4.4(d)). Amplitude $S = 1$, fréquence $f = 500Hz$. Le contour noir indique l'erreur quadratique de reconstruction $\epsilon = 0,04$. Le cercle rouge pointillé est de rayon $r = M/k$. Figure tirée de [42]

Cette zone de reconstruction est illustrée à la figure 4.4 tirée de la thèse de doctorat de Pierre Lecomte [42]. Un champ de pression cible (4.4(a) et 4.4(b)) émis par un monopole est tronqué dans le domaine ambisonique à l'ordre $M = 5$ (4.4(c) et 4.4(d)). À partir de ce calcul, la zone de reconstruction du champ telle que $\epsilon_5 = 0,04$ est délimitée par un trait plein noir. On peut alors observer que la sphère de rayon $r_{4\%} = \frac{M}{k}$ délimitée par les pointillés rouges est une bonne approximation de la zone de reconstruction valide, qu'on appellera par la suite le *sweet spot*.

4.1.6 Restitution d'un champ de pression grâce au domaine ambisonique

Une fois le champ de pression cible encodé dans le domaine ambisonique après captation ou simulation, il est possible de le restituer grâce à des haut-parleurs : c'est l'opération de décodage ambisonique. Dans le formalisme ambisonique, la géométrie sphérique est la plus naturelle pour répartir les haut-parleurs autour de la zone de reconstruction. Différentes méthodes de restitutions existent , parmi lesquelles on peut citer la méthode du mode-matching [171, 176] et la méthode simple source [175]. Si la règle de quadrature ayant servi à la disposition des hauts-parleurs sur la surface de la sphère permet de conserver le critère d'orthonormalité des harmoniques sphériques après discrétisation, ces deux formulations sont équivalentes. Or, la quadrature de Lebedev exploitée par Pierre Lecomte pour concevoir le sphère de spatialisation du laboratoire respecte justement ce critère. Pour cette raison, seule la méthode du *mode-matching* sera présentée ici.

On suppose que L haut-parleurs échantillonnent la sphère de spatialisation. Chaque haut-parleur est modélisé comme un monopôle associé à une amplitude de signal $s_l \in [s_1, s_2..., s_L]$. La contribution de chaque source peut être exprimée au centre de la sphère dans le domaine ambisonique. La somme de toutes les contributions élémentaires des haut-parleurs doit être égale au champ cible. Formellement, cette égalité se traduit dans le domaine ambisonique tronqué à l'ordre M par :

$$\sum_{l=1}^{L} s_l \sum_{m=0}^{M} i^m j_m(kr) F_m(kr_l) \sum_{n=-m}^{m} Y_{mn}(\theta_l, \phi_l) Y_{mn}(\theta, \phi) = \sum_{m=0}^{M} i^m j_m(kr) \sum_{n=-m}^{m} B_{mn} Y_{mn}(\theta, \phi) \qquad (4.8)$$

Où les F_m représentent les filtres de champs proches, introduits par Daniel dans le cadre de l'ambisonie [176]. Ces filtres modélisent la distance finie de la source à l'origine du repère.

Les harmoniques sphériques étant orthonormées, chacun des termes (m, n) de cette somme peut être identifié séparément (d'où le nom *mode-matching*). Pour trouver les composantes s_l, le théorème d'additivité [172], permettant de passer d'un produit scalaire entre deux ondes décomposées sur une

base ambisonique à un polynôme de Legendre d'ordre m, est nécessaire :

$$\begin{cases} \sum_{n=-m}^{m} Y_{mn}(\theta_1, \phi_1) Y_{mn}(\theta_2, \phi_2) = P_m(\gamma_l) \\ \text{avec : } \gamma_l = \cos(\phi_1)\cos(\phi_2)\cos(\theta_1 - \theta_2) + \sin(\phi_1)\sin(\phi_2) \end{cases} \tag{4.9}$$

Grâce à cette formulation, il est possible d'obtenir l'amplitude du signal de chaque haut-parleur s_l dans le cas d'ondes planes à partir de γ_l, qui représente le produit scalaire entre le vecteur normé pointant dans la direction (θ_1, ϕ_1) de la source virtuelle, et le vecteur normé pointant dans la direction (θ_2, ϕ_2) du haut parleur l :

$$s_l = Sw_l \sum_{m=0}^{M} \frac{(2m+1)}{F_m(k, r_{spk})} P_m(\gamma_l) \tag{4.10}$$

De même, on obtient l'amplitude de chaque haut-parleur dans le cas d'une onde sphérique :

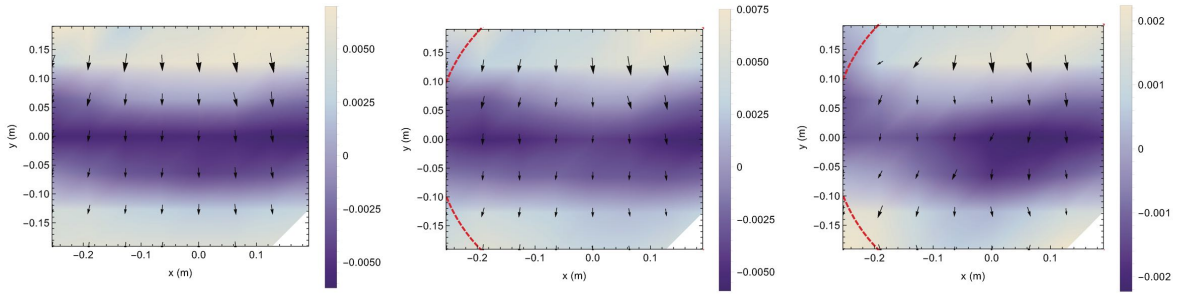
$$s_l = Sw_l \sum_{m=0}^{M} (2m+1) \frac{F_m(k, r_s)}{F_m(k, r_{spk})} P_m(\gamma_l) \tag{4.11}$$

En revanche, pour la restitution d'un enregistrement ambisonique, la solution implique de calculer un filtre radial permettant de compenser à la fois la diffraction par la sphère de captation et l'effet de champ proche des hauts-parleurs. Cette solution peut nécessiter d'y adjoindre une approche permettant de stabiliser les filtres obtenus [181].

4.1.7 Résultats obtenus grâce au spatialisateur *SpherBedev*

Cette section présente des résultats expérimentaux obtenus par Pierre Lecomte au cours de sa thèse grâce au spatialisateur 3D, dans le cas d'une source ponctuelle. L'objectif est ici d'illustrer le fait que la synthèse de champs réalisée à partir du spatialisateur 3D du laboratoire permet de reconstruire efficacement un champ cible. Pour cela, des mesures de caractérisation de champs de pression et

d'intensité sont réalisées à l'aide d'une antenne plane d'envergure de $32cm \times 45cm$, composée de 55 microphones, placée à l'équateur de la sphère de spatialisation. Le champ de référence, présenté à la figure 4.5(a), correspond à la mesure par cette antenne plane du champ acoustique (à 1 000 Hz) émis par un haut-parleur situé à 1,07 m du centre de la sphère de spatialisation. La figure 4.5(b) présente, quant à elle, le champ capté sur l'antenne plane, lorsque la sphère de spatialisation est utilisée pour simuler le champ de pression émis par cette source, par synthèse ambisonique d'ordre 5 (voir équation 4.11). Pour finir, la figure 4.5(c) présente la mesure du champ acoustique sur l'antenne plane, lorsque l'émission par le haut-parleur est mesuré par une antenne microphonique ambisonique (voir section 4.1.4) pour encoder le champ, puis décodé pour être restitué par la sphère de spatialisation 3D.



(a) Monopole (b) Synthèse (c) Restitution

FIGURE 4.5 – Partie réelle du champ de pression émis par un monopole capté sur une antenne plane. Cas d'un haut-parleur émettant un signal sinusoïdal à 1 000 Hz (4.5(a)). Cas d'un monopole simulé par le spatialisateur à l'ordre $M = 5$ (4.5(b)). Cas d'un haut-parleur capté dans le domaine ambisonique et restitué par le spatialisateur à l'ordre $M = 5$ (4.5(c)). Le cercle rouge pointillé est de rayon $r = 5/k$. Figure tirée de [42]

La comparaison de ces 3 champs mesurés permet de montrer que l'utilisation du spatialisateur 3D pour restituer les champs simulés par synthèse ambisonique (figure 4.5(b)) ainsi que les champs restitués après captation dans le domaine ambisonique (figure 4.5(c)) présentent une excellente concordance avec le champ de pression cible (figure 4.5(a)) dans la zone du *sweet spot*, que ce soit d'un point de vue de la pression ou du champ de vecteur d'intensité. En revanche, aucune précaution particulière n'ayant été prise pour ajuster le gain, les échelles de pression varient d'une captation à l'autre. Mais cela ne contredit en rien la validité de la reconstruction du champ de pression, qui est reconstruit à un facteur proportionnel d'amplitude près.

4.2 Constitution de la base de données mesurées

4.2.1 Contrainte de compacité des antennes microphoniques

Comme il a été présenté sur la figure 4.5, le système de restitution *SpherBedev* [42] disponible au laboratoire d'acoustique du Cnam, peut être utilisé de deux manières :

— pour réaliser la synthèse physique d'un champ de pression encodé préalablement grâce à un microphone ambisonique d'ordre 5 dédié *Memsbedev* [42],
— pour synthétiser un champ sonore parfaitement maîtrisé issu d'une source virtuelle.

Dans les deux cas, le domaine ambisonique permet de contrôler parfaitement le champ de pression. De plus, il est possible de lui faire subir des transformations (rotations, symétries...) comme exposé en section 4.1.4. Enfin, par linéarité du champ de pression, de nouvelles scènes sonores peuvent être créées en superposant plusieurs champs dans le domaine ambisonique, réalisé comme un *mixage* simple de champs tridimensionnels. L'outil de spatialisation 3D offre ainsi un cadre idéal pour constituer des bases de données expérimentales, avec une variabilité potentiellement infinie de situations synthétisées.

Comme exposé précédemment, le dispositif présente une limite en hautes fréquences, où la zone de reconstitution précise des champs présente une extension spatiale de plus en plus réduite lorsque l'on monte en fréquence (voir équation 4.7). Ainsi, pour réaliser un apprentissage sur une antenne physique au sein du spatialisateur, il est nécessaire de la disposer au centre de la sphère de hauts-parleurs, et que l'extension physique de l'antenne reste contenue dans la zone du *sweet spot*. En se fixant une validité fréquentielle jusqu'à 4 000 Hz, cela correspond à un *sweet spot* possédant un rayon caractéristique de 7 cm environ, ce qui implique que les antennes microphoniques auxquelles on souhaite adjoindre une intelligence artificielle sont toutes des antennes dites *compactes*.

4.2.2 Géométries d'antennes microphoniques pour les bases de données expérimentales

L'approche proposée de localisation de sources par BeamLearning, est conçue pour que le réseau de neurones associé puisse construire les représentations nécessaires à partir des données brutes, indépendamment de la topologie de l'antenne microphonique, ainsi que du nombre de capteurs la composant.

Ainsi, hormis les contraintes de compacité des antennes, liées à la technique de synthèse ambisonique utilisée pour exposer les antennes à des champs variés et contrôlés, il est possible d'utiliser des géométries variées. C'est la raison pour laquelle nous avons choisi d'utiliser plusieurs topologies d'antennes microphoniques, afin d'analyser leur comportement et leurs performances d'apprentissage. Cette section présente brièvement leurs caractéristiques, qui sont récapitulées dans le tableau 4.1. Pour chacune de ces antennes, des données techniques plus complètes sont disponibles en annexe de ce document.

Nom	Géométrie	Nombre de micros	Répartition des micros	Taille carac.
Mini DSP	Circulaire	7	6 microphones répartis sur le périmètre de l'antenne, et 1 au centre	Rayon = 4,3 cm
CMA Cube	Tétraédrique	7	4 microphones positionnés sur les sommets et 3 au centre des arrêtes supérieures	Coté = 10 cm
Zylia	Sphérique	19	19 microphones répartis sur la surface de la sphère	Rayon = 4,9 cm

Tableau 4.1 – Résumé des antennes utilisées pour les expériences

Par ailleurs, puisque les jeux de données simulés ont également été constitués pour plusieurs topologies d'antennes microphoniques, 2 configurations expérimentales parmi les 3 récapitulées dans le tableau 4.1 présentent la même géométrie pour les expériences et les simulations (antenne Mini DSP et antenne CMA Cube). Des comparaisons pourront donc être menées entre résultats expérimentaux et résultats numériques, en s'affranchissant de l'influence de la géométrie des antennes. [2]

2. Contrairement à ce qui était initialement prévu, la géométrie d'antenne CMA Cube n'a pas pu être testée dans le spatialisateur 3D, à cause d'une combinaison d'imprévus entre mi-Janvier et Août 2020 (défaut matériel, travaux de rénovation du laboratoire, COVID-19)

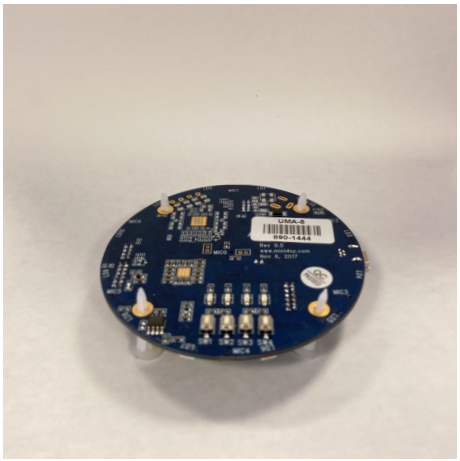


FIGURE 4.6 – Photo de l'antenne circulaire du constructeur Mini DSP

La première antenne qui a été utilisée dans la sphère de spatialisation est l'antenne circulaire du constructeur Mini DSP (voir annexe D). Elle est constituée de 7 microphones. Six microphones sont sur le périmètre de l'antenne et un est au centre. Son rayon fait 4,3 cm. De plus, sa géométrie relativement simple a permis de l'utiliser aussi lors des simulations. Elle est appelée antenne *mini DSP* dans tout le document.

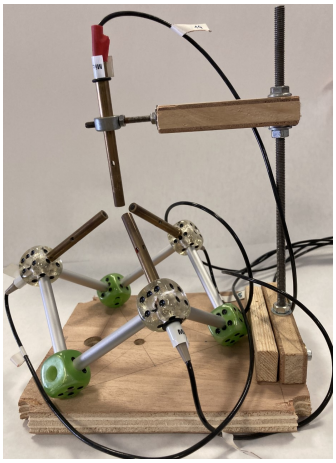


FIGURE 4.7 – Photo de l'antenne tétraédrique CMA Cube développée avec des microphones double couche lors de la thèse d'Aro Ramamonjy [14]

La deuxième antenne, prévue pour être utilisée dans la sphère de spatialisation, est une antenne en trois dimensions, possédant également 7 microphones. Son utilisation comparée à l'antenne MiniDSP

avait pour objectif d'évaluer l'influence de l'élévation des capteurs et de la diffraction par la structure de l'antenne sur les performances de localisation angulaire à 3 dimensions. La structure globale est un tétraèdre régulier de 3,2 cm de coté. Quatre microphones sont situés aux sommets du tétraèdre, et les trois derniers sont au milieu des arêtes supérieures. L'antenne *CMA Cube* a été développée avec des microphones double couche lors de la thèse d'Aro Ramamonjy [14]. Cette géométrie d'antenne a également été étudiée pour l'analyse de performance du réseau proposé, à partir de bases de données simulées numériquement. Les positions exactes des capsules microphoniques sont décrites en annexe E.

FIGURE 4.8 – Photo de l'antenne sphérique Zylia ZM-1

La troisième antenne exploitée pour la constitution de bases de données expérimentales est une antenne sphérique de microphones, conçue et commercialisée par la société Zylia (voir annexe F). Cette antenne est initialement conçue pour permettre un encodage ambisonique d'ordre 3 de champs acoustiques, avec des applications visant essentiellement la prise de son spatialisée et la navigation dans un champ acoustique pour la réalité virtuelle. L'antenne Zylia ZM-1, notée Zylia dans la suite du document, est une sphère rigide dotée à sa surface de 19 microphones MEMS numériques de dernière génération . Même si cette antenne est conçue pour obtenir une décomposition ambisonique du champ, les données d'entrées exploitées sont ici les données temporelles brutes de chaque canal microphonique, tout comme pour les autres antennes. D'autres travaux récents ont proposé l'utilisation de jeux de données ambisoniques en entrée de réseaux de neurones profonds pour traiter le problème de la localisation [63, 79], ou pour le problème de comptage de locuteurs en environnement réverbérant [182]. Cependant, l'approche BeamLearning se voulant totalement indépendante de la géométrie

de l'antenne, nous avons fait le choix de ne pas exploiter cette représentation afin de démontrer que notre approche de *joint feature learning* à partir de données brutes est pertinente. Le rayon de l'antenne Zylia est de 4,9 cm. Puisque les microphones sont disposés à la surface d'une sphère rigide, la diffraction par la structure de l'antenne est ici prédominante, et même s'il existe des codes de calcul de réponses impulsionnelles de salles prenant en compte ce paramètre [183], nous avons fait le choix ici de n'étudier ce type d'antenne que pour la partie expérimentale de ce travail [3] .

L'intérêt de la constitution d'une base de données expérimentale réside dans le fait qu'ici, les réponses en fréquences des microphones composant les antennes ne sont plus idéales, et, même avec des microphones appairés, il existe une faible disparité de réponses fréquentielles en amplitude et en phase, contrairement à l'hypothèse utilisée pour la constitution de bases de données simulées. Par ailleurs, la diffraction induite par la structure des antennes et des trépieds utilisés est nécessairement inclue dans les champs de pressions mesurés par les microphones. Pour des méthodes basées sur une approche *modèles*, il est particulièrement difficile de prendre en compte ce type de paramètre, alors que pour des méthodes basées sur une approche *données* telle que celle proposée ici, l'avantage réside dans le fait que le réseau apprend à exploiter également ces paramètres au cours de l'apprentissage. Ainsi, même lorsque les microphones d'une antenne ne sont pas étalonnés, la phase d'entraînement du réseau permet de réaliser un étalonnage intrinsèque des microphones [144], pour compenser ces réponses et affiner les estimations de localisation. Une discussion plus complète sur ce point sera menée dans le chapitre suivant (chap. 5).

4.2.3 Environnements acoustiques du spatialisateur 3D pour les jeux de données expérimentaux

Contrairement à ce qui a été développé pour les bases de données simulées numériquement, le spatialisateur 3D n'a pas été exploité ici pour simuler des environnements acoustiques différents de celui dans lequel la SpherBedev est installée au laboratoire, même si cela est possible techniquement [184]. L'adaptation des travaux initiés par Pierre Lecomte sur la compensation de réponse de salle à la simulation d'environnements variés par ambisonie nécessite des développements spécifiques complémentaires, qui sont prévus à l'issue de ma thèse de doctorat, mais n'ont pas été abordés dans le

3. la librairie SMIRGen possède en effet des temps de calcul excessifs par rapport aux besoins en termes de volumes de données à générer

cadre de ce travail. Ainsi, pour l'entraînement et le test du comportement du réseau de neurones basé sur l'approche BeamLearning, les environnements acoustiques sont ceux de la salle dans laquelle est installée le spatialisateur 3D du laboratoire. Cette salle est une salle traitée acoustiquement et désolidarisée du reste du bâtiment (boite dans la boite), sans pour autant être une salle anéchoïque. Pour la phase d'entraînement, le plafond n'était pas traité (dalle béton brute), le sol était recouvert d'une moquette épaisse, et les 4 parois latérales étaient masquées par des dièdres absorbants et des rideaux. Les dimensions de la salle sont : 4,3 m de longueur, 3,85 m de largeur et 3,02 m de hauteur sous plafond.

Dans cette configuration, une mesure de durée de réverbération T_r par bandes d'octaves a été réalisée par la méthode de la source interrompue, afin d'estimer les coefficients d'absorption moyens des murs, comme présenté en section 3.5. On trouve un T_r moyen de 0,1 s (les valeurs par bande d'octave sont données en tableau 4.2) et un α_{moy} de 0,8 s. La salle peut donc être considérée comme très absorbante, mais du fait que le plafond n'était pas traité acoustiquement, une réflexion marquée a donc lieu sur cette paroi, et la situation se rapproche d'une environnement semi-anéchoïque (inversé, puisque c'est habituellement le sol qui est non traité dans les salles semi-anéchoïques).

À l'intérieur de la salle, un bureau et la station informatique de pilotage de la sphère étaient présents à proximité de la sphère de spatialisation.

De manière à tester la robustesse de la méthode lorsque l'environnement acoustique est modifié et vérifier que notre approche ne réalise pas un sur-apprentissage de l'environnement d'entraînement, le traitement acoustique de la salle a été modifié pour la phase de test du réseau gelé après entraînement : tous les matériaux placés le long des parois latérales ont été déplacés, 4 structures de type *bass-trap* ont été ajoutées, et un traitement acoustique a été placé au plafond, avec un plénum de 10 centimètres. Par ailleurs, le mobilier et la station de pilotage du dispositif de spatialisation ont été déplacés dans la salle. Dans cette situation, la réflexion très marquée au plafond a donc été fortement atténuée, et l'environnement de mesure a été modifié, tout en restant dans une situation de salle « sourde », avec une durée de réverbération faible dans tout le domaine de fréquence visé.

Fréquence centrale de bande d'octave	T_r (s) avant travaux	T_r (s) après travaux	α_{moy} avant travaux	α_{moy} après travaux
125	0,13	0,16	0,52	0,46
250	0,18	0,12	0,41	0,55
500	0,12	0,8	0,55	0,7
1000	0,12	0,8	0,55	0,7
2000	0,09	0,8	0,66	0,7
4000	0,08	0,8	0,7	0,7

Tableau 4.2 – Résumé des durées de réverbération et coefficients d'absorption moyens de la salle par bande d'octave avant (configuration avec plafond en béton brut) et après avoir rajouté un matériaux absorbant au plafond

4.2.4 Signaux et synthèse de sources virtuelles par la sphère de spatialisation

D'un point de vue expérimental, compte tenu des travaux de démolition et de réhabilitation des salles de cours à l'étage au dessus des équipements du laboratoire depuis Février 2020, les acquisitions de bases de données expérimentales ont été conçues pour être réalisées pendant 12h consécutives, de nuit. Par conséquent, les séquences d'acquisition de jeux de données expérimentaux ont été conçues pour s'exécuter de manière automatique et continue, sans intervention humaine après avoir été lancées.

Le pilotage des paramètres de synthèse du spatialisateur SpherBedev est réalisé à l'aide d'un programme codé dans le langage Pure Data, qui permet de sélectionner séquentiellement des positions de sources virtuelles pour lesquelles le champ acoustique va être synthétisé par méthode ambisonique d'ordres élevés, ainsi que de sélectionner le type de signal émis par cette source virtuelle, parmi ceux présentés en section 3.2.4. Les instructions séquentielles sur les paramètres des sources virtuelles dont le champ est synthétisé correspondent à des instructions *timecodées*, pilotant par un message basé sur le protocole OSC [185] les coordonnées de la source (R, θ, ϕ), la durée d'émission du signal par cette source, et les temps de *pause* entre deux synthèses de champs successives. Ces messages OSC sont reçus en temps réel par la suite logicielle Ambitools développée en language Faust pour le pilotage du

spatialisateur *SpherBedev* [42, 44]. [4]

Pour ce pilotage des positions de sources virtuelles synthétisées par le spatialisateur, nous exploitons exactement le même principe de tirage aléatoire de positions que celui décrit en section 3.2.3 pour les bases de données simulées. Pour chaque source virtuelle synthétisée par ambisonie d'ordres élevés, les signaux correspondent à une séquence d'une durée de 1 seconde, avec une enveloppe trapézoïdale en amplitude, illustrée à la figure 4.9. Sur cette seconde de signal émis, 0.1 seconde est dédiée à une rampe d'attaque et une rampe d'extinction de l'amplitude du signal, afin d'éviter tout phénomène de *clic*, lié à la discontinuité imposée à la position de la membrane des haut-parleurs au début et à la fin des séquences.

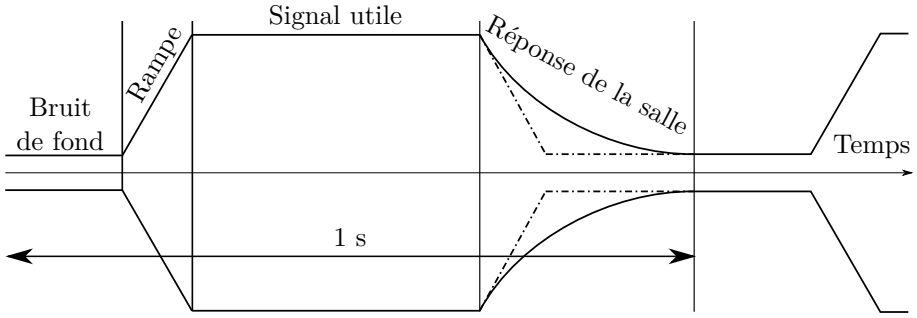


FIGURE 4.9 – Enveloppe schématique des signaux enregistrés

Ainsi, pour chaque couple de position de source et de signal, 1 seconde de signal utile est exploitable avec une amplitude maximale, permettant ainsi l'extraction, pour chaque position de source, de plus de 20 trames différentes de 1 024 échantillons (taille d'entrée pour le réseau de neurones présenté 2.1.2). Par ailleurs, entre chaque synthèse de champs, un silence d'une durée de 0,5 s est imposée. Puisque cette durée de silence entre deux synthèses de champs excède largement la durée de réverbération de la salle de restitution, nous nous assurons ainsi que la captation d'un champ correspondant à une source virtuelle n'est pas impactée par la réverbération de la précédente. Afin de maximiser la diversité de trames audio obtenues pour un même fichier *.wav* émis par les sources virtuelles synthétisées par le spatialisateur, l'échantillon de début de lecture est tiré aléatoirement pour chaque synthèse de champs.

4. Toute la partie pilotage de la sphère grâce au langage Faust ne fait pas partie de ce travail, et ne sera donc pas abordée en détail.

Compte tenu de ces spécifications, une acquisition automatisée de jeux de données d'une durée de 12 heures en période nocturne permet d'accumuler 43 200 synthèses ambisoniques de positions de sources différentes, dans lesquelles on peut extraire, pour chacune, une vingtaine de trames différentes de 1 024 échantillons.

Les positions angulaires des sources virtuelles étant définies par rapport à un repère centré sur le centre de la SpherBedev, le plus grand soin doit être pris lors du positionnement de l'antenne, afin de ne pas introduire de biais de position lors des acquisition des jeux de données. L'antenne est positionnée grâce à un niveau laser 3 axes, et elle n'est pas démontée pendant toute la phase d'acquisition.

4.2.5 Découpe et étiquetage des données acquises par les antennes dans le spatialisateur

Les paramètres d'émission sont gérés par la station de pilotage du spatialisateur SpherBedev. L'acquisition des signaux de sortie des antennes microphoniques est quant à elle gérée sur une autre station. Ici, plutôt que de réaliser un enregistrement séquentiel, l'acquisition des données multicanales est réalisée en une seule fois, dans un fichier audio multicanal d'une durée de 12 heures, et ce fichier est ensuite *découpé* et étiqueté. Le fait de réaliser cette acquisition dans une seul fichier audio de plusieurs heures implique des contraintes techniques, qui nécessitent d'être éclaircies ici.

Dans un premier temps, les formats de fichiers audio sans pertes sont pour la plupart limités à une taille maximale, essentiellement liée historiquement au fait que les spécifications de ces formats ont été définies lorsque les systèmes informatiques étaient limités à 32 bits. En particulier, le format *.wav* n'est pas conçu pour excéder une taille de fichier de 4 Go. Parmi l'ensemble des formats d'encodage de fichiers audio, l'un des rares formats adaptés à l'enregistrement sans limite de durée et de nombre de canaux, est le format *.rf64*, qui est beaucoup plus flexible pour les objectifs liés à nos acquisitions automatisées de larges quantités de données. À titre d'exemple, les acquisitions nocturnes réalisées avec l'antenne sphérique Zylia *pèsent* chacune 112 Go.

Par ailleurs, pour une approche d'apprentissage supervisé, il est nécessaire d'attribuer un label à chaque portion de ce fichier correspondant à une synthèse de champ différente. Ce label correspond essentiellement à la position de la source virtuelle synthétisée par le spatialisateur 3D. L'ordre des

positions étant défini par un fichier *.txt* en entrée du programme Pure Data pilotant les paramètres de spatialisation, il suffit en principe d'appairer les instructions de positions de sources (et tous les éléments utiles pour la base de donnée JSON présentée en section 4.2.6) et les portions de signal multicanal acquis.

Pour cela, il est nécessaire d'être en mesure de sélectionner chaque portion de l'acquisition dans le fichier de 12 heures, ce qui revient à diviser celui-ci en 43 200 portions utiles (voir figure 4.9). Pour cela, puisque deux stations informatiques différentes sont utilisées pour le pilotage de la sphère de spatialisation et pour l'acquisition sur les antennes microphoniques, nous avons fait le choix de ne pas exploiter les timecodes de pilotage. En effet, l'horloge de la station d'acquisition et de la station de pilotage sont très légèrement différentes (les mesures réalisées ont permis de mettre en évidence une différence d'environ 1% entre les deux fréquence d'horloge, ce qui est très commun, mais correspond à une dérive d'environ 5 secondes sur une période de 12 heures). La solution qui a été retenue pour tronçonner le signal repose donc sur une fonction de détection automatique de seuils dans le fichier audio de 12 heures. Puisque la synthèse et la captation reposent toutes les deux sur l'utilisation du serveur de son temps réel Jack [186], une solution reposant sur l'utilisation de NetJack [187] et du protocole OSC [185] permettra pour les acquisitions réalisées à la suite de ma thèse de simplifier cette approche, en enregistrant directement les labels comme une piste audio supplémentaire véhiculée par protocole réseau jusqu'à la station d'acqusition connectée à l'antenne microphonique.

4.2.6 Gestion de la base de données : les schémas JSON

Même les architectures de réseaux de neurones profonds les mieux *taillés* pour une tâche donnée nécessitent une base de donnée d'apprentissage conséquente et adaptée pour généraliser au mieux un problème et converger vers une solution efficace. En cela, la base de données d'apprentissage est l'une des briques primordiales pour toute méthode reposant sur des techniques de Deep Learning. Pour pouvoir entraîner un réseau sur un grand nombre de données pilotées par des paramètres haut niveau, comme le type de salle, leur géométrie, le type de source, le type de signal émis, ou le type d'antenne microphonique, tout en offrant une gestion modulaire et aisée de ces bases de données, il est primordial d'utiliser un système d'archivage performant. Pourtant, la gestion des bases de données est un problème rarement abordé dans le domaine de l'intelligence artificielle. Dans la plupart des cas, les

données sont fournies dans une arborescence de dossiers et de sous-dossiers. Cette approche a le mérite de pouvoir être prise en main très rapidement pour des bases de données très simples, quelque soit le langage de programmation ou le *framework* de Deep Learning utilisé. Cependant, dans un contexte tel que la tâche d'apprentissage visée dans le cadre de cette thèse, le fait de simplement sauvegarder les différents fichiers audio dans une arborescence de dossier se révèle être une gestion beaucoup trop basique pour être efficace, puisque limitante lorsque l'arbre n'est pas parfaitement connu à l'avance. De même, la navigation entre branches n'est pas aisée. Enfin, cette approche est par essence hiérarchisée entre les différents critères des données, alors que la tâche et les jeux de données construits nécessitent de pouvoir constituer rapidement des jeux de données sur la base de plusieurs paramètres à la fois, et ce, de manière flexible et aisée.

Pour toutes ces raisons, on propose d'utiliser un système de gestion de base de données (SGBD) afin de permettre le stockage de toutes les informations pertinentes liées à la sauvegarde de chaque fichier, mais également l'ajout de caractéristiques sur certains fichiers de manière simple, et d'offrir une navigation aisée entre les fichiers selon n'importe quel critère. L'objectif n'est pas de détailler ici les 12 règles de Codd [188], mais seulement de justifier le choix du format utilisé pour la gestion de la base de données.

Plusieurs formats ont été étudiés pour satisfaire ces objectifs : le CSV (*Comma-separated values*) [189], le XML (*Extensible Markup Language*) [190] et le JSON (*JavaScript Object Notation*) [143]. Le CSV n'étant pas assez facilement modulaire, a rapidement été abandonné puisqu'il n'offrait pas la flexibilité que nous recherchions. Le choix entre le XML et le JSON s'est fait sur un critère de simplicité d'utilisation. Le JSON a un format d'écriture correspondant au format *dict* de Python. De plus la librairie *pandas* [191], souvent utilisée dans la littérature, ne reconnaît pas nativement le format XML. Ainsi, le choix définitif s'est porté sur le JSON. De plus, ce format est moins *verbeux* que le XML, et très facilement lisible.

En particulier, le format JSON permet la création d'un schéma permettant de définir la forme sous laquelle les informations sur les données doivent être sauvegardées. Ainsi, les bases de données ne présentent ni d'informations redondantes, ni d'informations mal répertoriées, ni d'informations

manquantes. Il a de plus été choisi de fixer l'ensemble des données admissibles, avec leur type ainsi que l'ensemble des données nécessaires à minima pour pouvoir rajouter un fichier dans la base de données. Grâce à cette approche, les données peuvent être triées selon n'importe quel critère de manière quasi instantanée, ce qui permet d'étudier finement mais simplement l'influence du type de salle, de sources, et de signaux utilisés pour l'entraînement du réseau pour une tâche de localisation angulaire. Pour information, les schémas JSON construits pour les bases de données de ce projet sont fournis explicitement en annexe C. Toutes les informations des données, issues de simulation ou d'acquisitions expérimentales, sont donc sauvegardées et gérées grâce à ce langage.

4.3 Synthèse des apports principaux liés à ce chapitre

Le spatialisateur 3D par ambisonie d'ordres élevés SpherBedev du laboratoire LMSSC est un dispositif permettant de proposer une solution originale et efficace à la constitution de jeux de données expérimentaux, avec un principe de synthèse offrant une précision et un réalisme physique tout à fait adaptés à notre problème. Ces jeux de données expérimentaux, grâce auxquels les réseaux de neurones sont entraînés, présentent également l'avantage de contenir les caractéristiques propres aux antennes microphoniques auxquelles on adjoint une intelligence artificielle : contrairement aux jeux de données simulées ou aux approches par modèles classiquement utilisées, ces mesures permettent implicitement d'intégrer la diffraction de la structure de l'antenne et de son support de fixation, mais également la réponse en fréquence imparfaite et individuelle de chaque capteur composant l'antenne, sans avoir explicitement à la mesurer ou à la modéliser. Ainsi, les données obtenues sont au plus près des signaux qui seront effectivement captés lors de l'utilisation des antennes microphoniques intelligentes en situation réelle, ce qui permet d'envisager une intelligence artificielle individualisée en fonction de chaque dispositif, reposant sur la même architecture de réseau de neurones profond.

En ce qui concerne le processus de synthèse de champs acoustique exploité ici, le formalisme d'ambisonie d'ordre 5 utilisé ici assure au spatialisateur la validité de la reconstruction physique du champ de pression au centre de la sphère de spatialisation, dans une zone de 7 cm de rayon au minimum, dans le cas d'un signal à 4 000 Hz. Cet ordre de grandeur est donc tout à fait compatible avec les antennes microphoniques compactes utilisées dans le cadre de cette thèse de doctorat.

Le choix de plusieurs topologies d'antennes répondant à ce critère a été justifié, et un protocole de mesure expérimental original a été présenté, permettant de réaliser de manière automatisée l'acquisition et l'étiquetage de plusieurs dizaines de milliers de champs émis par des sources dont la position est pilotée sans intervention humaine. Compte tenu du volume et de la variété de données nécessaires pour effectuer une tâche de localisation de sources par apprentissage supervisé, cette automatisation est primordiale, et permet d'envisager la constitution de jeux de données réalistes expérimentaux, aussi variés que nécessaire.

Grâce aux jeux de données expérimentaux ainsi constitués, il est désormais possible d'éprouver les performances de l'approche BeamLearning, non seulement dans un contexte de simulations numériques, mais également dans le cas de captations réelles, et de comparer ces performances à celles d'algorithmes provenant d'approches *modèles*. Cette analyse est menée dans le cadre du chapitre 5 qui suit.

Chapitre 5

Analyse des performances de localisation offertes par l'approche BeamLearning

Si l'on considérait une théorie comme parfaite, et si l'on cessait de la vérifier par l'expérience scientifique, elle deviendrait une doctrine, (*Introduction à l'étude de la médecine expérimentale*, Claude Bernard, 1865)

Contenu du chapitre

Au cours des trois années de ma thèse de doctorat, l'approche BeamLearning a évolué itérativement, en commençant avec un problème de classification angulaire à 2 dimensions en champ libre pour des signaux monochromatiques, jusqu'à un problème de régression angulaire traitant le problème de la détermination de DOA en 3 dimensions en milieu réverbérant. Ces évolutions ont été accompagnées de jeux de données variés, et la structure du réseau de neurones a été itérativement améliorée, jusqu'à atteindre celle décrite au chapitre 2 de ce document. Au delà du choix de l'approche de classification ou de l'approche de régression pour résoudre le problème de localisation angulaire, le type de signaux émis par les sources à localiser, et le type d'environnement de mesure dans lequel est disposée l'antenne microphonique sont autant de paramètres qui définissent la complexité du problème physique à résoudre, et ont motivé des améliorations sensibles du réseau de neurones profond conçu spécifiquement pour réaliser ces tâches. Plutôt que de présenter les résultats obtenus par chaque version du réseau comme ils l'ont été au cours du développement, dans un souci de synthèse, ce chapitre présente les performances de localisation pour chacun des problèmes donnés, avec une seule et unique version du réseau de neurones profond (celui présenté au chapitre 2), pour les différents problèmes suivants, du plus simple au plus complexe :

— DOA 2D en champ libre ou en environnement traité acoustiquement par classification angulaire,
— DOA 2D par régression en environnement quelconque (traité ou réverbérant),
— DOA 3D par régression en environnement quelconque (traité ou réverbérant).

Pour chacun de ces trois volets, les performances seront étudiées à la fois pour des jeux de données obtenus par simulations numériques (voir chapitre 3), ainsi que pour des jeux de données obtenus expérimentalement (voir chapitre 4).

5.1 Détermination de DOA 2D par classification angulaire pour des sources monochromatiques

5.1.1 Étude d'une situation idéale : champ libre, sans bruit de mesure

Dès les premières semaines de cette thèse de doctorat, le problème de localisation a été abordé à l'aide d'un cas d'étude simple, permettant d'analyser le comportement de l'approche d'apprentissage

supervisé pour la localisation de sources : celui-ci correspond au problème idéal de localisation de sources monochromatiques, en champ libre, sans bruit de mesure. Pour ce problème, aucune difficulté ne se pose, puisque c'est un cadre théorique idéal. En revanche, même si ce cadre est très idéalisé, il permet de mettre en évidence le comportement des sorties du réseau. Par ailleurs, dans le cas de figure étudié ici, les sources monochromatiques utilisées pour la base de données d'apprentissage ont volontairement été restreintes à un petit ensemble de fréquences, correspondant aux fréquences centrales des bandes d'octaves, dans la gamme fréquentielle visée pour nos applications : [125, 250, 500, 1000, 2000,4000] Hz. Pour rappel, ce domaine est essentiellement défini pour respecter la gamme fréquentielle de reconstruction valide du spatialisateur 3D SpherBedev présenté au chapitre 4.

Un résumé des caractéristiques de cette base de données d'apprentissage est fourni dans le tableau 5.1. Les signaux microphoniques utilisés comme informations d'entrée du réseau sont ici simulés grâce aux outils présentés dans le chapitre, dans une situation très simple, puisque le milieu de mesure simulé est ici parfaitement anéchoïque. L'antenne utilisée pour ces simulations est l'antenne circulaire Mini DSP à 7 microphones, présentée en section 3.2.1.

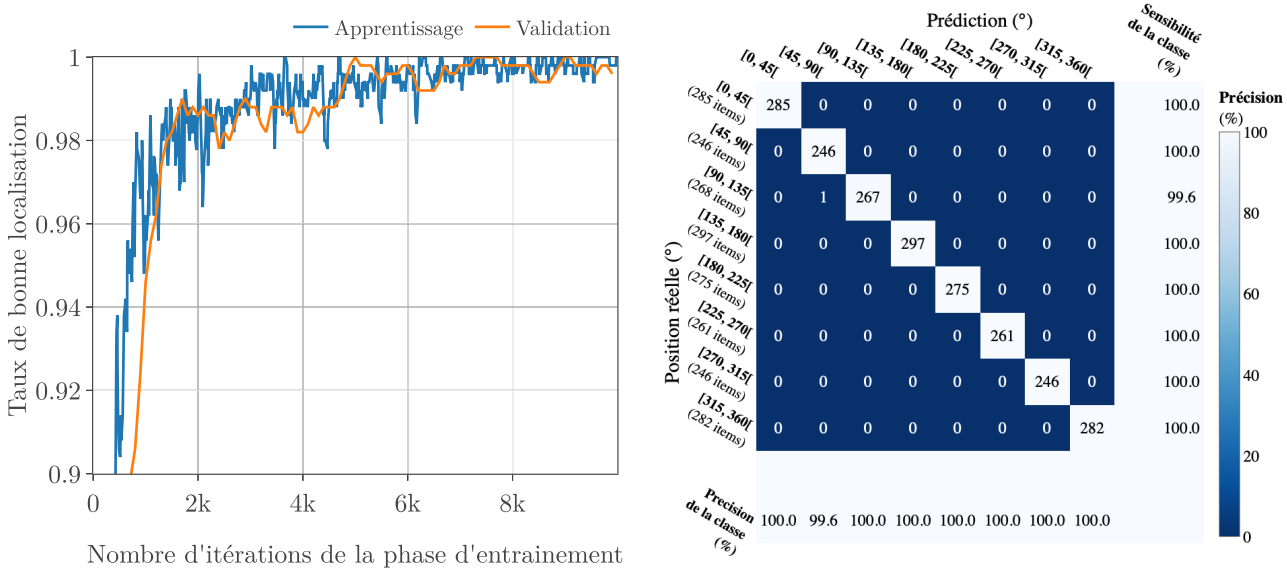
Données	Base de données	Environnement	Signal	Antenne	RSB
	Simulée	Champ libre	Sinus pur	Mini DSP.	$+\infty$
Résultats	Sortie	Nombre de classes	Précision	Nb. itérations	Sur apprentissage
	Classification	8	99,8%	10 000	Non

Tableau 5.1 – Récapitulatif synthétique des paramètres pour l'apprentissage présenté en section 5.1.1

Au total, pour constituer le jeu de données sous-jacent, 38 400 positions de sources ont été utilisées. Les positions de chacune de ces sources sont tirées aléatoirement dans un tore de rayon $2 \pm 0,5$ m, comme expliqué en section 3.2.3. Afin d'augmenter la diversité des exemples, même si on s'intéresse dans cette sous-section exclusivement à un problème de DOA 2D, une faible variabilité en élévation est ajoutée : ainsi, l'ouverture angulaire du tore $d\Phi$ vaut ici 7°. Pour chacune de ces positions, les captations du champ de pression sont simulées pour une émission à chacune des fréquences centrales des bandes d'octaves considérées, ce qui correspond à une base de données constituée de 230 400 exemples de

captations de champs sur l'antenne testée. L'apprentissage est ici réalisé pendant 10 000 itérations, chacune d'entre elles correspondant à la présentation en entrée du réseau d'un lot d'apprentissage aléatoire de 100 exemples de captations de champs de pression. Compte tenu du volume de la base de données, l'apprentissage est donc réalisé sur 4,3 époques [1] , correspondant à 1h45 de calcul sur une carte GPU Nvidia 1080Ti du serveur de calcul utilisé.

Pour cette situation simple, comme attendu, on peut observer sur la figure 5.1(a) qu'au terme de l'apprentissage, la précision obtenue pour la classification angulaire à 8 zones atteint une valeur de 99 à 100 %, tant sur la base de données d'apprentissage que sur les données de validation (données non présentées pendant les itérations d'entrainement, impliquant la mise à jour des variables du réseau et du mécanisme de rétropropagation des erreurs pour l'optimisation).



(a) Courbe d'apprentissage de BeamLearning (b) Matrice de confusion à partir de données de test

FIGURE 5.1 – Performances de l'approche BeamLearning dans le cas de classification de données simulées monochromatiques en champ libre sans bruit. (a) : Courbe de convergence d'apprentissage du réseau obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement. (b) Matrice de confusion obtenue sur l'ensemble du jeu de données de validation, pour la dernière itération de l'apprentissage.

1. Dans le domaine de l'apprentissage supervisé, le nombre d'époques correspond au nombre de fois que le jeux de données d'apprentissage est présenté dans son intégralité lors de la phase d'optimisation (on conserve également ce formalisme d'époques lorsque on fait appel à un mécanisme d'augmentation de données, qui peut modifier légèrement les données des exemples d'une présentation à une autre).

Afin d'analyser plus finement ces performances de classification, la figure 5.1(b) présente la matrice de confusion obtenue sur l'ensemble du jeu de données de validation à l'issue de l'apprentissage. Sur cette figure, on peut observer que seule une source parmi les 2 159 sources présentées au réseau a été classée dans un secteur angulaire différent de celui auquel elle appartient réellement. De plus, la figure 5.1(a) prouve que l'apprentissage converge avant les 10 000 itérations, et on obtient ces résultats très satisfaisants dès 4 500 itérations. Bien entendu, ces excellents résultats ne reflètent à ce stade que le fait que le problème présenté est idéalisé, et qu'il ne pose aucun problème à résoudre, que ce soit avec une approche modèle conventionnelle ou une approche d'apprentissage supervisé comme celle proposée ici. Dans la section suivante, un apprentissage similaire est réalisé, cette fois-ci en exploitant des données captées par l'antenne, avec un rapport signal à bruit dégradé.

5.1.2 Ajout de bruit de mesure pour une classification de DOA 2D de sources monochromatiques en champ libre

Afin d'étudier une situation plus réaliste, il est nécessaire de prendre en compte l'influence du bruit de mesure sur chacun des canaux microphoniques. Pour cela, la situation étudiée à la section précédente est reprise, toujours pour une situation de localisation en champ libre, en ajoutant du bruit sur chacune des voies d'entrée du réseau. Cet ajout correspond classiquement à la modélisation du bruit de fond observé sur les canaux microphoniques, provenant à la fois du bruit électrique de sortie intrinsèque à chacun des capteurs de l'antenne [2] . Ces bruits sont en général modélisés par des signaux aléatoires, et décorrélés entre toutes les voies microphoniques. Ainsi, pour chaque exemple de la base de données exploitée pour l'apprentissage et la validation, une nouvelle réalisation d'un signal aléatoire correspondant à une loi de type bruit blanc est générée pour chaque canal microphonique.

Comme ces signaux sont générés à la volée au cours de l'entraînement du réseau, même lorsqu'un élément de la base de données est présenté une nouvelle fois en entrée du réseau lorsque le nombre d'itérations d'entraînement dépasse une époque, le bruit ajouté à la mesure sur l'antenne est différent de celui rajouté lors des époques d'entraînement précédentes. Puisque les signaux d'entrée du réseau sont filtrés par un filtre passe bande sélectif entre 100 Hz et 4 000 Hz, les signaux de bruits blancs ajoutés sur chaque canal pour chaque élément subissent le même filtrage (voir sec. 2.1.2). L'énergie

2. En aucun cas ce bruit blanc ne peut modéliser la présence d'une autre source perturbatrice dans la pièce. Sinon les bruits seraient au contraire corrélés entre les voies microphoniques.

de chaque bruit blanc filtré est ensuite calculé, et comparée à l'énergie E_s du signal non bruité du microphone auquel il est attribué. L'amplitude du bruit est ensuite fixée aléatoirement, modifiant ainsi l'énergie E_b du bruit de fond de chaque capteur, de manière à ce que le rapport signal sur bruit défini par l'équation 5.1 soit supérieur à une valeur à ne pas dépasser :

$$RSB = 10 \log \left(\frac{E_s}{E_b} \right) \tag{5.1}$$

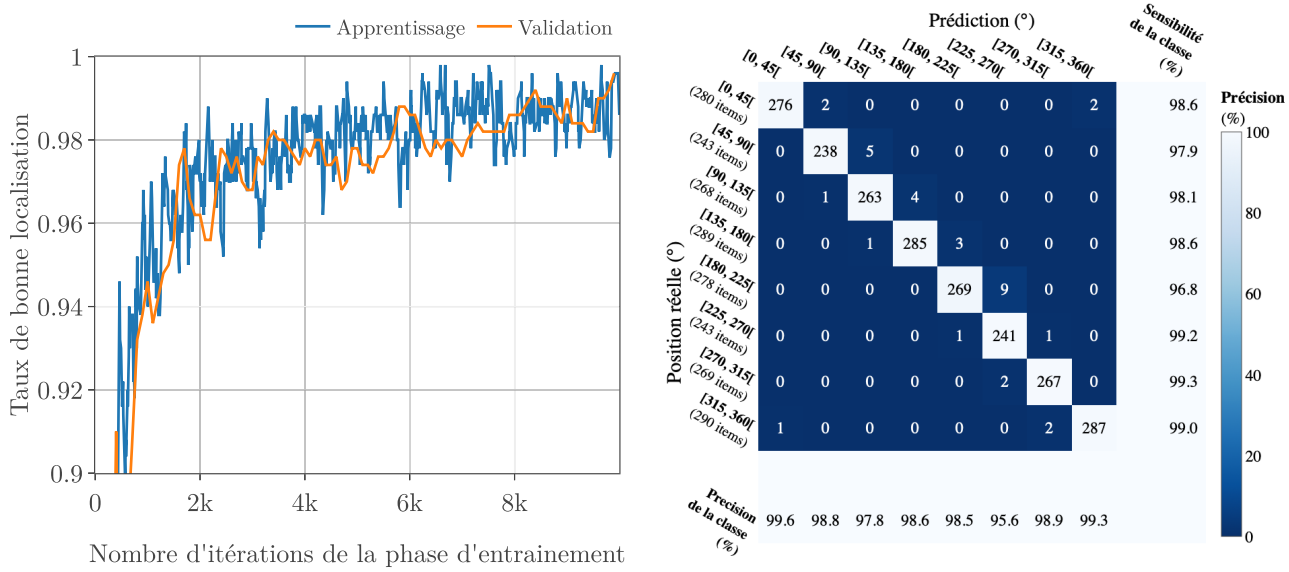
Ainsi, lorsque le RSB défini est de 0 dB, le bruit est *aussi fort* que le signal correspondant uniquement au champ émis par la source à localiser et capté sur le microphone de l'antenne. Lorsque le RSB vaut $+\infty$, on retrouve la situation étudiée à la section 5.1.1.

Lorsque le RSB minimal autorisé est de 20 dB par exemple, l'ensemble des scalaires tirés aléatoirement pour modifier l'amplitude du bruit ajouté à chaque canal pour chaque élément présenté en entrée du réseau permet d'obtenir un ensemble de valeurs de RSB pour toutes les voies microphoniques et tous les exemples présentés en entrée du réseau, qui soient contenus dans l'intervalle $[20, +\infty[$. Puisque le bruit de fond modélisé est essentiellement lié au bruit de fond de l'environnement de mesure et au bruit de fond intrinsèque des capteurs de l'antenne, il apparaît naturel que ce RSB soit identique pour toutes les voies microphoniques de l'antenne, pour un élément donné présenté en entrée du réseau. En revanche, puisque l'objectif est ici d'entraîner le réseau à s'affranchir de ce bruit de mesure sans a priori sur son amplitude par rapport au champ émis par la source à localiser, la procédure proposée est conçue pour offrir une variabilité de RSB en fonction des éléments de la base de données, mais également en fonction des époques d'entraînement.

Afin de vérifier la robustesse de la localisation de sources à l'ajout de bruit de mesure et comparer les résultats à une situation idéale, les mêmes conditions que dans la section précédente sont utilisées (voir tableau 5.2). Ici encore, les sources à localiser sont des sources monochromatiques de fréquences [125, 250, 500, 1000, 2000,4000] Hz. La base de données constituée correspond toujours à 38 400 positions possibles, dans un tore de rayon $2 \pm 0,5$ m et d'ouverture angulaire en élévation de $\pm$ 7°, afin d'estimer la DOA 2D des sources par une approche de classification angulaire à 8 classes azimutales. Le RSB minimal autorisé lors de l'apprentissage est de 20 dB.

Données	Base de données	Environnement	Signal	Antenne	RSB
	Simulée	Champ libre	Sinus pur	Mini DSP.	> 20 dB
Résultats	Sortie	Nombre de classes	Précision	Nb. itérations	Sur apprentissage
	Classification	8	99,8%	10 000	Non

Tableau 5.2 – Récapitulatif synthétique des paramètres pour l'apprentissage présenté en section 5.1.2



(a) Courbe d'apprentissage de BeamLearning (b) Matrice de confusion à partir de données de test

FIGURE 5.2 – Performances de l'approche BeamLearning dans le cas de classification de données simulées monochromatiques en champ libre Performances de l'approche BeamLearning dans le cas de classification de données simulées monochromatiques en champ libre sans bruit. (a) : Courbe de convergence d'apprentissage du réseau, obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement. (b) Matrice de confusion obtenue sur l'ensemble du jeu de données de validation, pour la dernière itération de l'apprentissage.

La figure 5.2 présente les résultats obtenus pour cet apprentissage, et l'analyse de la figure 5.2(a) révèle qu'ici encore, au terme des 10000 itérations d'entrainement, la précision de localisation atteint toujours une valeur de 99 à 100 %, même si les données d'entrée sont bruitées. En revanche, l'analyse de la matrice de confusion obtenue à partir du jeu de données de validation à l'issue de la dernière itération révèle une légère perte de performances, puisque la précision obtenue pour chaque classe est

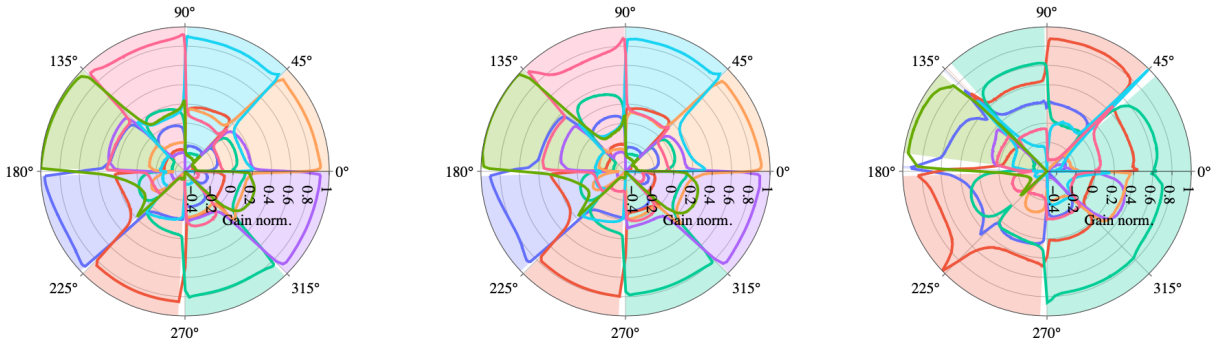
ici légèrement réduite, à une valeur de 98 à 99 %. Les résultats restent toutefois très satisfaisants, puisque seules 33 sources parmi les 2 159 sources présentées au réseau ont été classées dans un secteur angulaire différent de celui auquel elle appartient réellement, mais le secteur de classification est systématiquement contigu au secteur réel d'appartenance de ces sources. De plus, on peut voir qu'aucune zone de l'espace n'est significativement sur-représentée par le réseau, puisque les sensibilités sont toutes très proches les unes des autres (à droite de la figure 5.2(b)). De même, aucune zone de l'espace n'est moins bien localisée que les autres, car les précisions obtenues pour chaque classe possèdent toutes des valeurs très similaires. La convergence du réseau 5.2(a) est néanmoins légèrement plus longue que précédemment, puisque les résultats en test se stabilisent à partir de 6 500 itérations (soit après 1h10 de calcul sur une carte GPU Nvidia 1080Ti).

5.1.3 Analyse de la directivité du réseau

Dans le cas de la localisation de sources par classification, le vecteur de sortie contient la réponse de chaque neurone de sortie (voir sec. 2.2.1). Ainsi, pour connaître le diagramme de directivité global du réseau de neurones, il faut superposer les diagrammes de directivité de chaque neurone. La figure 5.3 présente ces diagrammes de directivité pour différentes fréquences. Les fréquences choisies sont ici 2 000 Hz, 2 200 Hz, et 2 500 Hz, permettant ainsi de visualiser le comportement à une fréquence faisant partie du jeu de données d'apprentissage, et à deux fréquences s'écartant de celles-ci, jamais présentées en entrée du réseau pendant la phase d'entraînement. Sur la figure 5.3, la couleur de fond des portions de diagrammes angulaires représentent quel neurone de sortie réagit le plus dans la direction donnée (c'est à dire la classe angulaire prédite par le réseau). Les courbes de couleurs représentent quant à elles la réponse de chaque neurone dans toutes les directions.

L'analyse de cette figure révèle que pour chaque direction, il y a au moins un neurone de sortie qui présente une réponse de forte amplitude (entre 0,7 et 1). Cette observation permet de comprendre qu'aucune direction n'est délaissée par le réseau. En revanche, l'analyse du comportement du réseau pour les 3 fréquences permet de mettre en évidence le fait que la directivité de chaque neurone de sortie du réseau varie fortement avec la fréquence.

Pour rappel, pour le cas d'apprentissage présenté dans cette section, seules les fréquences centrales



(a) 2 000 Hz (b) 2 200 Hz (c) 2 500 Hz

FIGURE 5.3 – Diagrammes de directivité de BeamLearning à différentes fréquences lors d'une classification à 8 classes. Apprentissage avec bruit (jusqu'à 20 dB de RSB) sur des fréquences pures ([125, 250, 500, 1000, 2000,4000] Hz) en champ libre.

de bandes d'octaves entre 125 Hz et 4 000 Hz ont été utilisées pour constituer la base de données d'apprentissage de sources monochromatiques. Pour la fréquence 2 000 Hz (figure 5.3(a)) on observe que chaque neurone réagit de manière prépondérante exclusivement dans une direction particulière. Comme les 8 classes utilisées pour discrétiser l'espace sont équi-répartis, chaque neurone est actif dans un huitième d'espace. En revanche, en dehors de leur zone d'activité, la réponse de chaque neurone de sortie du réseau chute drastiquement, ce qui ne laisse aucune ambiguïté sur la réponse globale du réseau et permet de comprendre les excellents résultats obtenus en termes de sensibilité et de spécificité pour cette fréquence.

Lorsque l'on s'écarte légèrement de cette fréquence et que l'on présente en entrée du réseau un champ monochromatique à une fréquence non présente pendant la phase d'entraînement (2 200 Hz, figure 5.3(b)), le comportement du réseau est quasiment identique à celui de la figure 5.3(a) à 2 000 Hz. Ce résultat met en évidence une relative robustesse du réseau à une variation de fréquence : l'algorithme permet de localiser efficacement à des fréquences légèrement différentes de celles présentes dans la base de donnée d'apprentissage. En revanche, lorsque la différence de fréquence du champ mesuré est trop importante par rapport aux fréquences d'entraînement, comme pour la figure 5.3(c) à 2 500 Hz, cette robustesse chute drastiquement, puisqu'on observe seulement 4 classes prédites en sortie du réseau, pour 8 classes possibles. Par exemple, la couleur vert d'eau, correspondant à la zone

angulaire contenue dans l'intervalle [270°, 315°], est pourtant prédite à tord par le réseau pour quatre zones angulaires. À cette fréquence trop écartée de celles utilisées pour l'entraînement du réseau, le réseau de neurones profond classera donc (à tord) dans la direction [270°, 315°] des sources provenant de directions dans les secteurs angulaires suivants : [315°, 0°], [0°, 45°] et [90°, 135°].

Cet exemple didactique permet ainsi de démontrer l'importance de la constitution des jeux de données d'entraînement, puisque ce n'est pas ici le réseau en tant que tel qui est réellement en cause dans ce type d'erreur de classification, mais le jeu de données utilisées pour l'entraîner. Même si le réseau offre une relative robustesse à une faible variation de fréquence par rapport à celles de la base de données d'apprentissage, lorsque les signaux sont trop différents fréquentiellement, les résultats ne sont plus pertinents. Cette observation pousse évidemment à développer une base de données d'entraînement constituée de signaux beaucoup plus variés d'un point de vue fréquentiel, afin d'obtenir une technique de localisation qui soit efficace sur une large gamme de fréquences. Cette observation motivera les jeux de données exploités à partir de la section 5.2.3 de ce manuscrit. Mais avant d'aborder cet élément, l'analyse des performances de classification angulaire à 2D fera l'objet d'une discussion sur les avantages offerts par l'approche BeamLearning pour la localisation expérimentale, lorsque les fonctions de réponse en fréquence des capteurs composant l'antenne sont inconnues (voir section 5.1.4). Pour mettre en évidence cette propriété, une validation expérimentale de la méthode sera proposée à l'aide de mesures en environnement semi-anéchoïque, sans étalonnage préliminaire des capteurs de l'antenne utilisée (voir 5.1.5).

5.1.4 Étalonnage implicite des capteurs de l'antenne grâce à l'apprentissage

L'un des problèmes inhérents à toutes les méthodes *conventionnelles* de localisation de sources acoustiques reposant sur une approche *modèle*, est le nombre d'hypothèses simplificatrices faites pour pouvoir modéliser le problème. Le point commun à toutes les méthodes *modèles* de la littérature exploitant une antenne microphonique repose sur la supposition que les réponses individuelles en amplitude, en phase (et en directivité) de chacun des capteurs composant l'antenne sont parfaitement connues. En effet, les signaux de ces capteurs sont systématiquement supposés comme idéaux et proportionnels au champ de pression mesuré, par compensation des réponses individuelles obtenues à l'aide d'une phase d'étalonnage individuel [192].

Au delà de cette hypothèse forte et de la contrainte pour l'expérimentateur, la diffraction par la structure même de l'antenne et de son support de fixation est en général négligée, mise à part dans les rares cas où celle-ci est parfaitement connue et calculable analytiquement (cas des antennes sphériques rigides, par exemple). Ce point primordial est pourtant en général survolé dans la littérature scientifique, alors que la précision des méthodes *modèles* repose en grande partie sur la connaissance fine des caractéristiques individuelles des microphones de l'antenne [193]. Bien entendu, dans la plupart des cas, les acousticiens expérimentateurs sont parfaitement conscients de l'importance de l'étalonnage individuel en phase et en amplitude des capteurs microphoniques d'une antenne, mais cette tâche est en pratique fastidieuse et nécessite la plupart du temps un protocole long et un soin particulier [194].

Malheureusement, pour certaines technologies émergentes de microphones basés sur des systèmes sur puces (microphones MEMS), il n'existe aujourd'hui aucun consensus sur le protocole d'étalonnage, ni même de norme internationale de mesurage des caractéristiques de réponse en fréquence, puisque le fait que ces composants de taille très réduites soient intégrés sur des circuits imprimés complique sensiblement les protocoles d'étalonnage, même si des solutions originales ont été proposées par la communauté scientifique [195]. Fort heureusement, le processus de fabrication et de sélection des microphones MEMS offre une variabilité très réduite entre différents capteurs issus d'un même lot de production [196], ce qui explique pourquoi certains auteurs choisissent même de ne pas étalonner individuellement ces capteurs lorsqu'ils sont intégrés en très grand nombre sur une antenne [196].

Dans le cas des microphones électrostatiques de métrologies exploités plus classiquement dans les antennes microphoniques, il existe des normes précises d'étalonnage (NF EN 61094), avec un protocole précis offrant des incertitudes relativement réduites pour un étalonnage individuel des capteurs, lorsque les installations nécessaires sont à disposition [194]. En revanche, aucune procédure standardisée n'existe pour un étalonnage des capteurs sur la structure de l'antenne, ce qui empêche ainsi de compenser l'éventuelle diffraction induite par cette structure.

Pourtant, une prise en compte approximative des caractéristiques des capteurs et de la présence de la structure de l'antenne peut entraîner une augmentation sensible des erreurs de localisation de

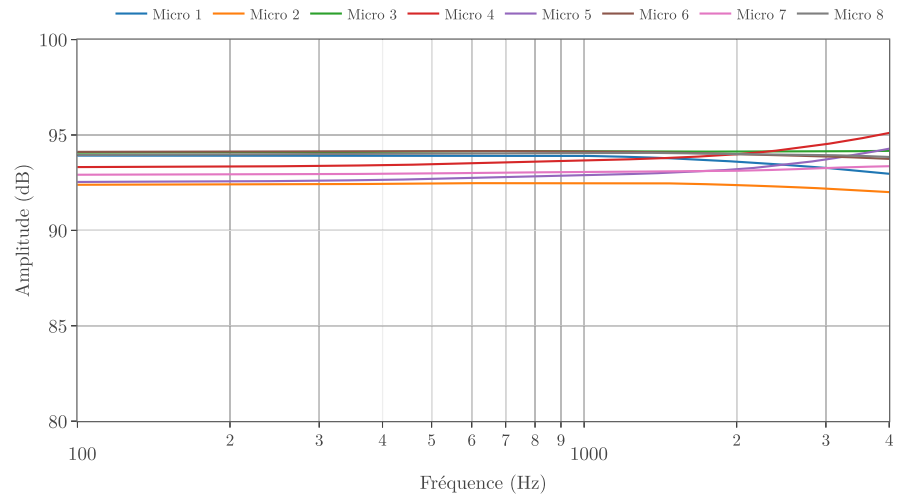
sources lorsque les approches *modèles* sont utilisées. À l'opposé des approches *modèles*, l'approche de localisation par apprentissage proposée dans cette thèse de doctorat offre une solution élégante à la prise en compte intrinsèque de l'étalonnage des microphones et de la structure de l'antenne [3] , sans nécessiter de protocole lourd d'étalonnage. En effet, les réponses individuelles des capteurs (amplitude, phase, directivité intrinsèque et induite par la diffraction) sont naturellement incluses dans les champs mesurés par l'antenne au cours de la constitution des bases de données [144].

Afin de mettre en évidence cette propriété originale qui constitue l'une des forces des approches reposant sur les données, plusieurs expériences de localisation ont été réalisées à partir d'un jeu de données simulées, et d'un ensemble de mesures de réponses en fréquences individuelles de microphones à capsule à électret du laboratoire. Ainsi, l'utilisation du jeu de données simulées correspond aux mesures qui seraient réalisées par un ensemble de microphones *parfaits*, ou parfaitement corrigés par leurs réponses individuelles. Les mesures qui seraient réalisées par un ensemble de microphones *réels* correspondent quant à elles à l'utilisation du jeu de données simulées, auxquelles on a appliqué les fonctions de réponses en fréquences individuelles des capteurs.

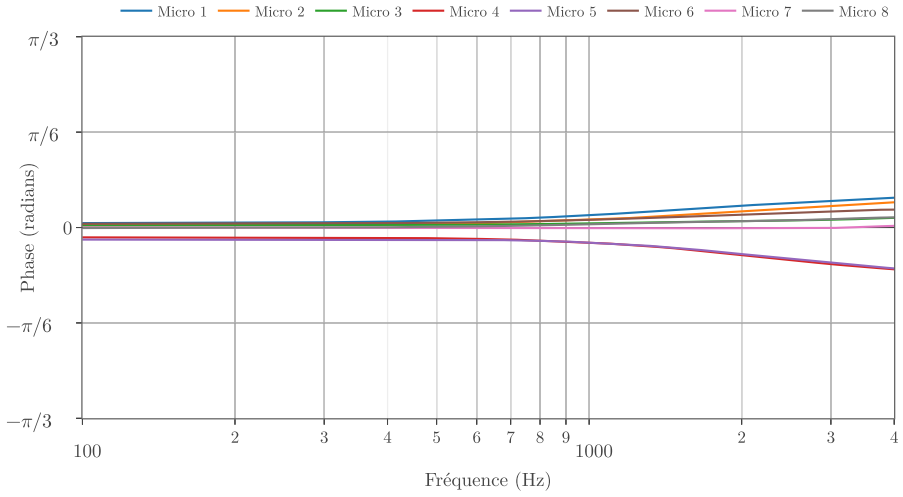
La figure 5.4 présente les huit réponses en fréquence (FRF) utilisées dans cette section, en particulier leur module et leur phase. Ces réponses ont été obtenues au laboratoire par un procédé d'étalonnage individuel en tube en laiton, pour des microphones 1/4" issus du même lot, et fabriqués par le CTTM à partir de capsules à électret de qualité. On peut observer sur ces figures que ces microphones, même s'ils sont issus d'un même lot de production par un organisme spécialisé, possèdent une (faible) variabilité de réponse en phase et en amplitude dans le domaine de fréquence qui nous intéresse ici.

À l'aide des jeux de données simulés et de ces réponses en fréquence individuelles de capteurs, on peut alors tester les performances de différents algorithmes de localisation de sources, calculés/entraînés pour des données issues de capteurs *idéaux* (ce qui correspond à une compensation parfaite des réponses individuelles de capteurs réels) ou pour des données issues de capteurs *réels*, sans compensation de leurs réponses individuelles.

3. lorsqu'elle est basée sur des jeux données expérimentales



(a) Module



(b) Phase

FIGURE 5.4 – Réponse en fréquence (FRF) mesurées de microphones 1/4" ICP à électret, produits par le CTTM et utilisés au laboratoire.

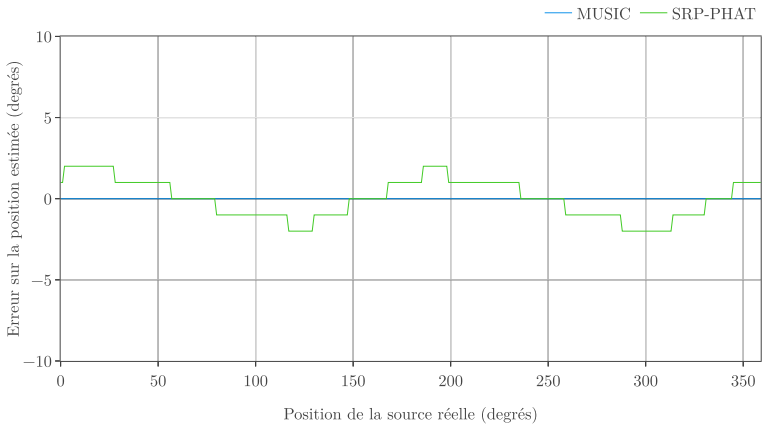
Les trois algorithmes qui sont testés ici avec ces deux jeux de données sont :

— l'algorithme MUSIC, présenté en section 1.1.4,

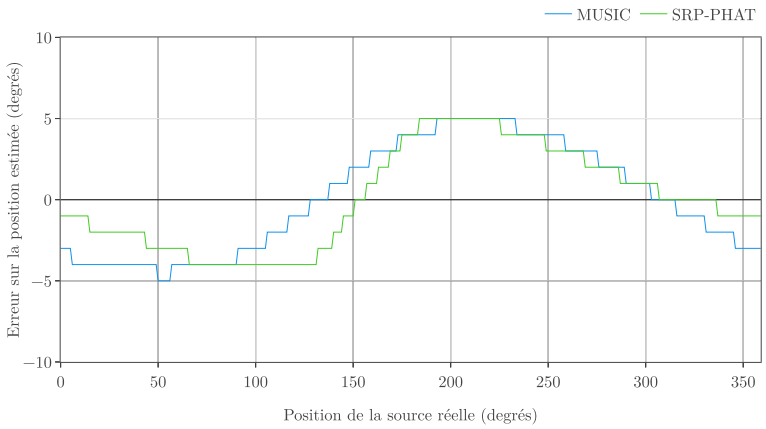
— l'algorithme SRP-PHAT, présenté en section 1.1.4,

— le réseau de neurones associé à l'approche BeamLearning, détaillé au chapitre 5 de cette thèse.

Les performances de localisation obtenues dans toutes ces situations sont présentées sur les figures 5.5 et 5.6. La figure 5.5 a été obtenue en appliquant successivement les méthodes MUSIC et SRP-PHAT à des mesures en champ libre du champ émis par 360 positions de sources réparties uniformément sur un cercle à une distance de 10 mètres du centre de l'antenne. Pour cette expérience, c'est l'antenne circulaire à 8 microphones décrite en section 3.2.1 qui a été utilisée. Les signaux émis par les sources sont non bruités et correspondent à une somme de sinus purs aux fréquences [125, 250, 500, 1 000, 2 000,4 000] Hz.



(a) Performances des algorithmes MUSIC et SRP-PHAT avec des microphones *idéaux*



(b) Performances des algorithmes MUSIC et SRP-PHAT sans compensation des FRF *réelles* des microphones.

FIGURE 5.5 – Comparaison des algorithmes MUSIC et SRP-PHAT : (a) dans le cas de microphones *idéaux* – (b) dans le cas de signaux signaux issus de microphones sans compensation des FRF *réelles*.

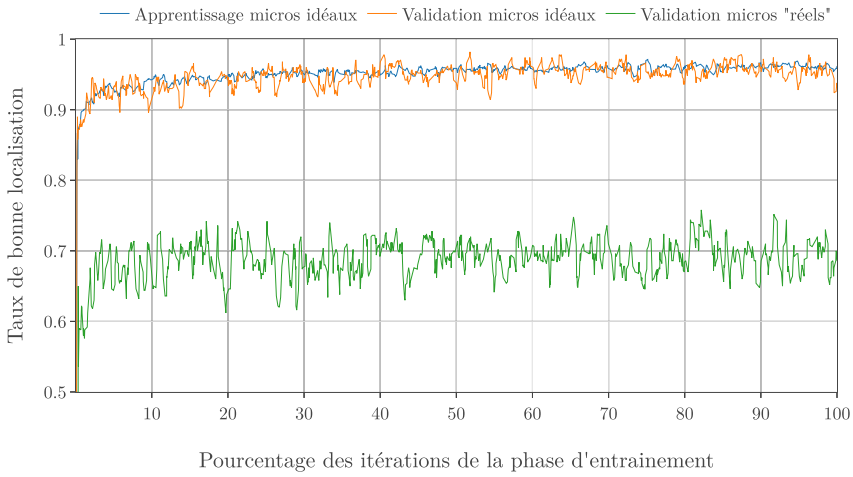
Comme attendu, on observe sur la figure 5.5(a) que dans une situation aussi idéale, les algorithmes MUSIC et SRP-PHAT offrent des excellentes performances de localisation avec des microphones *idéaux* (c'est à dire en compensant parfaitement les réponses individuelles en phase et en amplitude de chacun des capteurs). En revanche, avec des mesures par des microphones *réels* sans compensation des réponses individuelles des capteurs, les performances des deux algorithmes se trouvent dégradées, même si on est ici dans une situation où les capteurs sont issus d'un même lot de production, de technologie identique, et de réponses en amplitude et en phase respectant les tolérances standards.

La figure 5.6 représente, quant à elle, les courbes de convergence d'apprentissage du réseau et leur validation sur un jeu disjoint du jeu de données d'apprentissage, pour deux situations d'entraînement qui reproduisent les conditions utilisées pour la figure 5.5 (mesures en champ libre, sans bruit de fond, par une antenne circulaire à 8 microphones) :

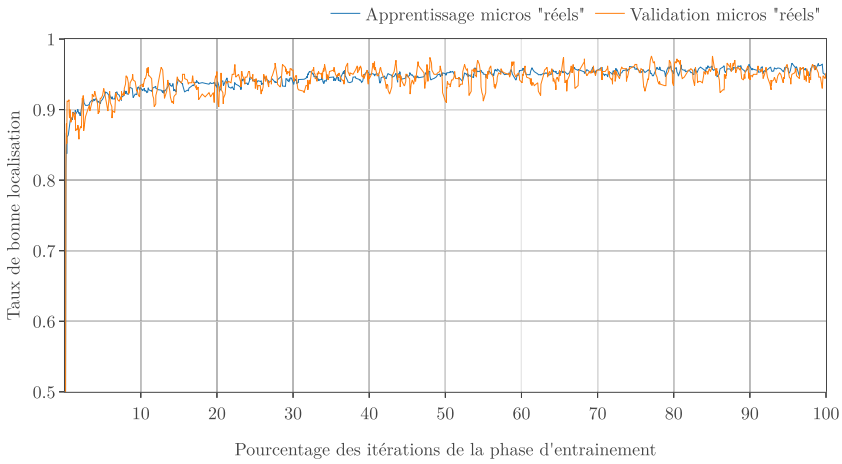
— la figure 5.6(a) correspond à un entraînement avec des mesures par des microphones *idéaux* (c'est à dire en compensant parfaitement les réponses individuelles en phase et en amplitude de chacun des capteurs),
— la figure 5.6(b) correspond à un entraînement avec des mesures par des microphones avec des mesures par des microphones *réels*, sans compensation des réponses individuelles des capteurs.

En premier lieu, en comparant les performances obtenues sur la convergence de l'apprentissage sur le jeu de données d'entraînement (courbes bleues), on constate qu'il n'y a aucune différence significative de performances entre ces deux situations : le réseau converge vers la même précision, avec la même vitesse de convergence pour les figures 5.6(a) et 5.6(b), ce qui signifie que la méconnaissance de la réponse individuelle des capteurs n'est absolument pas un obstacle à la tâche de localisation par apprentissage supervisé, contrairement aux approches modèles.

Sur la figure 5.6(a), on observe également que lorsqu'on a entraîné le réseau avec des données issues de capteurs *idéaux* et qu'on réalise une validation sur des jeux de données disjoints n'ayant pas servi à l'entraînement provenant de capteurs dont on a compensé les réponses individuelles, les performances en validation restent excellentes (courbe orange, 5.6(a)). En revanche, tout comme pour les approches



(a) Courbe d'apprentissage de l'approche BeamLearning avec des microphones *idéaux* ; Test avec des microphones *idéaux* ou avec des microphones sans compensation des FRF *réelles*.



(b) Courbe d'apprentissage de l'approche BeamLearning avec des microphones sans compensation des FRF *réelles*.

FIGURE 5.6 – Comparaison des performances d'apprentissage de l'approche BeamLearning : (a) dans le cas de microphones *idéaux* – (b) dans le cas de signaux signaux issus de microphones sans compensation des FRF *réelles*.

modèles, si on teste les performances sur un jeu de données issues de capteurs *réels* sans prise en compte des réponses individuelles, les performances sont dégradées (courbe verte, 5.6(a)).

Par opposition, on observe que lorsque le réseau est entraîné à l'aide de données issues de capteurs *réels* sans compensation de la réponse en fréquence des capteurs, la validation réalisée sur des jeux

de données disjoints n'ayant pas servi à l'entraînement provenant de capteurs sans compensation des réponses en fréquences, les performances en validation restent excellentes : le réseau a ainsi appris à corriger les réponses individuelles des capteurs tout en apprenant à localiser les sources, sans procédure supplémentaire d'étalonnage. Ce point permet de mettre en avant l'un des avantages primordiaux de l'apprentissage sur des bases de données expérimentales. Par ailleurs, même si ce point n'est pas mis en évidence par les expériences décrites ici, la compensation offerte par le réseau ne se restreint pas aux courbes de réponses en fréquence et en amplitude, mais elle concerne également la directivité des microphones de l'antenne, ainsi que la diffraction par le corps de l'antenne. Ce point sera validé expérimentalement à la section 5.3.6 avec une antenne sphérique à corps rigide.

Afin d'être parfaitement complet, et pour comparer les performances de localisation pour le problème de classification en secteurs angulaires pour les différentes approches, les résultats obtenus par les approches modèles MUSIC et SRP-PHAT présentées sur la figure 5.5 sont utilisées pour déterminer le taux de bonne localisation dans 8 secteurs angulaires. Le tableau 5.3 récapitule ces différents résultats et permet ainsi de mettre en exergue la propriété d'étalonnage implicite offerte par l'approche BeamLearning : tandis que les méthodes modèles présentent une dégradation de 5 à 6% lorsque les réponses individuelles des capteurs de l'antenne ne sont pas connues par rapport à une situation où les réponses individuelles sont connues, l'approche BeamLearning ne présente qu'une dégradation de 0.2% de performances, qui n'est pas statistiquement représentative.

Méthode de localisation	Capteurs idéaux (compensation parfaite des réponses individuelles)	Capteurs réels (sans compensation des réponses individuelles)
MUSIC	0%	6,66%
SRP-PHAT	1,66%	6,11%
BeamLearning	3,6%	3,8%

Tableau 5.3 – Pourcentage d'erreur de classification de 360 sources en 8 classes, pour des approches modèle (MUSIC, SRP-PHAT) et BeamLearning, à partir de sommes de sinus purs aux fréquences [125, 250, 500, 1000, 2000,4000] Hz ($RSB = +\infty$).

Il est essentiel de noter que pour mettre en évidence ces comportements, nous avons ici exploité des données simulées, auxquelles nous avons adjoint (ou non) des étalonnages de capteurs réels, afin de mettre en évidence de manière didactique la propriété d'étalonnage implicite offerte par une approche d'apprentissage. Bien entendu, en pratique, cette propriété prend tout son sens lorsque la base de données est constituée à partir de données expérimentales (voir chapitre 4). La sous-section suivante permet d'exploiter ce cadre, puisqu'elle a pour objectif d'analyser les performances d'apprentissage pour une détermination expérimentale de DOA par classification angulaire avec une antenne *réelle* non étalonnée, dans un environnement partiellement traité acoustiquement, avec une paroi parfaitement réfléchissante.

5.1.5 Détermination expérimentale de DOA 2D par classification, dans une salle partiellement traitée acoustiquement

À la suite de ces résultats encourageants obtenus à partir de simulations numériques pour la détermination de DOA par classification angulaire, le même processus a été testé pour des données mesurées à l'aide d'une antenne réelle. Ici, le jeu de données d'apprentissage a été constitué en appliquant le protocole décrit au chapitre 4, en plaçant l'antenne Mini DSP présentée en section 4.2.2 au centre du spatialisateur 3D SpherBedev. Le fait d'entraîner le réseau de neurones profond à l'aide de ces données mesurées permet ainsi de tester immédiatement l'optimisation des variables d'apprentissages pour des mesures réalisées par des capteurs non étalonnés, et permet de mettre en pratique le résultat prouvé à la section précédente (5.1.4).

Pour cette validation expérimentale, il est essentiel de noter que l'environnement de mesure est plus complexe que dans les cas précédemment présentés. Au cours de la première année de mon doctorat, la salle dans laquelle est installée le spatialisateur 3D n'était que partiellement traitée acoustiquement sur ses parois latérales et sur le sol. Le plafond de la salle, quant à lui, est une paroi de béton brut plane, qui s'approche d'une paroi parfaitement réfléchissante. On est donc ici dans une situation qui s'approche plus d'une mesure en salle semi-anéchoïque *dégradée* (le traitement acoustique des parois n'est pas aussi performant que celui d'une salle semi-anéchoïque et seule une moquette est présente au sol), que d'une mesure en champ libre. Sur ce sujet, une confrontation de résultats obtenus à partir de simulations et d'expériences pour un problème de détermination de DOA par une approche de

régression sera présentée en section 5.2.2.

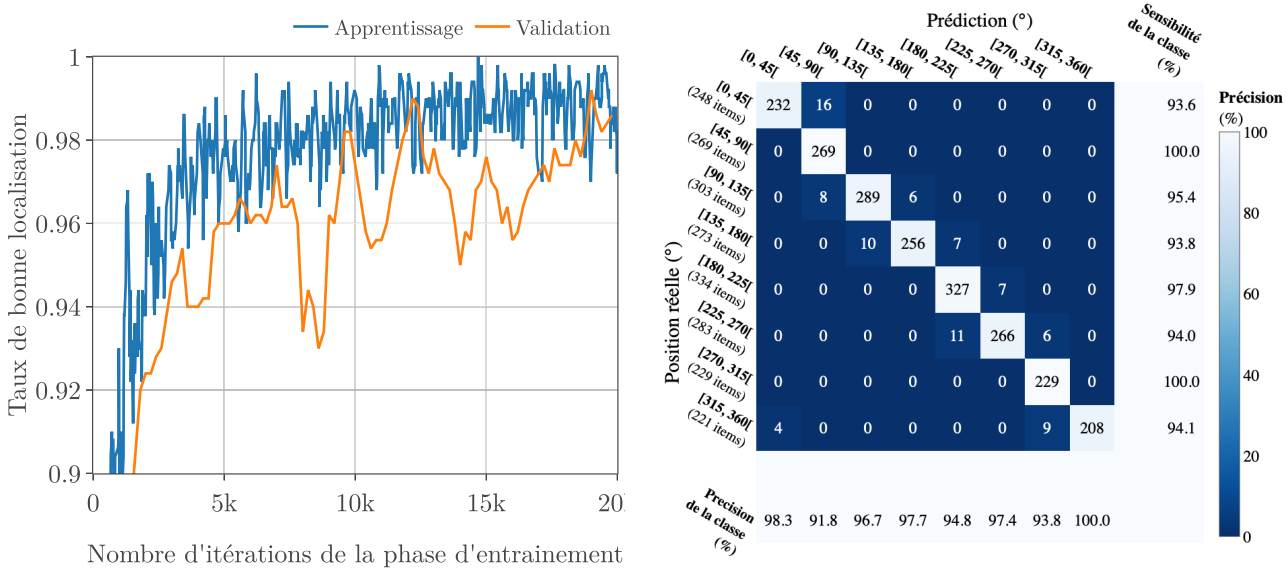
	Base de données	Environnement	Signal	Antenne	RSB
Données	Expérimentale	Salle partiellement traitée,une paroi parfaitement réfléchissante	Sinus pur	Mini DSP.	> 20 dB
	Sortie	Nombre de classes	Précision	Nb. itérations	Sur apprentissage
Résultats	Classification	8	98%	20 000	Léger

Tableau 5.4 – Récapitulatif synthétique des paramètres pour l'apprentissage présenté en section 5.1.5

Pour la détermination de DOA expérimentale par une approche de classification qui nous intéresse ici, avec antenne réelle à bas coût, sans étalonnage préalable, toutes les caractéristiques de l'antenne et des capteurs MEMS la composant sont incluses dans les signaux mesurés pour la constitution de la base de données d'apprentissage, depuis la diffraction du corps de l'antenne, jusqu'à la réponse en fréquence de chaque microphone. Au cours de l'entraînement du réseau, l'optimisation des paramètres des couches neuronales permet alors de résoudre le problème de localisation, tout en compensant les caractéristiques propre de l'antenne utilisée lors des enregistrements.

La figure 5.7 présente les résultats obtenus pour cet apprentissage, et l'analyse de la figure 5.7(a) révèle qu'avec ces données expérimentales en environnement partiellement traité et en présence d'une paroi parfaitement réfléchissante, la précision de localisation atteint une valeur supérieure à 98% sur la base de données d'entraînement à l'issue des 20 000 itérations d'entraînement. Les taux de bonnes localisations sont légèrement dégradés pour la base de données de validation, disjointe de la précédente, et non utilisée pour l'optimisation des couches neuronales au cours de l'apprentissage. Par ailleurs, on constate sur la matrice de confusion présentée à la figure 5.7(b) que la précision et la sensibilité restent satisfaisantes pour chacun des 8 cadrans angulaires, même si les performances sont dégradées par rapport à une situation de champ libre, et que l'entraînement nécessite un nombre d'itérations globalement plus grand que pour les situations numériques précédentes. Par ailleurs, en comparant les

courbes de convergence d'apprentissage obtenues en section 5.1.2 à celles obtenues ici (figure 5.7(a)), on peut remarquer que pour des données expérimentales en environnement traité partiellement et en présence d'un paroi réfléchissante, les performances obtenues sur le jeu de données de validation au cours de l'entraînement se superposent moins bien avec celles obtenues sur le jeu de données d'apprentissage, et que les performances de validation se stabilisent elles aussi après un plus grand nombre d'itérations d'entraînement. Ces phénomènes indiquent que la généralisation des résultats à partir des données d'apprentissage est plus difficile à réaliser pour le réseau qu'en champ libre.



(a) Courbe d'apprentissage de BeamLearning (b) Matrice de confusion à partir de données de test

FIGURE 5.7 – Performances de l'approche BeamLearning dans le cas de la classification de données mesurées monochromatiques dans une salle traitée acoustiquement avec un plafond réfléchissant et un $RSB \geq 20$ dB. (a) : Courbe de convergence d'apprentissage du réseau, obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement. (b) Matrice de confusion obtenue sur l'ensemble du jeu de données de validation, pour la dernière itération de l'apprentissage.

En revanche, l'analyse de la matrice de confusion sur la figure 5.7(b) révèle que les erreurs de prédiction de secteurs angulaires concernent exclusivement des secteurs angulaires contigus, puisque celles-ci sont toutes situées sur les sur-diagonales et les sous-diagonales de la matrice. En pratique, ces erreurs concernent des sources situées dans l'immédiate proximité de la frontière séparant deux classes. L'erreur commise est donc en réalité faible d'un point de vue angulaire. Par exemple, si

une source est située à un azimut $\theta = 134°$, il est possible qu'elle fasse partie des 6 points classés à tort dans le secteur angulaire [135, 180[. Dans ce cas précis, à 1 degré près, la source aurait été classée dans le bon secteur angulaire. Cet exemple pointe une limite fondamentale de l'approche de classification pour résoudre le problème de localisation, qui ne peut qu'être encore plus importante avec une augmentation du nombre de classes. En effet, plus le nombre de classes augmente, plus le nombre de frontières entre classes augmente, et avec lui le nombre de cas *litigieux*. L'approche par régression, qui permet d'obtenir une valeur unique plutôt qu'un secteur angulaire, apparaît alors plus appropriée pour réaliser une localisation angulaire précise. Ce type d'approche sera donc exploitée dans toute la suite du manuscrit.

5.2 Détermination de DOA 2D par une approche de régression

Tout comme dans la section précédente, cette section a pour objectif d'analyser les performances de localisation, en explorant de manière didactique l'influence de plusieurs paramètres liés aux bases de données, afin de fournir des éléments d'analyse pertinents permettant de comprendre le comportement du réseau. Ainsi, la complexité du problème de localisation sera itérativement augmentée, jusqu'à proposer une approche de détermination de DOA 2D en environnement bruité et réverbérant. Les analyses fournies seront issues à la fois de simulations numériques et de données expérimentales, et la méthode proposée sera confrontée à des algorithmes de localisation de sources performants basés sur des approches *modèles*.

5.2.1 Localisation en champ libre, avec bruit de mesure, pour des sources monochromatiques

Dans cette section, l'approche de localisation par régression est illustrée à partir d'un problème simple, en environnement de type champ libre, avec la présence de bruit de mesure sur les capteurs. Comme pour les sections précédentes, les signaux émis par les sources à localiser sont des signaux monochromatiques, aux fréquences centrales de bandes d'octaves entre 125 Hz et 4 000 Hz. Comme dans la section 5.1.2, un bruit de mesure d'amplitude aléatoire est ajouté aux signaux des capteurs microphoniques, limité à un RSB supérieur à 20 dB. Un résumé synthétique des caractéristiques de la base de données d'apprentissage et des performances obtenues est fourni dans le tableau 5.5.

L'approche d'estimation de DOA 2D par régression utilisée ici offre la possibilité d'analyser la

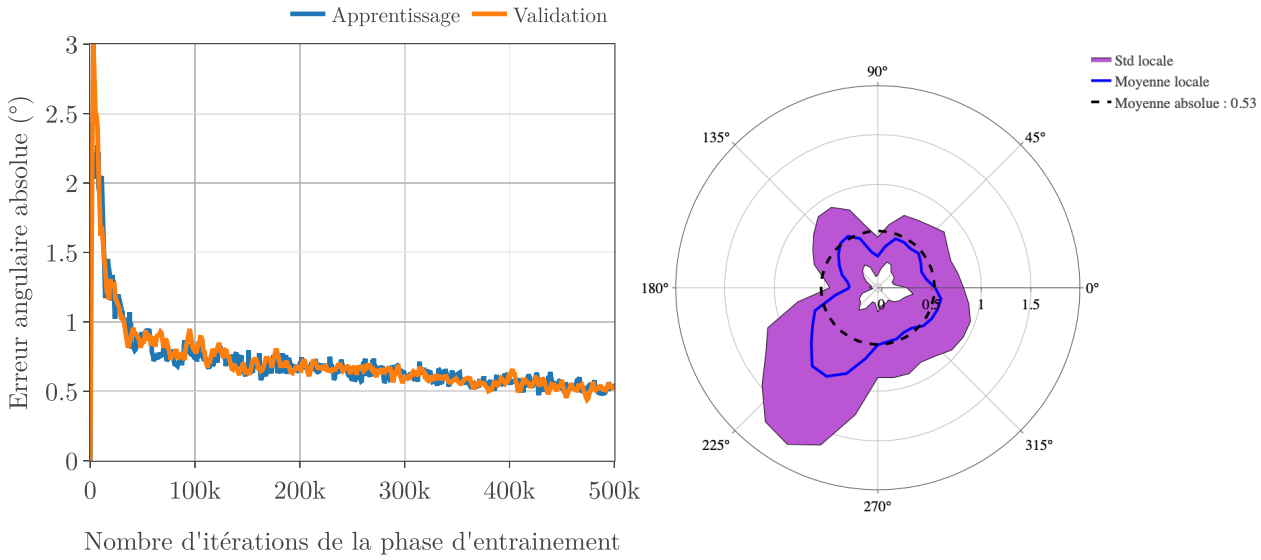
Données	Base de données	Environnement	Signal	Antenne	RSB
	Simulée	Champ libre	Sinus pur	Mini DSP.	> 20 dB
Résultats	Sortie	Angle estimé	Précision	Nb. itérations	Sur apprentissage
	Régression	azimut à 360°	$0,5°$	500 000	Non

Tableau 5.5 – Récapitulatif synthétique des paramètres pour l'apprentissage présenté en section 5.2.1

convergence du réseau au cours de son entraînement, non plus en termes de taux de bonne localisation dans des secteurs angulaires, mais en termes d'erreur angulaire absolue entre la position $\tilde{\theta}$ estimée par le réseau et la position θ réelle de la source. Cette grandeur permet immédiatement d'identifier l'erreur angulaire commise par le réseau au cours de son entrainement, comme sur la figure 5.8(a), mais également d'analyser statistiquement ces erreurs en fonction de la position des sources, comme sur la figure 5.8(b).

L'analyse de ces figures démontre, ici encore, que pour un cas simple de localisation en champ libre avec bruit de mesure, la méthode proposée offre des excellentes performances de localisation, puisque l'erreur absolue moyenne à l'issue de 500 000 itérations d'entraînement, atteint environ $0,5°$, et que seules quelques positions de sources situées autour de 225° possèdent une erreur d'estimation de l'ordre de 1°. En revanche, pour l'approche de régression, la convergence du réseau est plus lente que pour la classification, d'un facteur 25 environ, avec 4 jours environ d'entraînement sur un GPU 1080Ti pour atteindre les 500 000 itérations. Toutefois, sans attendre la fin de l'apprentissage, la précision est très vite satisfaisante, puisque les résultats restent, d'après la figure 5.8(a), en dessous de 2° d'erreur à partir de 6 000 itérations (soit 2h de calcul sur notre station de calcul équipée de cartes graphiques Nvidia 1080 Ti).

Puisque les données d'entraînement sont restreintes à des signaux monochromatiques pour un faible nombre de fréquences, par analogie avec la discussion menée à la section 5.1.3, il est également intéressant de ne pas se limiter à l'analyse des performances du réseau pour un jeu de données de test à ces fréquences. En effet, pour le problème de la classification, l'analyse de la directivité des neurones de sorties pour des fréquences s'écartant de celles utilisées pour entraîner le réseau a révélé



(a) Courbe d'apprentissage de BeamLearning

(b) Précision angulaire moyenne et écart type

FIGURE 5.8 – Performances de l'approche BeamLearning dans le cas de localisation par régression de données simulées monochromatiques en champ libre avec un $RSB \geq 20$ dB. (a) : Courbe de convergence d'apprentissage du réseau, obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement. (b) Représentation polaire statistique des erreurs angulaires absolues, obtenues à l'issue de la dernière itération d'entraînement, sur un jeu de données test correspondant à 4 800 sources réparties uniformément autour de l'antenne.

que les performances se dégradaient, puisque la base de données n'offrait pas une diversité suffisante en fréquence pour rester performante sur tout le domaine de fréquence. Ici, un constat similaire peut être réalisé. À cet effet, et pour illustrer le phénomène, la figure 5.9 présente l'erreur moyenne absolue obtenue sur 360°, pour des fréquences allant de 100 à 4 000 Hz par pas de 100 Hz. Ainsi sur cette représentation, les fréquences 125 Hz, et 250 Hz qui étaient représentées dans le jeu de données d'entraînement ne sont pas représentées sur la figure, et un grand nombre de sources test présentent volontairement un contenu fréquentiel différent de celui du jeu de données d'entraînement. L'analyse de la figure 5.8 révèle que seuls les sources aux fréquences ne s'écartant pas de plus de 200 Hz des fréquences utilisées pour le jeu de données d'entraînement, présentent des erreurs angulaires inférieures à 2°, et que seules les valeurs exactes des fréquences d'entraînement atteignent une erreur angulaire absolue de $0,5°$. En dehors de ces domaines fréquentiels, les performances de localisation angulaire par régression sont particulièrement dégradées, ce qui rejoint les observations réalisées pour le pro-

blème de classification. À titre d'exemple, les erreurs angulaires absolues peuvent atteindre 90° pour la fréquence de 2 500 Hz. Cette valeur n'est d'ailleurs pas anodine car elle correspond à l'espérance de l'erreur obtenue en prenant une position estimée aléatoirement. Il est donc clair que malgré les très bonnes performances obtenues pour des fréquences proches des fréquences utilisées pour le jeu de données d'entraînement, l'algorithme n'est pas suffisamment robuste en fréquence. Ce problème est strictement le même qu'avec une approche de classification par secteurs angulaires (5.1.3), mais il était alors moins prononcé. Ici, les mêmes causes entraînant les mêmes effets, ce n'est pas le réseau en tant que tel qui est responsable de cette dégradation, mais le jeu de données utilisé pour l'entraînement, ce qui révèle l'importance de construire un jeu de données présentant une diversité importante de contenus fréquentiels, comme ce sera réalisé à partir de la section 5.2.3.

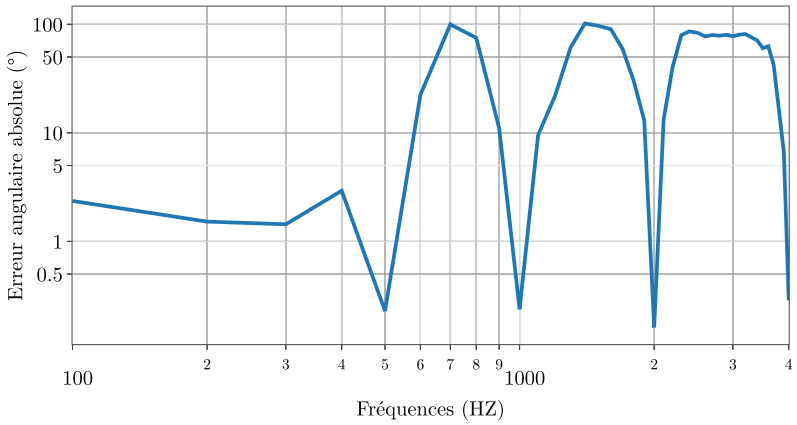


FIGURE 5.9 – Erreur angulaire absolue moyenne de localisation, lorsque le jeu de données d'apprentissage n'est constitué que de données monochromatiques aux fréquences centrales des bandes d'octaves de 125 Hz à 4 000 Hz, pour des sources de validation émettant un signal monochromatique de fréquence allant de 100 à 4 000 Hz, par pas de 100 Hz

5.2.2 Comparaison des performances de localisation obtenues en présence d'une paroi parfaitement réfléchissante, à partir de données simulées et de données mesurées

Dans le cas de la classification, l'analyse des performances de localisation obtenues expérimentalement à la section 5.1.5 a révélé qu'une légère baisse de la précision de détermination de DOA pouvait être liée à l'environnement plus complexe dans lequel l'expérience a été réalisée. Afin de confirmer cette hypothèse, un jeu de données issues de simulations numériques est constitué à partir d'un environ-

nement semi-anéchoïque avec un sol parfaitement réfléchissant, puisque l'environnement dans lequel les mesures en 5.1.5 ont été réalisées est proche de cette situation. Le réseau de neurones profond est alors entraîné suivant les mêmes conditions qu'en 5.2.1 à partir de ce jeu de données avec influence du sol réfléchissant, et les performances de localisation sont comparées avec celles obtenues à l'aide d'un entraînement du réseau fait à partir de données simulées de champ libre d'une part, et d'un apprentissage à partir des données expérimentales d'autre part, présentées en section 5.1.5. Dans ces trois cas, la même antenne (Mini DSP) est utilisée, physiquement ou numériquement, et les signaux émis sont des sinus purs aux fréquences [125, 250, 500, 1000, 2000,4000] Hz.

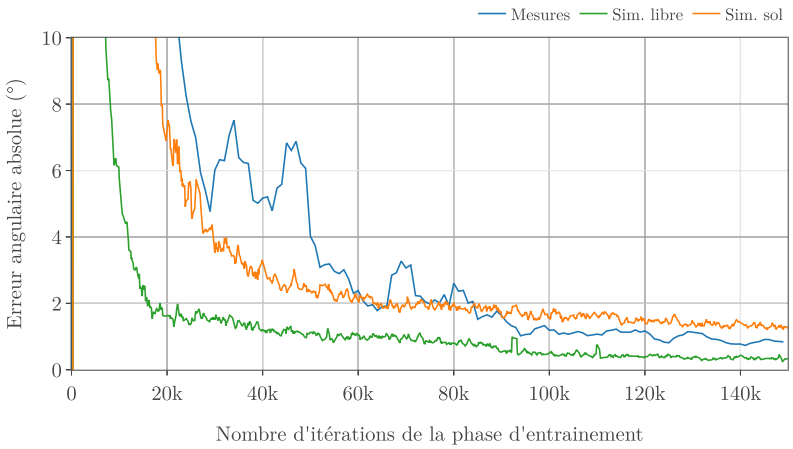


FIGURE 5.10 – Performances de l'approche BeamLearning pour des données : simulées en champ libre (vert), simulées en espace semi-infini avec un sol parfaitement réfléchissant (orange), expérimentales dans une salle où le sol et les murs sont traités acoustiquement, mais le plafond est parfaitement réfléchissant (bleu). Fréquences utilisées : [125, 250, 500, 1000, 2000,4000] Hz

La figure 5.10 synthétise les performances obtenues dans ces trois cas. Il apparaît bien qu'en simulation, le fait d'avoir un sol réfléchissant rend l'apprentissage moins aisé, même si les performances de localisation sont toujours très satisfaisantes : les performances de localisation se stabilisent autour de 1° d'erreur dans le cas du sol réfléchissant, contre $0,5°$ dans le cas de signaux simulés en champ libre. Dans le cas de l'apprentissage effectué sur des données mesurées, les performances se réduisent légèrement par rapport à celles observées sur un apprentissage dans le cas d'un sol réfléchissant. Cette diminution n'est pas liée à la *qualité* des données, ni au fait que ces données soient mesurées à partir de capteurs non étalonnés, puisqu'il a été prouvé en section 5.1.4 et 5.1.5, que les propriétés d'étalonnage

implicite offertes par l'approche BeamLearning permettaient de ne pas altérer les performances de localisation. En revanche, contrairement aux simulations, les murs et le sol ne sont pas parfaitement absorbants, puisque la salle dans laquelle est installée le spatialisateur 3D n'est que partiellement traitée acoustiquement. On peut donc raisonnablement suspecter que ce sont les réflexions des sources primaires sur les murs qui rendent le problème de localisation plus difficile à résoudre. Cette tendance justifie le fait que ce travail de thèse de doctorat se soit ensuite orienté vers la localisation de sources acoustiques en environnement réverbérant (à partir de la section 5.2.5).

En revanche, malgré ce léger écart de performances, les caractéristiques de la convergence de l'apprentissage à partir de données simulées ou expérimentales sont équivalentes. Cette constatation permet donc de valider le fait qu'il est pertinent dans notre cas d'étudier les tendances et les caractéristiques de la méthode BeamLearning à partir de jeux de données obtenus par simulations numériques, et d'extrapoler ces résultats pour des données mesurées, tout en prenant le soin régulièrement de les valider expérimentalement, comme ce sera le cas dans les prochaines sections de ce manuscrit.

5.2.3 Augmentation de la robustesse dans le domaine fréquentiel visé

Comme exposé en section 5.2.1, pour obtenir un comportement homogène des performances de détermination de DOA grâce à l'approche BeamLearning, il est nécessaire que les jeux de données d'entraînement possèdent une variabilité de contenu fréquentiel suffisante. Dans la suite du manuscrit, les jeux de données seront constitués non plus de champs émis par des sources monochromatiques, mais seront issus de sources émettant des signaux beaucoup plus variés. Puisque les jeux de données monochromatiques ayant servi aux analyses des sections précédentes étaient constitués de six signaux différents, et même si cela n'est pas indispensable, nous avons choisi ici de conserver le même nombre de signaux différents (voir tableau 3.3 pour référence). Pour rappel, ces signaux sont :

— un enregistrement de type *Cocktail party*, constitué de voix féminines,
— un bruit de klaxon, pour conserver des exemples dans la base de données qui aient un comportement très tonal,
— quatre extraits d'enregistrements de musique symphonique, afin d'obtenir des exemples présentant une densité spectrale très variée et une dynamique en amplitude importante.

À partir de ce jeu de données, qu'on appellera *multi signaux*, et en s'appuyant sur le même type de procédure d'entraînement que précédemment avec les paramètres récapitulés au tableau 5.6, les performances de détermination de DOA 2D par une approche de régression sont tracées sur la figure 5.11.

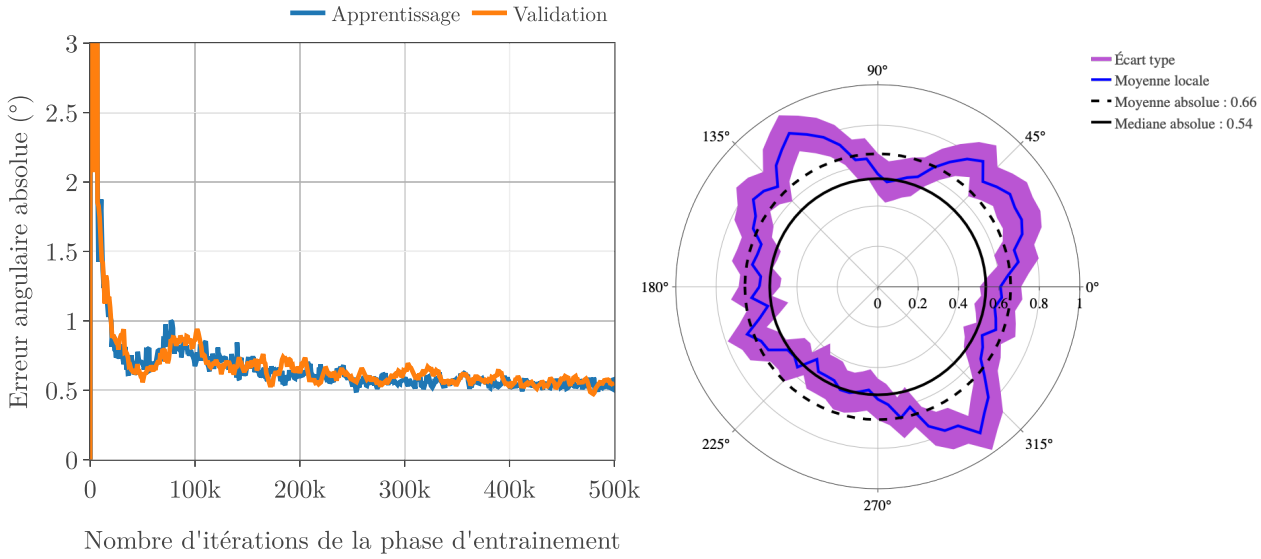
Données	Base de données	Environnement	Signal	Antenne	RSB
	Simulée	Champ libre	Multi signaux	Mini DSP.	> 20 dB
Résultats	Sortie	Angle estimé	Précision	Nb. itérations	Sur apprentissage
	Régression	azimut à 360°	$0,5°$	500 000	Non

Tableau 5.6 – Récapitulatif synthétique des paramètres pour l'apprentissage présenté en section 5.2.3

L'analyse des résultats sur les figures 5.11(a) et 5.11(b) permet de conclure qu'avec ce jeu de données plus varié en termes de contenu fréquentiel, les performances de détermination de DOA obtenues restent quasiment identiques à celles obtenues sur des signaux mono-fréquentiels à la section 5.2.1. L'erreur absolue moyenne de localisation de $0,56°$ obtenue après 500 000 itérations d'entraînement du réseau est une excellente performance.

La figure 5.11(b) présente plus spécifiquement les résultats de l'approche BeamLearning, après apprentissage, sur un ensemble de 4 800 exemples de test disjoints de ceux ayant servi à l'entrainement du réseau. L'analyse de la précision angulaire prouve que, même si quelques directions présentent une erreur moyenne locale de l'erreur angulaire absolue plus élevée que les autres directions, ces erreurs restent systématiquement contenues en dessous de 1° d'erreur, et ce, pour toutes les valeurs d'azimut. Ces moyennes locales sont calculées pour des secteurs angulaires de 5°, ce qui correspond à 110 sources en moyenne par secteur. Enfin, la dispersion de l'erreur, calculée sur les 4 800 exemples de test pour une meilleure représentativité statistique, est tracée en violet autour de la courbe de l'erreur moyenne locale. Cette dispersion, d'environ 1°, est relativement resserrée, ce qui met en évidence le fait que les erreurs sont statistiquement bien concentrées autour de la moyenne. Toutefois, quelques grosses erreurs d'estimations subsistent, jusqu'à 100° sur certains signaux, mais ces exemples sont statistiquement non

représentatifs puisque l'analyse aux grands nombres réalisée ici démontre qu'ils n'impactent que très peu l'écart-type de l'erreur d'estimation angulaire.



(a) Courbe d'apprentissage de BeamLearning

(b) Précision angulaire moyenne et écart type

FIGURE 5.11 – Performances de l'approche BeamLearning dans le cas de localisation par régression : données simulées à partir de différents signaux tels que du bruit de *cocktail party*, un klaxon ou de la musique classique, en champ libre avec un $RSB \geq 20$ dB. (a) : Courbe de convergence d'apprentissage du réseau obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement. (b) Représentation polaire statistique des erreurs angulaires absolues, obtenues à l'issue de la dernière itération d'entraînement, sur un jeu de données test correspondant à 4 800 sources réparties uniformément autour de l'antenne.

Par analogie avec l'analyse ayant mené au tracé de la figure 5.9, et afin de vérifier que l'utilisation d'un jeu de données d'entraînement constitué de champ de pression issus de signaux au contenu fréquentiel plus varié et plus large bande permet d'obtenir de meilleures performances de localisation dans l'ensemble du domaine fréquentiel, l'erreur moyenne angulaire est tracée de nouveau pour des captations de champ monochromatiques aux fréquences allant de 100 à 4 000 Hz, par pas de 100 Hz. La figure 5.12 récapitulant ces résultats, prouve que conformément aux attentes, l'erreur de localisation est beaucoup plus homogène que sur la figure 5.9 sur une grande plage fréquentielle : elle reste contenue entre 1 et 2 degrés entre 200 Hz et 2 000 Hz environ, sans pertes de performances dans ce domaine de fréquence. En dehors de cette plage fréquentielle, les résultats se dégradent, en particulier en haute fréquence. Cette dégradation s'explique aisément *a posteriori* par le manque de contenu spectral dans

les exemples du jeu de données constitué au delà de de la fréquence de 2 500 Hz dans la base de données d'apprentissage. En effet, ces fréquences, même si elles peuvent être présentes sous forme d'harmoniques, ne représentent qu'une faible proportion de l'énergie des signaux utilisés ici. Sur le même principe que ce qui a été illustré ici, il suffit donc de constituer des jeux de données d'entrée au contenu plus *présent* au delà de 2 500 Hz pour obtenir une robustesse de localisation sur tout le domaine de fréquences.

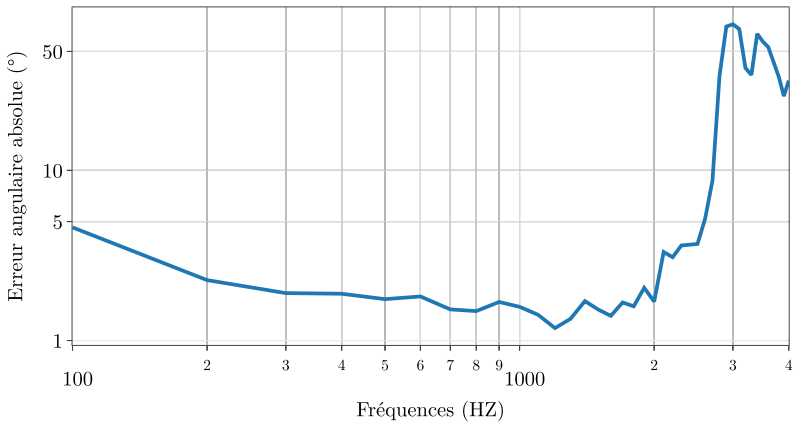


FIGURE 5.12 – Erreur angulaire absolue moyenne de localisation, sur des sinus purs allant de 100 à 4 000 Hz par pas de 100 Hz, lorsque le jeux de données d'entraînement est multi-signaux

5.2.4 Comparaison des performances de l'approche BeamLearning avec les algorithmes MUSIC et SRP-PHAT en champ libre

La variabilité de signaux constituant le jeu de données d'entraînement ayant permis d'obtenir une localisation robuste dans le domaine fréquentiel visé, l'objectif de l'analyse menée dans cette section est d'analyser les résultats obtenus grâce au réseau de neurones entraîné grâce aux paramètres de la section 5.2.3, avec les résultats obtenus grâce à des approches de localisation des sources plus conventionnelles reposant sur une approche de type *modèle*. Pour cette comparaison, nous proposons ici d'utiliser les algorithmes MUSIC [5] et SRP-PHAT [11], reconnus dans la communauté scientifique pour leur bonnes performances de localisation et leur robustesse au bruit de mesure. Pour ce faire, les trois méthodes sont testées en réalisant une tâche de localisation de sources avec un RSB allant de -1 dB à 40 dB. Pour rappel, le réseau a ici été entraîné à l'aide d'un jeu de données multi-signaux, avec un RSB d'entraînement supérieur à 20 dB. La figure 5.13 permet de comparer les résultats obtenus,

pour des sources émettant des signaux de type *Cocktail party*, n'ayant jamais été présentées au réseau BeamLearning lors de sa phase d'entraînement.

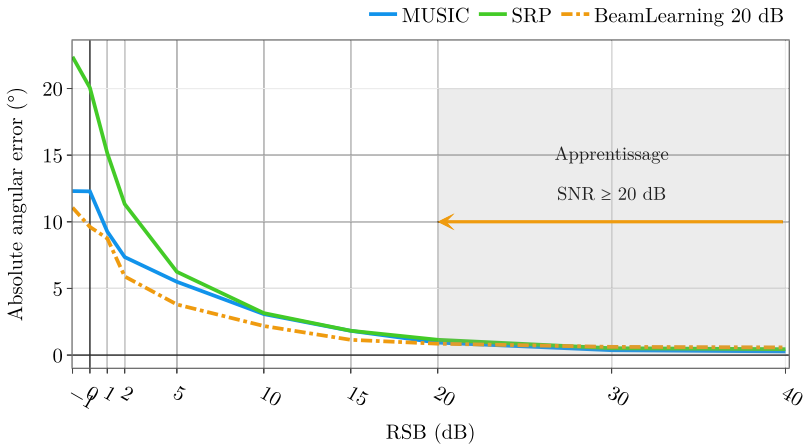


FIGURE 5.13 – Comparaison entre les algorithmes MUSIC, SRP-PHAT, et le réseau de neurones profond proposé, entrainé sur un jeu de données multi-signaux en champ libre, avec un RSB supérieur à 20 dB.

L'analyse de la figure 5.13 permet de mettre en exergue deux grandes tendances. En premier lieu, comme attendu, quelque soit la méthode utilisée, plus le RSB diminue, plus l'erreur de localisation est importante. Même si le réseau a été entraîné avec un RSB supérieur à 20 dB pour le rendre robuste au bruit de mesure, ce phénomène est bien entendu également observé pour l'approche par apprentissage, qui nécessite, comme les approches *modèles*, une bonne qualité d'enregistrement pour pouvoir localiser une source dans le plan. En revanche, on peut observer que les performances obtenues sont globalement équivalentes entre l'approche apprentissage et les deux approches modèles proposées, même si l'approche BeamLearning offre une meilleure robustesse à très faible RSB.

Quand le RSB est supérieur à 30 dB, donc dans des cas de mesures très favorables, les algorithmes qui utilisent des modèles sont très légèrement meilleurs (de 1 ou 2 dixième de degré, ce qui n'est pas particulièrement limitant). La tendance s'inverse pour des RSB entre -1 dB et 30 dB : même si les algorithmes MUSIC et SRP-PHAT sont conçus pour rester robustes au bruit de mesure, l'approche BeamLearning permet de localiser les sources avec une meilleure précision que les approches modèles, y compris dans des cas de mesures très défavorables avec des caractéristiques de rapport signal à bruit

auxquelles le réseau n'a pas été entraîné.

Par ailleurs, au delà de ces performances de localisation très encourageantes face à des algorithmes reconnus issus de la littérature, l'approche par apprentissage offre également un avantage non visible sur le graphique : le temps de calcul. Les temps de calcul nécessaires pour déterminer la DOA 2D de 360 sources à partir de trames de 1 024 échantillons identiques sont récapitulés dans le tableau 5.7. Alors qu'il faut en moyenne 4 min à l'algorithme MUSIC pour estimer 360 positions de sources, l'approche BeamLearning en inférence ne nécessite sur GPU que 2,1 s, soit 100 fois moins de temps environ. Si les temps de calculs sont plus élevés lorsque l'approche proposée est implémentée sur CPU, l'approche BeamLearning reste plus rapide que les deux méthodes modèles proposées. Pour bien comprendre l'intérêt de ce type de réduction drastique du temps de calcul pour des applications de localisation en temps réel, où l'implémentation GPU est tout à fait envisageable, il faut comparer ces temps caractéristiques à la longueur temporelle de chaque trame ayant permis les estimations, qui correspond à une durée de 23 ms. Ainsi, dans le cas de l'estimation de la position d'une seule source, l'estimation d'une seule source peut se faire plus rapidement (6 ms en moyenne) que l'enregistrement du signal avec l'approche par apprentissage, alors qu'une estimation par l'algorithme MUSIC ou SRP-PHAT nécessiterait une réduction de la résolution de la grille de recherche (et donc d'une perte de résolution et de précision) pour atteindre des objectifs d'estimation de position en temps réel.)

Algorithme	Temps de calcul CPU	Temps de calcul GPU
MUSIC	3 min 54s	
SRP-PHAT	16,7 s	
BeamLearning	**16,4 s**	**2,1 s**

Tableau 5.7 – Temps de calcul des algorithmes pour 360 sources

5.2.5 Localisation en environnement réverbérant, avec bruit de mesure

Les analyses menées dans les sections précédentes ont révélé que l'approche proposée offrait des performances de localisation intéressantes en champ libre, y compris en présence de bruit et sans

étalonnage préalable des capteurs de l'antenne microphonique. Le fait que ces performances soient *a minima* similaires aux méthodes MUSIC et SRP-PHAT en champ libre nous incite désormais à analyser le comportement du réseau de neurones proposé lorsqu'il s'agit de réaliser une tâche de détermination de DOA 2D par régression angulaire, en environnement réverbérant. Pour cela, cette section présente spécifiquement les performances de l'approche BeamLearning dans une salle de cours simulée à l'aide des outils développés au chapitre 3, de dimensions $10 \times 7 \times 3,7\ m^3$. La durée de réverbération de la pièce est d'une demie seconde, et les parois délimitant la salle sont modélisées simplement par un coefficient d'absorption constant sur chacune d'entre elles, d'une valeur de 0,312.

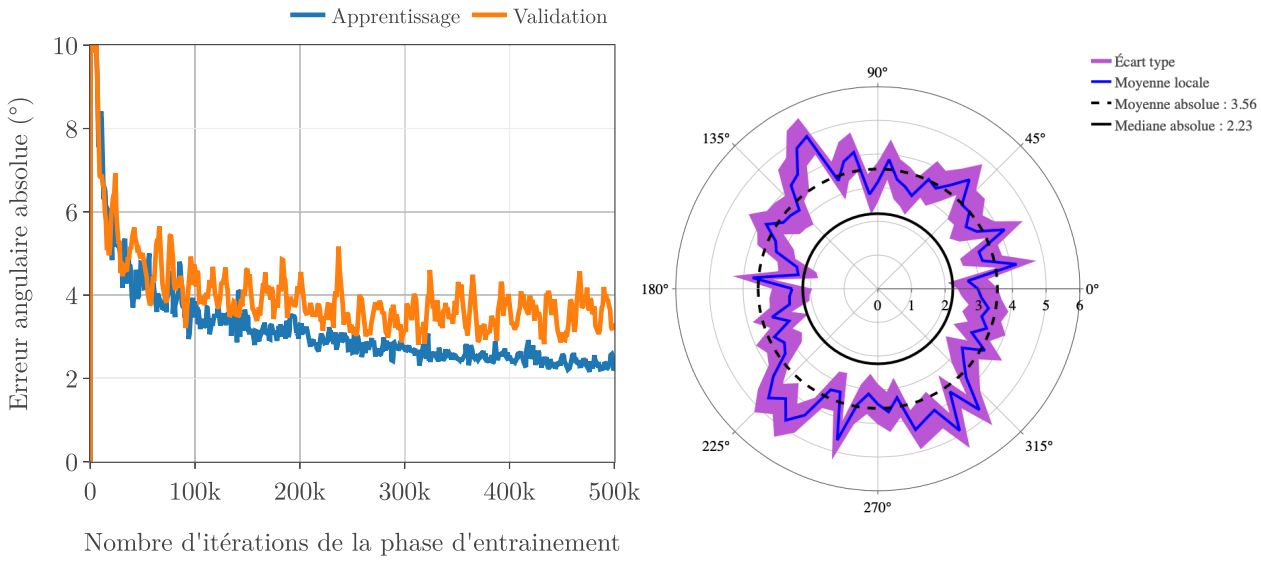
Dans cette section du manuscrit, nous nous restreignons à une tâche de détermination de DOA 2D, c'est à dire d'une détermination azimutale, pour une source dans la pièce – même lorsqu'elle n'est pas située dans le plan de l'antenne. Une étude plus complète de localisation 3D sera proposée dans la section 5.3. Pour ce faire, la base de données d'apprentissage est constituée selon les caractéristiques récapitulées dans le tableau 5.8. Le réseau est donc toujours le même, seul le jeu de données a été constitué à l'aide des méthodes développées au chapitre 3 pour entraîner le réseau à localiser des sources en présence de réverbération.

Données	Base de données	Environnement	Signal	Antenne	RSB
	Simulée	Salle (Tr = 0,5 s)	Multi signaux	Mini DSP.	> 20 dB
Résultats	Sortie	Angle estimé	Précision	Nb. itérations	Sur apprentissage
	Régression	azimut à 360°	$3,5°$	500 000	Léger

Tableau 5.8 – Récapitulatif synthétique des paramètres pour l'apprentissage présenté en section 5.2.5

La figure 5.14 permet d'analyser les performances de convergence au cours de l'entraînement, et d'analyser finement les performances de localisation sur un jeu de données de validation disjoint du jeu de données d'apprentissage, à l'issue de la dernière itération d'entraînement. Par comparaison avec les résultats obtenus en champ libre, on peut constater qu'ici, la précision angulaire obtenue sur des données de validation se sont légèrement dégradées, puisqu'elles atteignent une valeur de 3,5 ° environ (voir

figure 5.14(b)). De plus, on peut observer sur la figure 5.14(a) un léger phénomène de sur-apprentissage, puisque les courbes de convergence obtenues à partir des jeux de données d'apprentissage et des jeux de données de validation ne sont plus confondues. Toutefois, ce léger sur-apprentissage n'est pas critique pour les performances du réseau.



(a) Courbe d'apprentissage de BeamLearning

(b) Précision angulaire moyenne et écart type

FIGURE 5.14 – Performances de l'approche BeamLearning dans le cas de localisation par régression : données simulées en environnement réverbérant à partir de différents signaux tels que du bruit de *cocktail party*, un klaxon ou de la musique classique avec un $RSB \geq 20$ dB. (a) : Courbe de convergence d'apprentissage du réseau, obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement. (b) Représentation polaire statistique des erreurs angulaires absolues, obtenues à l'issue de la dernière itération d'entraînement, sur un jeu de données test correspondant à 4 800 sources réparties uniformément autour de l'antenne.

Enfin, l'analyse de la figure 5.14(b), obtenue comme précédemment à partir de 4 800 exemples de validation disjoints des données utilisées pour l'entraînement, permet d'analyser plus finement les performances sous un angle statistique local et global. En premier lieu, la moyenne locale reste très proche de la moyenne globale sur l'ensemble des valeurs d'azimut testées, ce qui indique qu'aucune direction n'est statistiquement prédominante et que malgré la réverbération dans l'environnement de mesure, les performances restent relativement homogènes d'un point de vue spatial. Par ailleurs, l'écart-type des erreurs angulaires d'estimation reste lui aussi homogène sur tout le domaine angulaire,

avec une valeur de l'ordre de 1 °. Pour finir, il est particulièrement instructif d'observer que la valeur de la médiane des erreurs d'estimations, en trait noir continu sur la figure 5.14(b), est inférieure de plus de 1° par rapport à la moyenne globale des erreurs (en trait pointillé), ce qui signifie que seul un très faible nombre d'observables mènent tout de même à une erreur d'estimation importante de position.

5.2.6 Influence du rapport signal à bruit utilisé lors de la phase d'apprentissage.

Le fait de rajouter du bruit aux données lors de la phase d'apprentissage peut être vue comme une méthode d'augmentation de données, permettant au réseau de mieux généraliser les paramètres permettant de déterminer sa sortie [197]. Dans notre cas, comme vu dans la section 5.2.4, ce procédé permet à l'approche de rester robuste aux mesures dégradées par un bruit de fond, y compris dans des conditions plutôt défavorables. En pratique, l'analyse menée en champ libre et la confrontation aux méthodes MUSIC et SRP-PHAT en 5.2.4 a permis de mettre en évidence le fait que les variables d'apprentissages sont optimisées pour s'affranchir du bruit de fond et localiser efficacement dans ces conditions grâce à l'augmentation de données par ajout de bruit au cours de l'entraînement. Les algorithmes de localisation de sources acoustiques conventionnels reposant sur une approche *modèle* ont d'ailleurs pour la plupart vocation à être implémentés non pas dans un environnement contrôlé de laboratoire, mais sur des sites industriels ou dans des environnements potentiellement bruyants, où l'environnement est *a priori* moins propice à la mesure acoustique. L'approche BeamLearning se doit donc elle aussi d'être le plus robuste possible au bruit de fond.

Pour compléter l'analyse menée à la section 5.2.4, l'objectif est ici de comparer, en environnement réverbérant et bruité, les performances de localisation offertes par l'approche BeamLearning, à celles des algorithmes MUSIC et SRP-PHAT. Tout comme dans la section précédente, l'environnement de mesure est ici la salle de cours simulée, avec une durée de réverbération de 0,5 s. Les sources à localiser émettent ici un signal de type *cocktail party*, pour différents rapports signal à bruit allant de -1 dB à 40 dB.

Afin de pouvoir analyser l'influence des paramètres utilisés pour l'augmentation de données par ajout de bruit (voir section 5.1.2) au cours de la phase d'apprentissage, deux entraînements distincts du réseau ont été réalisés :

— un entraînement avec augmentation de données par ajout de bruit limité à un RSB $>$ 20 dB (courbe rouge sur la figure 5.15)

— un entraînement avec augmentation de données par ajout de bruit limité à un RSB $>$ 10 dB (courbe orange sur la figure 5.15).

Ces deux apprentissages ne diffèrent donc qu'au niveau du RSB minimal utilisé lors de la phase d'apprentissage, pour l'ajout de bruit de mesure sur les voies microphoniques (voir section 5.1.2) : le même jeu de données est utilisé pour les deux entraînements du même réseau, seule la méthode d'augmentation de données est différente.

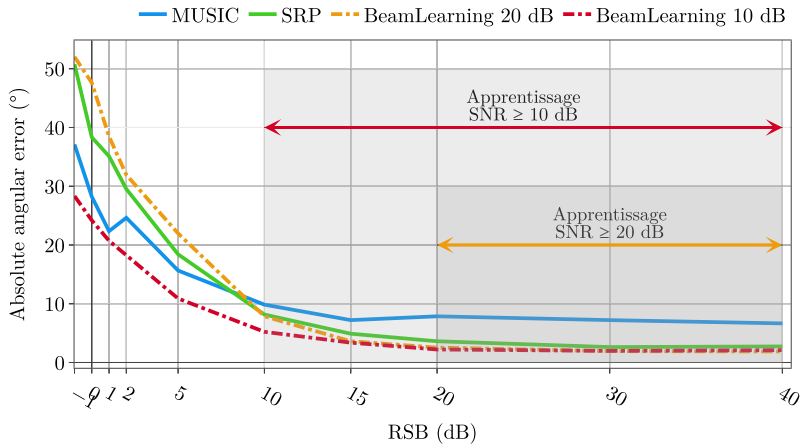


FIGURE 5.15 – Erreur angulaire absolue moyenne de localisation, sur des signaux de type *cocktail party* pour un $RSB \in [-1; 40]$ dB, pour des algorithmes issus de modèles (en traits plein) : MUSIC (bleu) et SRP-PHAT (vert) et pour l'approche BeamLearning (trait mixte), entraîné avec des jeux de données augmentés par ajout de bruit de mesure avec un $RSB > 20$ dB (orange) ou un $RSB > 10$ dB (rouge).

La figure 5.15 est construite en utilisant exactement le même processus qu'à la section 5.2.4 pour la figure 5.13, exception faite qu'ici, la localisation est réalisée en environnement réverbérant. Tout comme ce qui a été observé en champ libre, les 3 performances de localisation des 3 méthodes se dégradent naturellement lorsque le bruit de fond est élevé (RSB inférieur à 5 dB), même si elles sont toutes les trois conçues pour offrir une robustesse accrue à ce type de situation.

En ce qui concerne les approches conventionnelles reposant sur une approche modèle , l'algorithme MUSIC, basé sur une séparation en sous-espace bruit et sous-espace mesure, offre de meilleures performances que l'algorithme SRP-PHAT dans ce domaine à faible RSB. En revanche, l'approche sous-jacente à la méthode MUSIC est dégradée lorsque des sources corrélées sont prédominantes, comme c'est le cas avec les réflexions précoces en acoustique des salles à fort RSB, ce qui explique que pour un RSB supérieur à 15 dB, ce soit cette fois-ci la méthode SRP-PHAT qui surpasse la méthode MUSIC dans l'environnement réverbérant testé.

En ce qui concerne l'approche BeamLearning proposée, reposant quant à elle sur les données et un mécanisme d'apprentissage par Deep Learning, l'expérience réalisée ici permet d'étudier finement l'influence du paramètre de RSB minimal utilisé pour l'augmentation de données au cours de l'entraînement. Sur la figure 5.15, pour chacun des deux entraînements en orange et en rouge, on peut observer deux *zones* de RSB qui suivent un comportement plutôt intuitif compte tenu des jeux de données présentés au réseau au cours de l'entraînement.

On constate en effet que lorsque le bruit utilisé pour l'inférence sur les données de validation correspond à des valeurs de RSB similaires à celles utilisées au cours de l'entraînement, les performances sont très stables et offrent une très bonne précision de localisation dans les deux situations d'entraînement. Pour ces situations à *fort* RSB, en environnement réverbérant, on constate d'ailleurs que l'approche BeamLearning surpasse les performances de l'approche MUSIC d'environ 6 à 8° de précision, et présente des performances légèrement meilleures à la méthode SRP-PHAT. Ainsi, contrairement à MUSIC, le réseau semble non seulement avoir appris à s'affranchir du bruit de mesure et de la réverbération diffuse, mais également des réflexions précoces.

Dans des conditions de mesure beaucoup plus dégradées, c'est à dire pour un $RSB < 10\ dB$, c'est à dire pour une situation d'inférence avec un bruit de mesure bien supérieur à celui présenté au cours de l'entraînement, on constate que les performances du réseau dépendent plus fortement du choix réalisé pour l'augmentation de données par ajout de bruit. Tout naturellement, il apparaît que le réseau entraîné avec un $RSB > 10\ dB$ reste plus robuste à des situations très dégradées, surpassant d'ailleurs assez largement les approches MUSIC et SRP-PHAT dans ces conditions très défavorables.

Le réseau entraîné avec un $RSB > 20\ dB$, quant à lui, n'est pas assez robuste à très faible RSB, et présente des erreurs d'estimation supérieures aux approches *modèle*.

Algorithme	Erreurs (°)	
	$RSB = 20\ dB$	$RSB = 5\ dB$
MUSIC	7,8	15,6
SRP-PHAT	3,6	18,4
BeamLearning 20 dB	2,3	22,4
BeamLearning 10 dB	**2,2**	**10,8**

Tableau 5.9 – Performances des algorithmes MUSIC, SRP-PHAT et de l'approche BeamLearning extraites de la figure 5.15.

Afin d'illustrer ces phénomènes, sur le tableau 5.9, deux points particuliers sont extraits de la figure 5.15, permettant de démontrer la supériorité de la méthode BeamLearning par rapport à des approches *modèles* pourtant optimisées pour offrir une localisation précise en présence de bruit et de réverbération. Les valeurs de RSB choisies (20 dB et 5 dB) correspondent à des *régimes* de fonctionnement où la méthode SRP-PHAT ou la méthode MUSIC sont considérées comme performantes. On peut observer qu'en milieu réverbérant et en présence de réflexions précoces, le réseau de neurones profond proposé permet systématiquement d'obtenir les estimations de DOA 2D les plus précises, que ce soit à niveau de bruit de fond *raisonnable*, ou en présence d'un fort bruit de fond. Cette robustesse dans des conditions très dégradées est accessible en entraînant le réseau avec des données à RSB variables, jusqu'à un RSB maximal de 10 dB, ce qui offre une méthodologie parfaitement maîtrisée pour obtenir une détermination de DOA très satisfaisante.

Ces constatations ayant été réalisées pour une détermination d'azimut, il apparaît maintenant naturel d'étudier l'approche BeamLearning pour la détermination de DOA 3D. Pour cela, l'un des avantages de la méthode proposée ici réside dans le fait qu'aucune modification sur l'architecture du réseau n'est nécessaire, à l'exception de la fonction de coût permettant de réaliser l'optimisation au cours de l'entraînement.

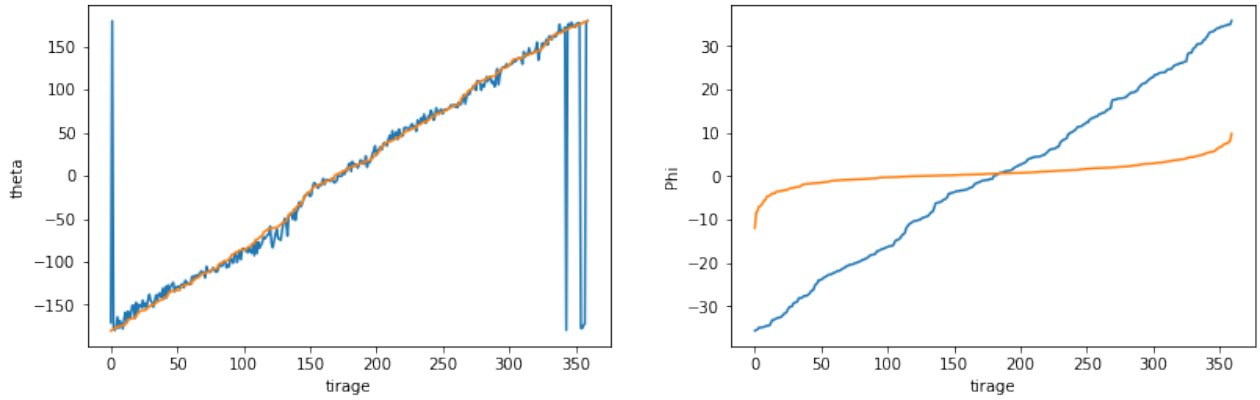
5.3 Détermination de DOA 3D en environnement réverbérant et bruité

5.3.1 Ambiguïté d'élévation avec une antenne plane

L'un des objectifs de l'approche BeamLearning et du réseau de neurones profond associé repose sur le fait d'exploiter le même réseau, quelque soit la géométrie de l'antenne, de l'environnement de mesure, ou du type de source à localiser. La seule modification – minime – associée au passage d'un problème d'estimation de DOA 2D à un problème 3D repose sur la dimension du vecteur de sortie, et l'adaptation de la fonction de coût à ce changement de dimension. Par conséquent, pour commencer, nous avons fait le choix d'analyser les performances de localisation dans une situation similaire à celles analysées dans la section précédente, avec une antenne plane. Il est pourtant bien connu que les techniques de localisation de sources conventionnelles reposant sur des approches *modèles* se heurtent à l'ambiguïté de détermination d'élévation du fait de la forte symétrie entre le demi espace supérieur et le demi espace inférieur à l'antenne. Pour une approche BeamLearning, il apparaissait intéressant d'observer le comportement du réseau pour une antenne plane, en 3D. Pour accentuer l'influence de la symétrie, nous avons choisi de nous placer dans un situation de champ libre, modélisée par une salle aux parois parfaitement absorbantes, et de tester la localisation pour des valeurs d'élévation comprises entre $-36°$ et $+36°$, qui présente l'avantage d'être parfaitement symétrique par rapport au plan de l'antenne, et de maximiser les ambiguïtés haut/bas au cours de l'entraînement.

Compte tenu de ce cadre géométrique fixé pour l'entraînement du réseau avec une antenne plane, nous avons donc constitué un jeu de données à partir de sources sonores situées dans un volume centré sur l'antenne, délimité par deux sphères tronquées aux pôles et de rayons respectifs 1 m 50 et 2 m 50.

Afin d'analyser les performances de localisation 3D dans ce cas d'étude de *transition*, les champs captés par l'antenne, correspondant à l'émission de 360 sources non comprises dans les données d'entraînement, ont été fournies en entrée du réseau à l'issue de sa convergence d'apprentissage. Ces résultats sont présentés sur la figure 5.16. Comme attendu, les méthodes d'apprentissage ne permettent pas de lever l'ambiguïté d'élévation puisque ce problème est inhérent aux données, qui sont identiques de part et d'autre de l'antenne. Une analyse des erreurs azimutales et d'élévations sur la figure 5.16 montre surtout une dégradation de la localisation au niveau de l'angle d'élévation. Les fortes disparités pour



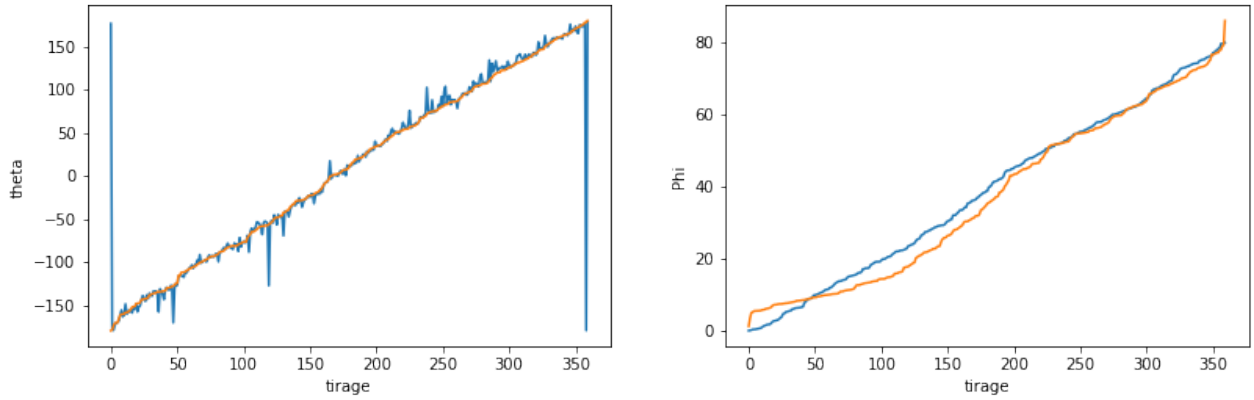
(a) Position réelle et estimée de l'angle azimutal (°) (b) Position réelle et estimée de l'angle d'élévation (°)

FIGURE 5.16 – Position réelle (bleue) et estimée par le réseau de neurones proposé (orange), des angles azimutal (gauche) et d'élévation (droite), pour 360 tirages successifs de sources en champ libre.

l'estimation azimutale, visibles sur la figure 5.16(a) ne sont dues qu'à des erreurs d'environ 360°, avec une source positionnée en $-180°$ mais estimée à $+178°$ par exemple, qui correspondent en réalité à une erreur d'estimation azimutale de seulement 2°.

En revanche, l'analyse des estimations en élévation de la figure 5.16(b) révèle la difficulté à localiser les sources en 3D avec une antenne plane : alors que les sources sont positionnées à toutes les élévations possibles dans le domaine $[-36; 36]°$, les estimations en élévation fournies en sortie du réseau entraîné avec une antenne plane sont quasi exclusivement réalisées dans le plan de l'antenne. Par rapport aux approches *modèles*, cette *ambiguïté* de détermination d'élévation avec une antenne plane se manifeste par un caractère assez original. En effet, les algorithmes de détermination de DOA 3D conventionnels présentent en général des erreurs aléatoires sur le signe de l'angle d'élévation, tandis que notre approche par optimisation converge vers une estimation d'élévation qui minimise les erreurs sur l'ensemble des lots d'apprentissage : la moyenne des élévations, qui est ici confondue avec le plan de l'antenne.

Pour confirmer cette hypothèse que les moins bonnes performances sont bien dues à un problème d'ambiguïté haut/bas lors de la captation des sources, une autre expérience numérique a été réalisée sur le même principe que la précédente, avec une antenne plane posée sur un sol parfaitement réfléchissant. Cette approche avait été proposée dans la thèse d'Aro Ramamonjy [14] avec d'autres méthodes



(a) Position réelle et estimée de l'angle azimutal (°) (b) Position réelle et estimée de l'angle d'élévation (°)

FIGURE 5.17 – Position réelle (bleue) et estimée (orange) des angles azimutal (gauche) et d'élévation (droite) pour une antenne **posée au sol**. Les positions des sources sont sur une demi-sphère.

de localisation de sources. Dans ce cas, la localisation ne se fait plus sur une sphère complète, mais seulement dans le demi espace défini par $\phi \geq 0$. En analysant sur la figure 5.17 les résultats obtenus pour l'estimation de DOA 3D de sources à l'aide d'une antenne plane posée sur un sol réfléchissant, on constate que le problème illustré par la figure 5.16 est bien levé : les angles d'azimut et d'élévation estimés correspondent mieux cette fois-ci aux DOA des sources à localiser. Sur ces courbes, des erreurs résiduelles subsistent, mais dans ce cas, l'apprentissage n'a pas été mené jusqu'à la convergence complète, puisque l'objectif était seulement de confirmer ce principe et de l'illustrer. Ainsi, même si l'approche de localisation de sources par BeamLearning est conçue pour être indépendante du type d'antenne de mesure, si celle-ci n'offre pas la diversité suffisante de données pour résoudre le problème de localisation, le problème d'optimisation restera conditionné à l'utilisation d'un jeu de données inadapté au problème, et mènera irrémédiablement à des performances dégradées.

Pour cette raison, et puisque la constitution de bases de données expérimentales nécessite de placer les antennes microphoniques à hauteur de l'équateur de la sphère de spatialisation, dans toute la suite du document, nous n'utiliserons donc plus que des antennes non planes, puisqu'elles permettent d'éviter le phénomène d'ambiguïté en élévation illustré ici.

5.3.2 Détermination de DOA 3D dans une salle réverbérante : expérience numérique

Afin de lever l'ambiguïté en élévation du problème, il faut utiliser une antenne dont les microphones sont disposés non plus dans un seul plan, mais dans un volume. Toutefois, pour être en mesure de faire des simulations sans avoir à simuler également les phénomènes de diffraction par le corps de l'antenne, il est nécessaire d'utiliser une antenne la plus *ouverte* possible. Or, au cours de sa thèse de doctorat au laboratoire, Aro Ramamonjy a développé, parmi d'autres prototypes, une antenne à géométrie tétraédrique constituée de microphones doubles couches [14]. Les caractéristiques de cette antenne, appelée CMA, sont données en section 3.2.1 (pour sa version *numérique*) et 4.2.2 (pour l'antenne *réelle*). Pour rappel, l'envergure de cette antenne tétraédrique antenne est d'une dizaine de centimètres de rayon, et les 7 microphones qui la composent sont espacés d'une distance de 3 *cm* environ. [4]

Pour cette nouvelle expérience de détermination de DOA 3D grâce à une antenne compacte, l'antenne est positionnée dans une salle de dimensions $10 \times 7 \times 3,7\ m^3$ simulée numériquement. La durée de réverbération de la salle est de 0,5 s, et le coefficient d'absorption est cette fois ci hétérogène entre les types de parois : $\alpha_{murs} = 0,312;\ \alpha_{plafond} = 0,412;\ \alpha_{sol} = 0,212$.

Le problème de localisation étant plus complexe, l'entraînement n'est ici réalisé que sur un jeu de données constitué de sources émettant des signaux de type *cocktail party* (voir section 3.2.4). Ainsi, plutôt que d'avoir 6 types de signaux différents pour chaque élément du jeu de données d'apprentissage comme c'était le cas pour les résultats exposés en section 5.2, le jeu de données utilisé ici n'est plus *multi signaux*, mais présente toujours une variabilité importante de contenus spectraux. L'entraînement est également accompagné par une procédure d'augmentation de données par ajout de bruit de mesure selon la procédure exposée en 5.1.2. Pour les expériences menées dans le cadre de cette sous-section, le tableau 5.10 récapitule de manière synthétique les paramètres liés à l'entraînement du réseau, et au jeu de données utilisé pour cet entraînement.

4. Cette antenne a le même nombre de capteurs que l'antenne Mini DSP utilisée pour la localisation dans le plan. Un des objectifs de la thèse était de vérifier expérimentalement sur différentes antennes les performances de l'approche BeamLearning. Malheureusement, suite à une accumulation de problèmes indépendants de notre volonté (défaillance matérielle, travaux de mise aux normes du laboratoire et pandémie internationale), ces comparaison n'ont pas pu être menées pour l'antenne CMA.

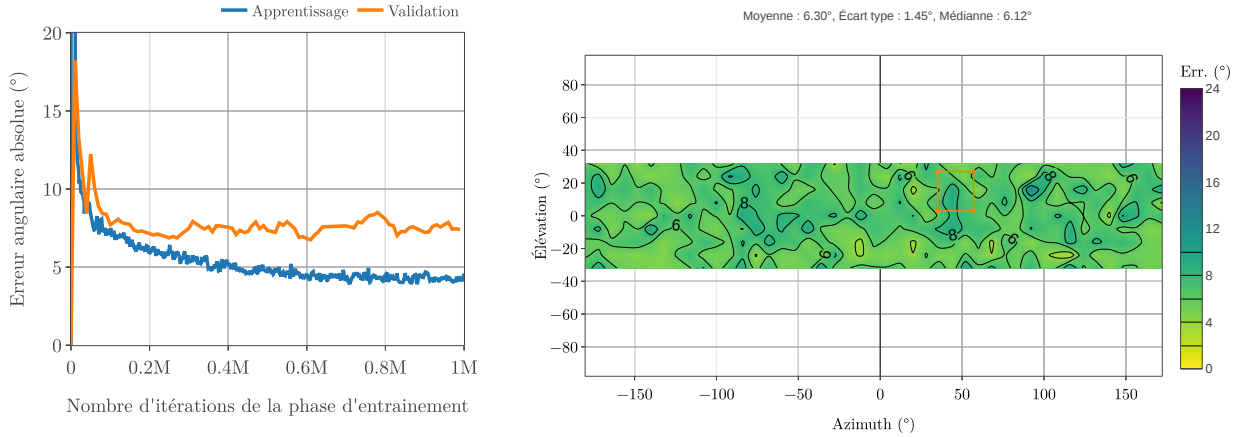
Données	Base de données	Environnement	Signal	Antenne	RSB
	Simulée	Salle (Tr = 0,5 s)	*Cocktail party*	CMA Cube	> 10 dB
Résultats	Sortie	Angles estimés	Précision (° solides)	Nb. itérations	Sur apprentissage
	Régression	azimut à 360° élévation à ±36°	7°	1 000 000	Oui

Tableau 5.10 – Récapitulatif synthétique des paramètres pour l'apprentissage présenté en section 5.3.2

Afin d'analyser finement les performances de localisation 3D avec cette antenne CMA tétraédrique et ce jeu de données d'entraînement, la figure 5.18 présente la convergence obtenue au cours de la procédure d'entraînement du réseau pendant 10^6 itérations (convergence sur le jeu de données d'apprentissage et sur le jeu de données de validation, disjoint du précédent et ne servant pas à optimiser les variables du réseau), ainsi qu'une carte d'erreurs représentées dans un diagramme (θ, ϕ) avec le réseau correspondant à la dernière itération de l'entraînement, et le jeu de données de validation.

L'analyse de la courbe de convergence d'apprentissage 5.18(a) révèle ici que les performances de localisation sur le jeu de données d'entraînement atteignent, à l'issue d'un million d'itérations, une erreur angulaire absolue moyenne de 4 degrés environ, ce qui ne représente qu'une faible dégradation des performances par rapport au problème de localisation angulaire à 2D étudié en section 5.2.

En revanche, il est primordial ici de constater que les performances obtenues sur le jeu de données de validation, beaucoup plus représentatives du comportement du réseau pour une inférence après gel du réseau, s'écartent progressivement des performances obtenues sur le jeu de données d'entraînement. Ce phénomène est typique d'un sur-apprentissage, et fera l'objet d'une analyse approfondie dans la section suivante, mais il mène à une dégradation des performances d'inférence en 3D par rapport au problème 2D sur ce jeu de données précis : ici, l'erreur angulaire absolue moyenne obtenue à la fin de l'entraînement atteint une valeur moyenne de 7° environ pour un jeu de données de validation avec un RSB aléatoire supérieur à 10 dB (voir figure 5.18(a)).



(a) Courbe d'apprentissage de BeamLearning. $RSB \geq 10dB$

(b) Précision angulaire solide (moyenne sur un cône de 8° d'ouverture). RSB=20dB

FIGURE 5.18 – Performances de l'approche BeamLearning dans le cas de DOA à 3D : données simulées en environnement réverbérant à partir de signaux de type *cocktail party* avec un $RSB \geq 10$ dB. (a) : Courbe de convergence d'apprentissage du réseau obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement. (b) Représentation polaire statistique des erreurs angulaires absolues, obtenues à l'issue de la dernière itération d'entraînement, sur un jeu de données test correspondant à 4 700 $\theta \in [0°; 360°[,\ \phi \in [-36°; 36°]$.

La figure 5.18(b), quant à elle, permet d'analyser le comportement des erreurs en fonction de la position angulaire 3D de la source à localiser. Pour construire cette cartographie, et de manière à pouvoir comparer quantitativement les résultats à ceux obtenus en 2 dimensions sur la figure 5.14, le jeu de données de test utilisé ici (disjoint des données ayant servi à l'entraînement du réseau) est constitué exclusivement de données avec un RSB de 20 dB. Pour ces données test, l'erreur angulaire absolue moyenne est de $6,3°$, et reste relativement homogène, puisque l'écart-type de l'erreur de localisation sur tout le volume est de seulement $1,45°$. Pour que la cartographie des erreurs d'estimation angulaire 3D soit statistiquement représentative, elle a été calculée à partir de la moyenne des erreurs angulaires absolues sur des secteurs angulaires de 8° par 8°, avec un moyennage exécuté sur 20 tirages de bruit blanc de mesure pour chaque position de sources testées. Cette analyse permet d'obtenir une statistique d'environ 100 positions de sources par secteur angulaire, correspondant à la dimensions des batchs utilisés pour l'évaluation de la convergence sur la figure 5.18(a).

Même si les performances de détermination de DOA 3D en environnement réverbérant et bruité obtenues ici sont légèrement en deçà des performances obtenues pour une tâche de détermination de DOA 2D, elles n'en restent pas moins plutôt satisfaisantes, d'autant qu'elles restent très homogènes pour toutes les valeurs d'azimut et d'élévation (voir figure 5.18(b)).

Afin de mieux comprendre la statistique obtenue sur la figure 5.18(b), une zone où les erreurs présente une inhomogénéité ponctuelle a été sélectionnée (cadre orange). Dans cette zone, l'erreur moyenne est d'environ 6 à 8°, et est calculée à partir d'un ensemble de positions de sources, dont les erreurs d'estimation angulaires individuelles sont tracées sur la figure 5.19.

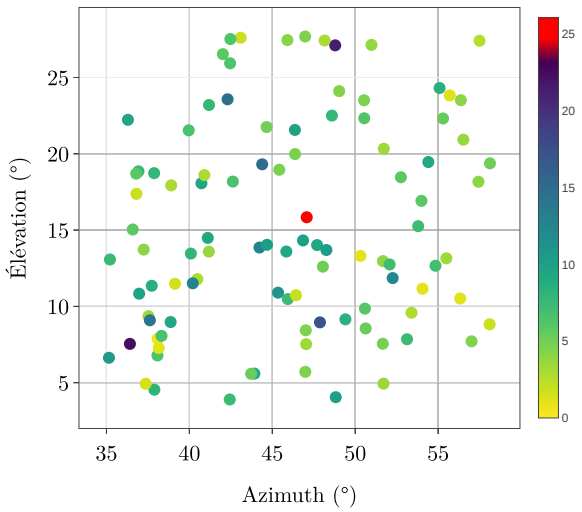


FIGURE 5.19 – Erreurs absolues moyennes de localisation, obtenues sur 20 tirages, pour des sources appartenant au cadre orange visible sur la figure 5.18(b). Le point rouge correspond à une valeur de 80°.

Sur cette figure 5.19, on peut observer qu'une seule position parmi les 101 positions possède une erreur d'estimation angulaire importante de 80° (représentée en rouge), et que la grande majorité des positions de source ont été estimées avec une erreur inférieure à 8°. Cette figure nous rappelle ainsi que l'objectif principal des méthodes d'optimisation sous-jacentes aux techniques d'apprentissage par Deep Learning consiste à minimiser l'erreur d'estimation, en moyenne, et sur un grand nombre de données

d'entrée, ce qui n'empêche pas d'obtenir des erreurs ponctuellement élevées, mais statistiquement non représentatives. En effet, même si nous sommes ici en présence d'un phénomène de sur-apprentissage qu'il est important d'éviter, ces erreurs élevées d'estimation restent extrêmement rares et disséminées sur l'ensemble du volume angulaire : sur les 4 700 positions de sources différentes ayant servi à tracer la cartographie 5.18(b), seulement 0,5% d'entre elles ont été estimées par la réseau avec une erreur supérieure à 25°, ce qui explique le faible écart-type global d'erreurs angulaires sur tout le volume.

Malgré ces résultats encourageants, il n'en reste pas moins que si un soin particulier n'est pas apporté à la taille de la base de données d'entraînement, l'approche de détermination de DOA 3D par BeamLearning peut être sujette à un net phénomène de sur-apprentissage, qui mène à une dégradation des performances par rapport au problème de détermination de DOA 2D, liée en partie à l'augmentation de la complexité du problème à résoudre. Pour améliorer le comportement en inférence et les capacités de généralisation des performances obtenues sur des données d'entraînement à des données de test, il n'est bien entendu ici pas question de modifier le réseau, qui est conçu pour être adaptable au problème, mais plutôt de travailler sur le volume de données nécessaires pour constituer les jeux de données d'entraînement.

5.3.3 Étude de l'influence du volume du jeu de données d'entraînement sur les performances de localisation 3D

Pour l'entraînement du réseau à une tâche de détermination de DOA 2D (section 5.2), les sources n'étaient localisées que dans le plan, la variabilité en élévation des positions de sources pour constituer le jeu de données d'apprentissage n'était donc que de $\pm 7°$. Pour la détermination de DOA 3D qui nous intéresse ici, les jeux de données étudiés jusqu'ici possèdent une variabilité en élévation de $\pm 36°$. Ainsi, le volume du tore définissant l'ensemble des positions possibles de sources utilisée pour constituer les jeux de données d'entraînement des sources est donc multiplié par 5 environ. Par conséquent, pour conserver la même densité de sources par unité volumique, il faut augmenter dans la même proportion le nombre de sources à tirer aléatoirement dans le tore pour constituer le jeu de données, alors que l'expérience présentée dans la section 5.3.2 précédente a été réalisée avec un jeu de données d'entraînement constitué du même nombre d'éléments que pour les tâches de détermination de DOA 2D, ce qui peut apporter une explication simple au phénomène de sur-apprentissage observé dans la section précédente.

Par conséquent, afin d'illustrer l'influence du volume de données utilisées pour entraîner le réseau sur le phénomène de sur-apprentissage tout en limitant les temps de calcul nécessaires aux apprentissages, il a été choisi d'entraîner à nouveau le réseau, tous les autres paramètres restant identiques à ceux de la section 5.3.2, avec 3 jeux de données différents :

— un jeu de données x1 constitué de 38400 positions de sources dans le volume,
— un jeu de données x2 constitué de 75800 positions de sources dans le volume,
— un jeu de données x3 constitué de 115200 positions de sources dans le volume.

Pour ces 3 jeux de données d'entraînement, compte tenu de la variabilité du volume de données présentées au réseau, il est en revanche important de noter que la vitesse de convergence est différente : pour un jeu de données plus volumineux, le nombre d'itération nécessaires avant convergence est plus grand. Les courbes de convergence d'apprentissage du réseau tracées sur la figure 5.20 correspondent à un entraînement pendant 10^6 itérations sur le jeu de données x1 (avec l'apparition franche d'un phénomène de sur-apprentissage à partir de 600 000 itérations), 2×10^6 itérations sur le jeu de données x2 (avec l'apparition de sur-apprentissage suspecté à partir de de 1.2×10^6 itérations), et 2×10^6 itérations sur le jeu de données x3 (sans l'apparition véritable de sur-apprentissage). Ce nombre conséquent d'itérations nécessite un temps de calcul variant entre 12 et 20 jours suivant les modèles de cartes Nvidia 1080Ti exploitées dans le serveur de calcul que nous utilisons [5] .

Les courbes de convergences pour ces trois apprentissages sont présentées sur la figure 5.20. L'analyse de ces résultats révèle en premier lieu que l'augmentation du volume de données d'entraînement diminue légèrement les performances de localisation sur la base de données d'apprentissage : la convergence sur les données utilisées pour entraîner le réseau atteint une valeur de 4° pour le jeu de données x1, tandis qu'il atteint une valeur de 5° pour le jeu de données x3, avec le double d'itérations d'entraînement. En soi, cette légère diminution n'est absolument pas problématique, puisqu'elle est à mettre en regard de l'amélioration des performances sur le jeu de données de validation lorsque le réseau est

5. Ces temps de calcul conséquents justifient ainsi la raison pour laquelle nous n'avons pas exploité un jeu de données *x5* puisque l'objectif ici est exclusivement didactique.

entraîné sur un volume de données plus important.

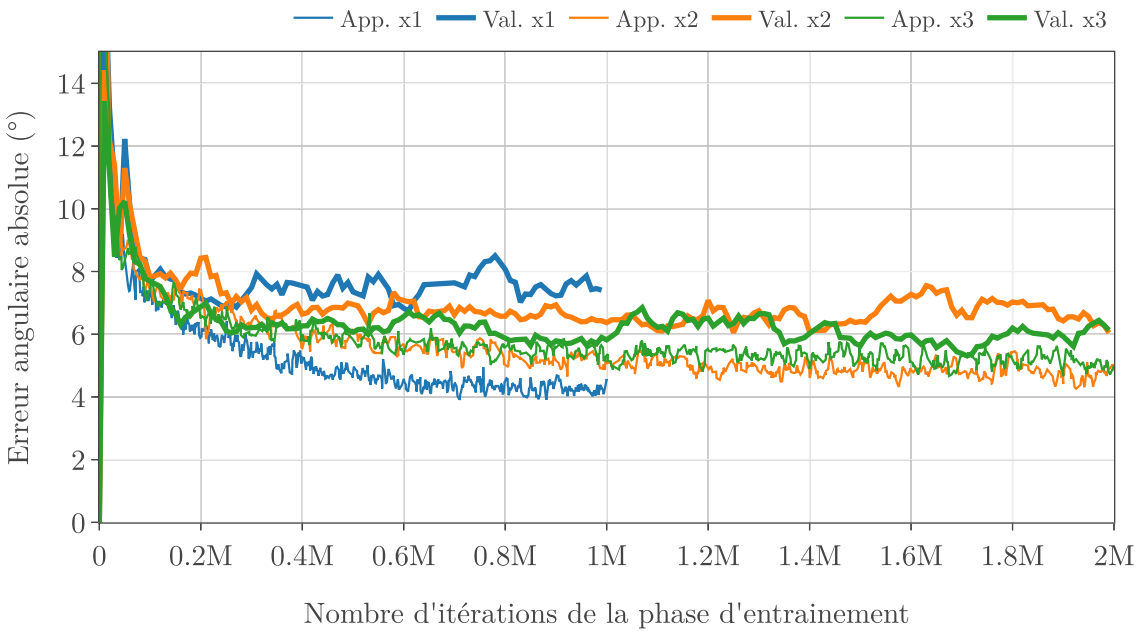


FIGURE 5.20 – Influence sur les performances du réseau du volume de données utilisé pour l'entraînement du réseau en phase d'apprentissage. En bleu : entraînement avec le jeu x1 (38 200 exemples) - en orange : entraînement avec le jeu x2 (76 400 exemples) - en vert : entraînement avec le jeu x3 (115 200 exemples). Pour chacun de ces trois entraînements, les courbes de convergence sont tracées en trait fin pour le jeu de données d'entraînement, et en trait gras pour le jeu de données de validation. Le signal utilisé pour l'apprentissage est un bruit de *cocktail party*. Un bruit aléatoire est rajouté pour l'apprentissage (RSB $\geq$ 10 dB).

Ainsi, l'utilisation du jeu de données x3 pour l'entraînement permet de supprimer tout phénomène de sur-apprentissage, et l'erreur angulaire absolue d'estimation converge pour le jeu de données de validation vers une valeur de 6° avec un entraînement sur le jeu de données le plus volumineux, tandis qu'il ne converge que vers une valeur de 8° avec un entraînement sur le jeu de données x1. L'augmentation du volume de données d'entraînement permet ainsi de rapprocher les performances obtenues pour l'entraînement et l'inférence sur des données jamais vues par le réseau. On s'assure ainsi, que la solution vers laquelle converge le réseau de neurones profond n'est pas spécifique aux données présentées lors de l'entraînement.

5.3.4 Validation expérimentale de la détermination de DOA 3D par Deep Learning

Les analyses menées dans les sous-sections précédentes à partir de champs de pression simulés numériquement ont permis de démontrer que l'approche BeamLearning permet d'estimer la DOA d'une source acoustique, y compris en milieu réverbérant et avec un bruit de mesure important. Pour cela, même si l'approche BeamLearning est indépendante de la topologie d'antenne utilisée, il est tout de même nécessaire d'utiliser une antenne microphonique permettant d'éviter des ambiguïtés dans les données mesurées, comme montré en section 5.3.1. En particulier, une antenne non-plane ou une antenne permettant d'exploiter la diffraction par la structure de l'antenne ou son proche environnement, apparaît comme une solution pertinente. Par ailleurs, nous avons également vu qu'en comparaison du problème de détermination de DOA 2D, le passage de l'estimation à 3 dimensions nécessite de constituer des jeux de données suffisamment volumineux pour résoudre ce problème plus complexe (voir section 5.3.3).

Fort de ces informations, l'objectif de cette sous-section est de réaliser une validation expérimentale de notre méthode d'estimation de DOA 3D, en utilisant le procédé de constitution de bases de données expérimentales par synthèse ambisonique décrit au chapitre 4. Pour cela, l'antenne microphonique testée ici est l'antenne Zylia-ZM1 (voir section 4.2.2 et annexe F), qui est une antenne sphérique rigide de 10 cm de diamètre environ, constituée de 19 microphones MEMS placés à la surface de la sphère. Le choix de cette antenne n'est bien entendu pas anodin, puisque la sphère rigide permet d'obtenir une diffraction importante par la structure de l'antenne, que les dimensions de l'antenne sont parfaitement compatibles avec les contraintes d'utilisation du spatialisateur, et que la géométrie se prête naturellement à une localisation tridimensionnelle de sources acoustiques.

Pour cette validation expérimentale, il est donc nécessaire d'entraîner le réseau à partir de données temporelles enregistrées sur l'antenne Zylia, avec une grande diversité de positions de sources, de contenu fréquentiel, de dynamique, et de RSB pour que le réseau soit robuste à ces paramètres afin de localiser efficacement une source en 3 dimensions. Pour cela, la procédure détaillée au chapitre 4 a été exploitée, en plaçant l'antenne Zylia au centre du spatialisateur 3D SpherBedev du laboratoire, qui a permis de générer en une nuit de mesures 43 200 champs spatialisés correspondant à des ondes planes

de directions différentes, avec des signaux spatialisés de type *cocktail party*. Pour cet apprentissage, l'environnement de la salle de spatialisation était toujours partiellement traité acoustiquement, avec un plafond parfaitement réfléchissant, de la moquette au sol, et des matériaux absorbants placés sur les parois latérales de la salle, tout comme à la section 5.1.5 et à la section 5.2.2. Parmi les 43 200 exemples de champs de pression mesurés correspondant à autant de positions angulaires de sources, 34 560 ont été utilisées pour constituer le jeu de données d'apprentissage, le reste des exemples étant utilisé pour la validation du réseau, et n'ont pas été utilisées pour l'entrainement et l'optimisation des variables du réseau de neurones profond. Le tableau 5.11 récapitule de manière synthétique les principales caractéristiques du jeu de données constitué et de l'entraînement réalisé grâce à ces données.

	Base de données	Environnement	Signal	Antenne	RSB
Données	Expérimentale	Salle partiellement traitée,une paroi parfaitement réfléchissante	*Cocktail party*	Zylia	> 20 dB
	Sortie	Angles estimés	Précision (° solides)	Nb. itérations	Sur apprentissage
Résultats	Régression	azimut à 360° élévation à 180°	4°	1 000 000	Léger

Tableau 5.11 – Récapitulatif synthétique des paramètres pour l'apprentissage présenté en section 5.3.4

Il est essentiel de noter ici que, malgré le fait que l'antenne Zylia-ZM1 soit initialement conçue pour réaliser des captations ambisoniques à l'ordre 3 de champs de pression acoustique, les données exploitées pour l'entrainement du réseau sont les données temporelles brutes des 19 canaux microphoniques de l'antenne, sans encodage ambisonique. En effet, l'objectif des travaux exposés dans cette thèse de doctorat est de proposer une méthode qui soit indépendante de la géométrie et de la topologie de l'antenne, et qui n'exploite que les données brutes mesurées par chacun des capteurs. Cette caractéristique représente une autre différence fondamentale avec les approches de localisation de sources *conventionnelles*, reposant sur des approches de type *modèles* : lorsqu'une antenne sphérique rigide est

utilisée, les approches de formations de voies, les algorithmes paramétriques, ou les algorithmes de séparation en sous-espaces nécessitent quasi-systématiquement d'être adaptés pour prendre en compte la diffraction par la sphère rigide, ou d'être transposés dans le domaine ambisonique [198]. Au contraire, la constitution et le traitement des données pour en faire un jeux de données utilisable pour l'approche BeamLearning est de l'ordre de 2 ou 3 jours, quelque soit la géométrie de l'antenne, sans modification du réseau ni de l'approche. Même si la phase d'apprentissage qui vient ensuite est plus longue (voir figure 5.21 : 10 jours sur une carte GPU Nvidia 1080Ti), cette optimisation ne requiert l'attention d'aucun opérateur, et n'a donc pas le même *coût* que le *temps de cerveau* nécessaire pour implémenter un algorithme reposant sur des modèles qui soit adapté à une géométrie particulière d'antenne diffractante.

Ces remarques ayant été faites sur les avantages d'un paradigme de Deep Learning pour la tâche de localisation de sources qui nous intéresse ici, la figure 5.21 permet d'analyser le comportement de la convergence au cours de la phase d'entraînement du réseau, à partir du jeu de données d'entraînement décrit précédemment, mesuré grâce à l'antenne Zylia. Comme pour chacune des courbes de convergence d'apprentissage présentées précédemment, la convergence est présentée au cours des 10^6 itérations d'entraînement, sur le jeu de données ayant servi à l'entraînement et sur le jeu de données de validation. L'analyse de ces résultats révèle qu'à l'issue de l'ensemble des itérations de la phase d'apprentissage, les performances de localisation sont très satisfaisantes, puisqu'elles convergent vers une erreur angulaire absolue de $2,5°$ environ sur le jeu de données d'entraînement.

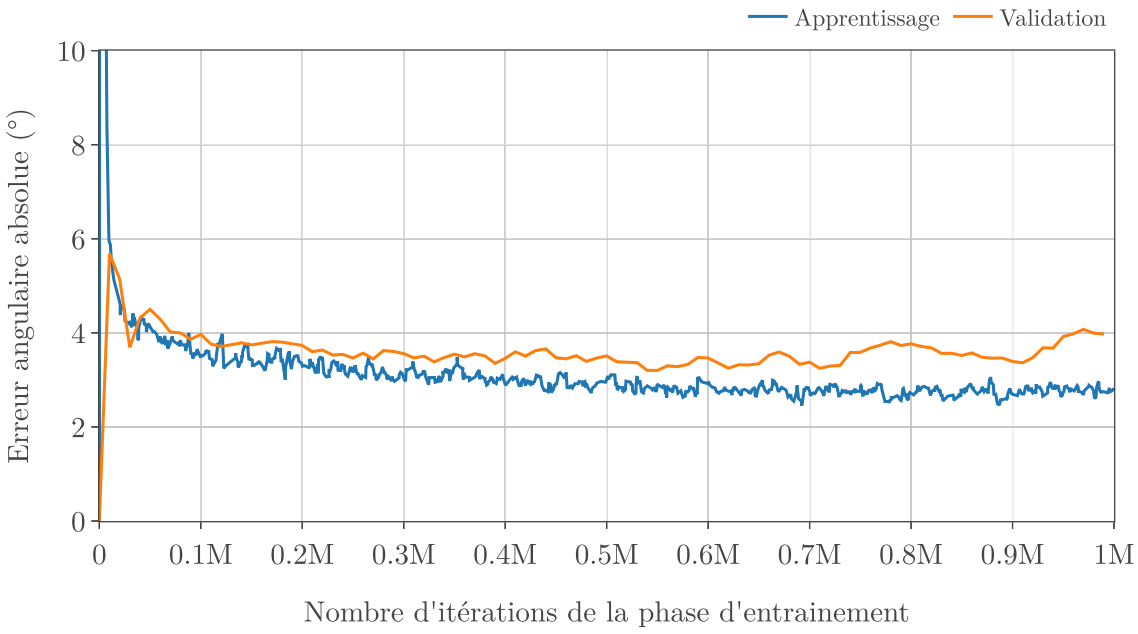


FIGURE 5.21 – Courbe de convergence d'apprentissage du réseau, obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement.

En revanche, on peut observer un phénomène de sur-apprentissage, puisqu'au delà de 600 000 itérations d'entraînement, les performances obtenues sur le jeu de données de validation tendent à se dégrader, tandis que les performances sur le jeu de données d'entraînement continuent à s'améliorer.

Compte tenu de cette observation, il est d'usage courant de réaliser une procédure *d'early stopping* (arrêt précoce) de l'entraînement [199], qui consiste à stopper l'entraînement avant que les performances du réseau ne se dégradent sur le jeu de données de validation. Par conséquent, en n'entraînant le réseau que sur 600 000 itérations, on obtient alors des performances d'estimation de DOA 3D atteignant une erreur de 3° sur le jeu de données d'entraînement, et de $3,5°$ sur le jeu de données de validation. La comparaison de ces très bonnes performances obtenues expérimentalement sur un jeu de données de validation à celles obtenues à partir de simulations numériques à la section 5.3.3 (voir figure 5.20) révèle qu'on obtient même ici des valeurs d'erreurs d'estimation plus faibles expérimentalement grâce à l'antenne Zylia. Cette amélioration sensible peut potentiellement provenir de plusieurs facteurs :

— en premier lieu, les données expérimentales ont été captées dans une pièce traitée acoustiquement, qui, même si elle présentait une forte réflexion au niveau du plafond, possède une durée de réverbération inférieure à la durée de réverbération de l'environnement simulé numériquement. Cette différence de complexité de l'environnement dans lequel est placé l'antenne peut ainsi contribuer à une modification des performances de localisation, comme exposé en section 5.2.2.

— par ailleurs, la structure même de l'antenne Zylia est diffractante, ce qui permet aux données utilisées au cours de l'entraînement de contenir des informations pertinentes pour la tâche de localisation de sources, contrairement à l'antenne tétraédrique CMA simulée, pour laquelle la structure est considérée comme parfaitement transparente acoustiquement. Sur ce point particulier, on peut dresser un parallèle formel avec le principe des HRTF en écoute binaurale [21] : l'être humain exploite, par apprentissage au cours de sa vie, les transformations apportées au champ de pression par sa tête, son pavillon, son buste, et son conduit auditif pour localiser des sources sonores, ce qui lui permet d'obtenir une performance de localisation bien supérieure à une situation où seuls deux capteurs seraient positionnés aux positions des oreilles, sans diffraction ou modification du champ. Dans un domaine moins relié à la psychophysique mais à l'utilisation de microphones, une étude récente [200] a permis de montrer qu'il était même possible de localiser des sources acoustiques à l'aide d'un seul microphone, en exploitant la diffraction du champ par des éléments dans l'environnement du capteur. Même si l'approche exploitée dans le cadre de cette thèse est fondamentalement différente de celle proposée dans [200], on peut raisonnablement suspecter que la diffraction offerte par la structure sphérique rigide de l'antenne Zylia permet ici également d'améliorer sensiblement les performances.

— dans une moindre mesure, entre l'étude numérique réalisée en section 5.3.2 et l'expérience réalisée ici, on peut évidemment constater que l'antenne CMA ne contient que 7 capteurs, tandis que l'antenne Zylia exploite 19 canaux microphoniques. Ainsi, le nombre d'informations disponibles en entrée du réseau est presque 3 fois plus important avec l'antenne Zylia. Cette augmentation du nombre de capteurs disponibles peut aider le réseau à construire une meilleure représentation de la grandeur de sortie permettant d'estimer la position angulaire de sources sonores. Aussi, la

section 5.3.5 est dédiée à l'étude des performances de localisation expérimentales en n'exploitant que 7 canaux microphoniques de l'antenne Zylia, afin d'évaluer l'influence de ce paramètre.

5.3.5 Influence du nombre de voies microphoniques de l'antenne intelligente

L'approche BeamLearning proposée permet d'obtenir une indépendance des traitements par rapport à la topologie de l'antenne et du nombre de voies microphoniques. Cependant, l'analyse des résultats expérimentaux obtenus avec l'antenne Zylia a révélé que les performances étaient améliorées par rapport aux études numériques menées avec l'antenne idéale CMA à 7 capteurs. Par conséquent, l'occasion se présente ici de tester les performances de l'antenne Zylia, en ne conservant que 7 voies microphoniques, dont les positions sur la surface de la sphère ont été choisies pour obtenir une géométrie similaire à l'antenne CMA (voir annexes E et F). Pour cela, puisque l'approche BeamLearning est parfaitement modulaire, aucune modification du code de calcul ou du réseau de neurones profond présenté au chapitre 2 n'est nécessaire, si ce n'est la dimension du tenseur d'entrée correspondant aux données microphoniques sur chacun des canaux.

Pour cela, un second apprentissage a été réalisé, avec les mêmes paramètres que ceux exposés dans la section précédente, en extrayant simplement 7 canaux microphoniques de l'antenne Zylia pour constituer le jeu de données d'entraînement et le jeu de données de validation. Les résultats obtenus sont représentés sur la figure 5.22 par les courbes de convergence de l'erreur angulaire absolue au cours de la phase d'entrainement, tant sur les données ayant servi à l'apprentissage que sur le jeu de données de validation. Pour faciliter la comparaison avec les performances obtenues dans la section précédente avec l'intégralité des voies microphoniques de l'antenne Zylia, les résultats de la figure 5.21 sont également tracés sur la figure 5.22.

L'analyse de la figure 5.22 met en évidence que le fait de n'utiliser que 7 capteurs parmi les 19 voies microphoniques de l'antenne Zylia dégrade légèrement les performances de localisation : l'erreur d'estimation angulaire moyenne à l'issue des itérations d'entraînement (ou après arrêt précoce à 600 000 itérations) augmente d'environ 0.5 degrés par rapport à un apprentissage réalisé à partir de l'intégralité des données disponibles grâce à l'antenne. Même si cette dégradation des performances

peut être considérée comme statistiquement représentative puisqu'elle est observée à la fois sur les jeux de données d'entraînement et les jeux de données de validation, la faiblesse de la dégradation peut justifier, lorsque le volume d'espace disque nécessaire au stockage des bases de données d'entraînement est primordial, de s'orienter vers l'utilisation d'une antenne possédant un nombre moins important de capteurs que l'antenne Zylia.

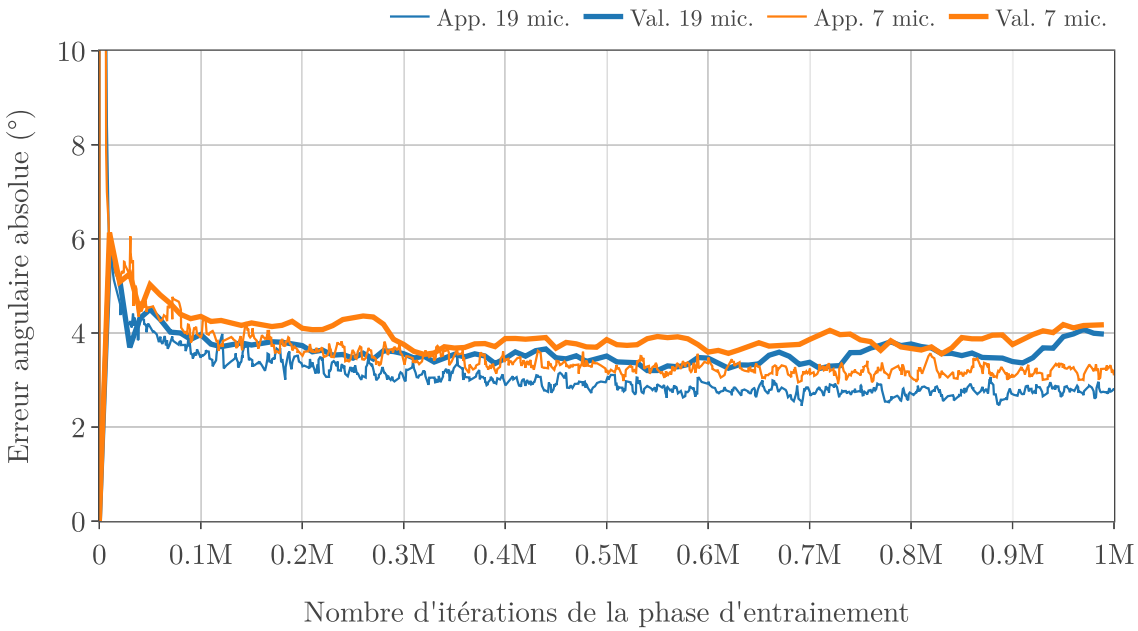


FIGURE 5.22 – Courbe de convergence d'apprentissage du réseau, obtenue à partir du jeu de données utilisé pour l'entraînement et du jeu de données de validation, disjoint du précédent, non utilisé pour l'entraînement. Prise en compte de toutes les voies microphoniques (en bleu) ou seulement de 7 voies (en orange).

En effet, si l'augmentation du nombre de canaux d'entrée ne modifie pas l'architecture générale du réseau de neurones profond, ni son empreinte mémoire lors de l'entraînement ou de l'inférence, le coût en terme de stockage des jeux de données pour l'entraînement lié à l'augmentation des canaux microphoniques est proportionnel au nombre de voies microphoniques exploitées. Par conséquent, si un arbitrage doit être fait sur le choix du nombre de microphones, avec comme objectif de minimiser l'espace disque nécessaire à l'entraînement du réseau sans dégrader de manière trop importante les performances, il est envisageable alors de s'orienter vers une antenne à 7 microphones plutôt qu'à 19.

Toutefois, une analyse plus fine de l'influence du nombre de capteurs pourrait être menée pour choisir le nombre de microphones le plus approprié pour représenter le meilleur compromis, ce qui n'est pas l'objectif dans le cadre de cette thèse.

5.3.6 Comparaison des performances d'estimation de DOA 3D avec l'algorithme SH-MUSIC

Compte tenu des résultats obtenus dans les sections précédentes, l'approche de localisation de sources par Deep Learning proposée dans cette thèse se révèle être particulièrement prometteuse. L'objectif initial de la thèse n'était pas de surpasser les méthodes *modèles*, mais d'analyser la faisabilité et la pertinence d'une approche reposant sur les données plutôt qu'une approche reposant sur des modèles. Cependant, les performances obtenues expérimentalement et numériquement, y compris dans des cas plutôt défavorables (environnement réverbérant, bruit de mesure d'amplitude élevée, non étalonnage des capteurs, ou diffraction par l'antenne) démontrent que le réseau de neurones profond développé offre une solution à la fois robuste et précise, qui pourrait rivaliser avec les méthodes conventionnelles les plus avancées.

L'antenne Zylia utilisée pour cette validation expérimentale étant optimisée pour réaliser un encodage ambisonique des champs mesurés, il apparaît ainsi naturel de s'intéresser aux méthodes *modèles* de la littérature les plus performantes pour réaliser une tâche de détermination de DOA avec ce type d'antenne sphérique rigide. La communauté scientifique a été extrêmement prolifique sur ce sujet, avec de nombreuses méthodes proposées, reposant sur des approches pouvant être classées de la même manière que ce qui a été proposé dans l'état de l'art de ce document à la section 1.1.4 [201]. Parmi l'ensemble de ces méthodes, l'une des approches les plus précises et les plus robustes, malgré son coût computationnel important, est la méthode SH-MUSIC [202, 203], à partir de laquelle plusieurs variantes ont été proposées, offrant toutes des performances de localisation similaires. Pour toutes ces raisons, nous avons donc choisi de confronter la méthode BeamLearrning à SH-MUSIC sur des données expérimentales.

En dehors du changement de paradigme modèle/données dont nous avons déjà largement discuté dans le reste du manuscrit, l'une des différences fondamentales ici repose sur le fait que l'approche

BeamLearning ne s'appuie pas sur une décomposition ambisonique du champ mesuré, mais bel et bien sur les données brutes des capteurs composant l'antenne sphérique, tandis que la méthode SH-MUSIC permet d'exploiter avantageusement les symétries de l'antenne, la diffraction par la structure sphérique, l'orthogonalité de la décomposition obtenue, et la réduction de dimension entre l'espace des positions des microphones et l'espace des composantes ambisoniques [202, 203].

Pour cette comparaison, l'approche BeamLearning exploite ici le réseau entraîné à partir de données expérimentales mesurées au mois de Février 2020 en section 5.3.4. Cet apprentissage a été réalisé à l'aide de la synthèse d'ondes planes par le spatialisateur SpherBedev, et des signaux de voix féminines de type *cocktail party*, dans la salle partiellement traitée, avec un plafond parfaitement réfléchissant.

Pour l'inférence réalisée ici, l'expérience a été réalisée au mois de Juillet, dans des conditions différentes de celles de l'apprentissage, de manière à se placer dans une situation réaliste d'un point de vue de l'inférence. En effet, entre Février et Juillet, nous avons réalisé une modification du traitement acoustique de la salle du spatialisateur 3D, avec l'installation d'un matériau absorbant large bande au plafond de la salle, avec un plénum de 10 cm. Par ailleurs, l'ensemble des matériaux absorbants placés sur les parois latérales ont été déplacés et modifiés. Dans une moindre mesure, les conditions de température de l'environnement de mesure sont nécessairement différentes entre un mois d'hiver et un mois de plein été, même si les installations expérimentales du laboratoire d'acoustique bénéficient d'une relative stabilité des conditions thermique et hygrométrique grâce à sa situation au second sous-sol du Cnam. Pour finir, l'ensemble du mobilier (table et station de pilotage de la SpherBedev) a également été déplacé à l'opposé dans la salle, toujours à proximité de la sphère de haut-parleurs. Ces éléments sont des surfaces réfléchissantes et potentiellement diffractantes, ce qui modifie donc fondamentalement les conditions expérimentales entre l'apprentissage et l'inférence. Par ailleurs, les données utilisées pour cette comparaison possèdent des caractéristiques différentes de celles utilisées pour l'entraînement du réseau, ce qui est volontairement un cas non-idéal pour l'approche BeamLearning par rapport à SH-MUSIC :

— tandis que l'entraînement du réseau a été réalisé grâce à des synthèses ambisoniques d'ondes planes, les sources à localiser pour cette validation expérimentale et confrontation à l'algorithme SH-MUSIC sont des ondes quasi-sphériques, provenant des 50 haut-parleurs individuels du

dispositif placés à une distance de 1,07 m de l'antenne Zylia, sans utilisation de synthèse de champ.

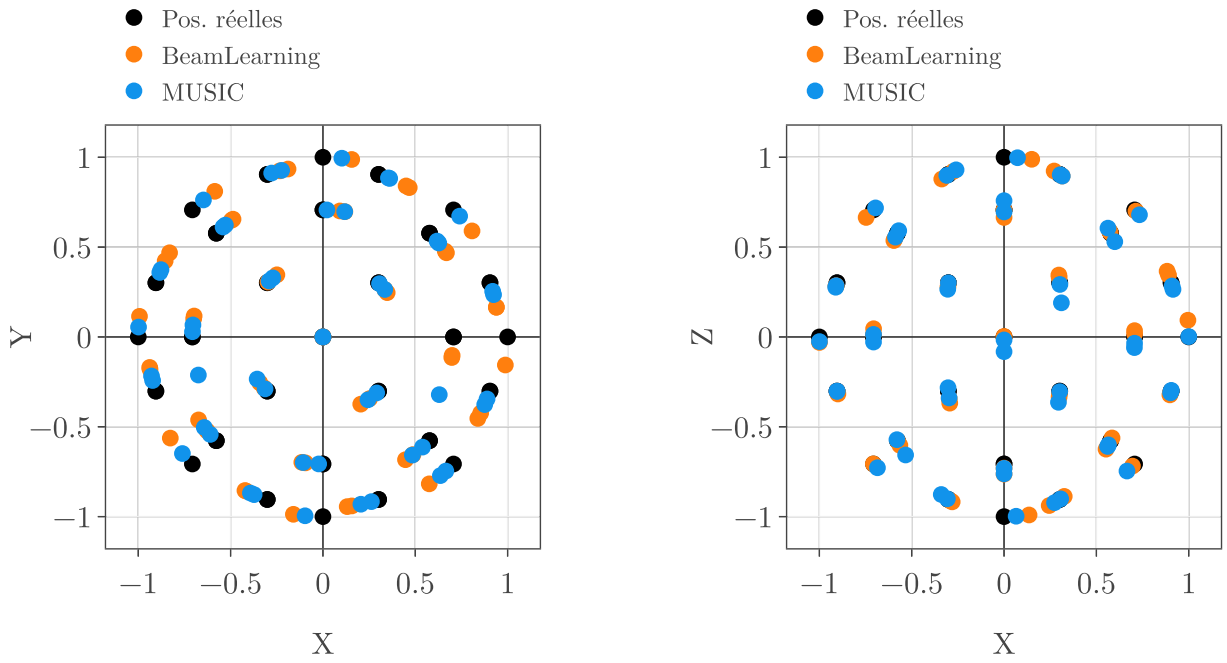
— tandis que l'entraînement du réseau a été réalisé à l'aide d'un enregistrement de type *cocktail party* de voix féminines en danois, les signaux émis par les 50 sources à localiser pour cette comparaison sont des enregistrements de voix masculines de personnels du laboratoire, en français.

— entre les mesures exploitées pour l'entraînement au mois de Février et les mesures utilisées pour la validation au mois de Juillet, l'antenne Zylia a bien entendu été retirée du centre de la sphère de spatialisation pendant toute la phase de travaux et de traitement acoustique de la salle, pour être positionnée à nouveau. Son positionnement est donc légèrement différent.

Pour chacune des 50 positions de sources (haut-parleurs Aurasound NS-W2), le champ de pression a été mesuré à l'aide des 19 voies microphoniques de l'antenne Zylia. Grâce à ces mesures, pour chacune des sources à localiser, 300 trames de 1 024 échantillons consécutives sont sélectionnées, permettant ainsi de réaliser 300 estimations indépendantes de la position de la source par les deux méthodes. En ce qui concerne la méthode SH-MUSIC, qui repose en premier lieu sur un encodage ambisonique à l'ordre 3 du champ mesuré, les 5 dernières trames de 1 024 échantillons ont été écartées de l'estimation pour chacune des 50 sources, puisque l'encodage ambisonique repose sur l'utilisation d'une transformée de Fourier directe et inverse sur des fenêtres de 8 192 échantillons. L'analyse des performances de localisation par la méthode SH-MUSIC a révélé que l'encodage ambisonique de la dernière fenêtre était incorrect, ce qui menait à des fortes erreurs de localisation qui auraient été inutilement défavorables à la méthode SH-MUSIC. Nous avons donc décidé d'écarter ces 5 dernières trames pour conserver une comparaison équitable entre les deux méthodes. Par conséquent, la méthode SH-MUSIC ne permet que 295 estimations de positions pour chaque source, ce qui n'impacte pas statistiquement la validité de l'analyse qui suit.

Après inférence de l'ensemble de ces positions de sources grâce à l'approche BeamLearning et à l'algorithme SH-MUSIC, on peut dans un premier temps calculer une valeur moyenne des estimations de positions 3D de chacun des 50 haut-parleurs pour ces deux méthodes. L'analyse de ces données moyennes a révélé que, malgré l'effort de positionnement au laser de l'antenne Zylia au centre de

l'antenne, il subsistait un léger biais de positionnement absolu de l'antenne par rapport au repère dans lequel sont définies les positions des haut-parleurs sur la structure de la Spherbedev. La figure 5.23 illustre cette observation pour la méthode SH-MUSIC (en bleu clair) et pour l'approche BeamLearning (en orange), et permet de les comparer à la position des 50 haut-parleurs dans le repère de la Spherbedev (en bleu).



(a) Mise en évidence du biais de l'angle azimutal du à une rotation de l'axe de l'antenne par rapport à l'axe de la sphère de spatialisation

(b) Mise en évidence du biais en élévation du à un placement du plan de l'équateur de l'antenne plus bas que le plan de l'équateur de la sphère de spatialisation

FIGURE 5.23 – Mise en évidence du biais d'estimation en azimut (a) et en élévation (b) du à une erreur de centrage de l'antenne Zylia dans la sphère de spatialisation

Pour mettre en évidence ce biais de positionnement en azimut, la figure 5.23(a) représente les estimations moyennes des positions de tous les haut-parleurs, projetées dans le plan équatorial (xOy), en supposant une estimation parfaite de l'angle d'élévation : en notant $\tilde{\theta}_i$ et $\tilde{\phi}_i$ la moyenne des positions azimutales et en élévation, estimées pour le haut-parleur i, et $(\theta_i; \phi_i)$ la position réelle du haut parleur, on trace pour chacun des haut-parleurs et pour les deux méthodes de localisation les positions suivantes :

$$\begin{cases} x_i = \cos\left(\tilde{\theta}_i\right) \times \cos\left(\phi_i\right) \\ y_i = \sin\left(\tilde{\theta}_i\right) \times \cos\left(\phi_i\right) \end{cases} \tag{5.2}$$

De même, pour mettre en avant dans la figure 5.23(b) le biais dans l'estimation de l'élévation des haut-parleurs, les positions sont projetées dans les plans $(x0z)$, en supposant exacte l'estimation azimutale :

$$\begin{cases} x_i = \cos\left(\theta_i\right) \times \cos\left(\tilde{\phi}_i\right) \\ z_i = \sin\left(\tilde{\phi}_i\right) \end{cases} \tag{5.3}$$

L'analyse de la figure 5.23(a) révèle que tous les points moyens d'estimation de position obtenus par la méthode SH-MUSIC sont biaisés par un angle β_1, alors que cette méthode ne présente théoriquement pas de biais lorsque l'antenne utilisée est compacte et que le critère d'échantillonnage spatial du champ est respecté pour l'encodage ambisonique, comme c'est le cas ici [203]. Par ailleurs, l'analyse de toutes les inférences réalisées à l'aide de BeamLearning au cours de cette thèse a également révélé que l'estimation d'azimut et d'élévation était sans biais, et que les erreurs étaient statistiquement centrées.

Compte tenu de ces deux observations, il apparaît évident que le repère local de l'antenne Zylia et le repère de la SpherBedev n'étaient pas parfaitement alignés au cours de la mesure. Par ailleurs, l'analyse des résultats présentés sur la figure 5.23(b) révèle que le biais en élévation lié au positionnement de l'antenne Zylia est nettement plus faible (inférieur à $0,5°$), puisque l'alignement avec le plan équatorial de la SpherBedev a été réalisé grâce à un niveau laser. D'un point de vue pratique, cette observation est parfaitement concordante avec le fait qu'il est beaucoup plus facile de rendre coplanaires les plans équatoriaux de l'antenne Zylia et de la sphère de spatialisation que d'aligner les méridiens de l'antenne et de la sphère de spatialisation. Par ailleurs, on peut remarquer que le biais de positionnement n'est pas le même pour la méthode SH-MUSIC et pour la méthode BeamLearning.

En effet, pour la méthode SH-MUSIC, le biais provient du défaut de positionnement de l'antenne pour l'expérience d'inférence, tandis que pour la méthode BeamLearning, le biais observé provient de l'écart entre le défaut de positionnement de l'antenne pendant la phase d'inférence et du défaut de positionnement de l'antenne pendant la phase d'enregistrement de la base de données d'entraînement. Ce point pratique de positionnement de l'antenne devra donc faire l'objet d'une amélioration à la suite de cette thèse de doctorat, même si elle n'est absolument pas limitante pour l'analyse qui nous intéresse ici. En effet, la précision des deux méthodes, et l'absence de biais théorique pour chacune d'entre elles, couplée à la très grande homogénéité des biais observés pour les 50 haut-parleurs, permettent de compenser aisément et précisément ce défaut de positionnement qui ne provient pas des méthodes de localisation, mais du protocole expérimental d'alignement 3D de deux référentiels.

Une fois ces biais corrigés, la précision des deux approches peut alors être comparée finement. Le tableau 5.12 résume ces performances de localisation, en présentant la moyenne, la médiane, et l'écart-type des erreurs angulaires absolues d'estimation réalisées par les deux méthodes, pour les 15 000 (resp. 14 750 pour la méthode SH-MUSIC) estimations de positions à partir de trames de 1 024 échantillons réalisées au cours de cette expérience.

Approche	Moyenne absolue (°)	Médiane absolue (°)	Écart type (°)	Temps de calcul CPU	Temps de calcul GPU
SH-MUSIC	4,40	3,50	7,03	1h53	
BeamLearning	**3,88**	**3,22**	**2,92**	**11min21**	**1min35**

Tableau 5.12 – Performances de localisation (erreur angulaire solide en °) des algorithmes MUSIC et de l'approche BeamLearning sur respectivement 295 et 300 estimations de position de chacun des 50 haut-parleurs de la sphère de spatialisation

L'analyse de ces résultats synthétiques révèle que malgré le changement d'acoustique de la pièce, le changement de contenu de la source et le changement de la nature même de la source, et sans exploitation explicite d'une projection ambisonique des données offerte par la géométrie de l'antenne, l'approche BeamLearning permet une localisation plus précise, mais aussi plus robuste que l'algorithme SH-MUSIC. En effet, même si la différence en terme d'erreur moyenne absolue n'est pas démesurée

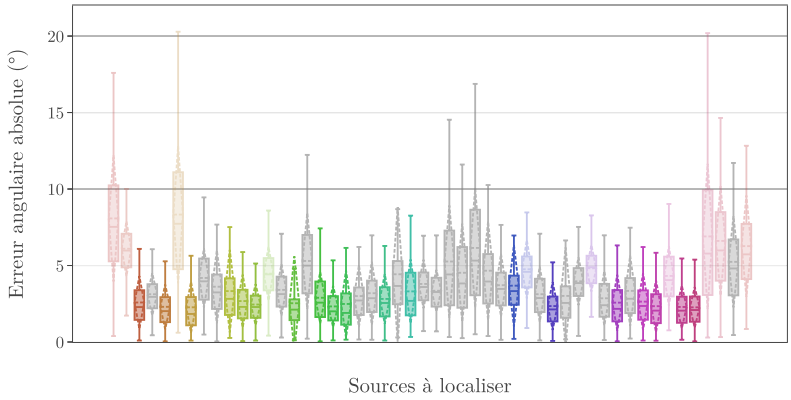
$(0,5°)$ et que les deux méthodes offrent toutes deux de très bonnes performances de localisation, il est particulièrement notable que la différence entre les écarts types des deux approches est relativement élevée. En effet, le fait que l'écart type de SH-MUSIC soit 4° plus élevé que celui de l'approche par BeamLearning ($7,03°$ contre $2,92°$), indique que l'estimation fournie par SH-MUSIC contient beaucoup plus de *grosses* erreurs de localisation. Enfin, pour calculer l'ensemble des 15 000 positions (14 750 dans le cas de SH-MUSIC), il a fallu 1h53 pour l'algorithme SH-MUSIC, contre seulement 1min35 dans le cas de l'approche par BeamLearning lorsqu'elle est implémentée sur GPU. Mais même lorsque l'approche proposée est implémentée sur CPU comme la méthode SH-MUSIC, le gain de temps reste très significatif. Cette énorme réduction de temps de calcul prouve que l'apport de l'approche proposée n'est pas seulement quantifiable en terme de précision angulaire, mais également en terme de temps de calcul.

La démarche utilisée pour cette comparaison permet par ailleurs d'analyser plus finement cette différence de comportement au delà de la simple analyse des erreurs moyennes sur l'ensemble des estimations. En effet, puisque la position des 50 sources a été estimée 300 (resp. 295) fois par chacune des deux méthodes, il est possible de mener une analyse statistique rigoureuse sur les erreurs angulaires commises par les deux méthodes.

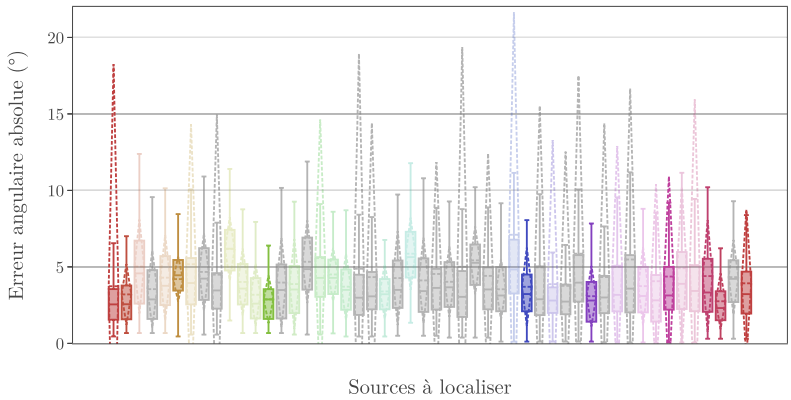
En particulier, une analyse statistique des erreurs a été menée pour chacun des haut-parleurs, afin de déterminer laquelle des 2 méthodes est plus *précise*, en comparant les ensembles statistiques de 295 à 300 points par méthode. Puisque les répartitions des erreurs ne sont pas gaussiennes mais particulièrement asymétriques, c'est donc vers le test non paramétrique de Wilcoxon-Mann-Whitney que nous nous sommes tournés, puisqu'il est conçu pour fournir un résultat pertinent dans ce cas [204]. L'objectif de ce test est dans un premier temps de déterminer si les erreurs angulaires commises par la méthode SH-MUSIC et la méthode BeamLearning sont statistiquement différentes. Pour 21 haut-parleurs parmi les 50 localisés (42% des sources), les deux méthodes offrent des performances de localisation similaires (valeur de p inférieure à 5×10^{-5}).

Pour les 29 haut-parleurs restants où les différences de distributions d'erreurs de localisation sont statistiquement représentatives, le test de Wilcoxon-Mann-Whitney a de nouveau été utilisé, afin de

déterminer quelle méthode offre les meilleures performances de localisation. À l'issue de ce test, on observe que 19 haut-parleurs ont été mieux localisés par la méthode BeamLearning (38 % des sources), et que seulement 10 haut-parleurs (20 % des sources) ont été mieux localisés par la méthode SH-MUSIC. Dans tous les cas, l'indice p indiquant la confiance statistique dans ce classement est inférieur à 5×10^{-5}.



(a) Analyse statistique *box and whisker* des erreurs angulaires de localisation 3D avec la méthode BeamLearning pour les 50 haut-parleurs testés (300 estimations de position par haut-parleur)



(b) Analyse statistique *box and whisker* des erreurs angulaires de localisation 3D avec la méthode SH-MUSIC pour les 50 haut-parleurs testés (295 estimations de position par haut-parleur)

FIGURE 5.24 – Représentation *box and whisker* des erreurs angulaires de localisation pour les 50 haut-parleurs, pour la méthode BeamLearning proposée et la méthode SH-MUSIC avec l'antenne Zylia. Les haut-parleurs en couleur franche sont ceux pour lesquels l'erreur angulaire est la plus faible par rapport à l'autre méthode, de manière représentative statistiquement. Les haut-parleurs pour lesquels les deux méthodes obtiennent des erreurs similaires sont représentés en gris. Les haut-parleurs pour lesquels la méthode testée est moins bonne que l'autre sont représentés en couleurs pastel.

La figure 5.24 permet de mieux comprendre ce classement obtenu avec la rigueur statistique nécessaire. Sur cette figure, les haut-parleurs pour lesquels aucune des deux méthodes n'est *meilleure* d'un point de vue statistique sont représentés en gris. Pour chacune des deux méthodes (BeamLearning en 5.24(a) et SH-MUSIC en 5.24(b)), les haut-parleurs ayant été mieux localisés par la méthode sont représentés en couleur franche, et les haut-parleurs pour lesquels l'autre méthode est plus précise sont représentés en couleur pastel.

Sur la figure 5.24, pour chacune des 50 sources, un représentation de type *box and whisker* est proposée, ce qui permet d'offrir une représentation compacte des erreurs commises sur l'ensemble des localisations par trames de 1 024 échantillons [205]. Sur cette représentation, chacune des boites (box) colorées représente l'ensemble des erreurs dans la gamme interquartile, définie par les données entre le 25^e percentile et le 75^e percentile des erreurs d'estimations angulaires. La médiane des erreurs est quant à elle représentée à l'intérieur de chacune de ces boites par une ligne horizontale, permettant d'observer que les répartitions statistiques sont plutôt concentrées vers des faibles valeurs, mais qu'il existe des estimations ponctuelles présentant de fortes erreurs d'estimation. Les *moustaches* représentent quant à elle l'intervalle défini par les données appartenant à la gamme du 9^e percentile au 91^e percentile, permettant de mieux percevoir la répartition des erreurs. En dehors de ces valeurs, les *outliers* ne sont pas tracés, mais il existe des estimations excédant la hauteur des moustaches. Par conséquent, pour chacune des boites, l'intervalle représentant l'écart-type des erreurs est représenté en flèches pointillées centrées sur la moyenne des erreurs angulaires pour chaque source à localiser.

Cette représentation compacte permet ainsi d'observer des tendances globales : pour la méthode BeamLearning, pour chacune des sources, la moustache inférieure représentant les données entre le 9^e quantile et le 25^e quantile atteint systématiquement une valeur très basse, ce qui signifie que pour un grand nombre de trames, la méthode BeamLearning offre des estimations avec très peu d'erreur absolue, y compris pour les sources possédant une valeur de médiane plus élevée que celle obtenue par la méthode SH-MUSIC.

Par ailleurs, on observe quasi-systématiquement des écarts-types d'erreurs commises beaucoup plus élevés avec la méthode SH-MUSIC qu'avec la méthode BeamLearning, ce qui signifie que le nombre

d'erreurs d'estimation angulaires élevées est plus important pour SH-MUSIC, même si dans l'ensemble les estimations de positions par trames mènent à une estimation globalement satisfaisante. On peut suspecter que la cause de ces erreurs ponctuelles élevées avec la méthode SH-MUSIC peut être liée à des réflexions précoces résiduelles dans la salle, qui sont connues pour dégrader les performances de localisation des méthodes de séparation en sous-espaces.

En revanche, la position centrale des boites (la médiane des erreurs angulaires pour chaque haut-parleur) est plus homogène pour l'ensemble des sources avec la méthode SH-MUSIC qu'avec l'approche BeamLearning. On observe d'ailleurs que les seules valeurs pour lesquelles la méthode SH-MUSIC l'emporte en termes de performances correspond aux quelques haut-parleurs qui s'écartent de la tendance globale offerte par BeamLearning. Il est intéressant de noter que parmi eux, on retrouve les haut-parleurs aux pôles de la sphère de spatialisation, et des haut-parleurs à proximité de la station de pilotage et de l'écran lors de la phase d'apprentissage (déplacés pendant la phase d'inférence). Cette observation tend à montrer que la modification des éléments interagissant avec le champ de pression à proximité des haut-parleurs ne permet plus au réseau, qui a appris à compenser leur présence pour la localisation, d'être aussi précis que pour d'autres positions de sources dans la salle (ce qui n'est évidemment pas le cas pour la méthode SH-MUSIC, qui est ici dans un cas plutôt idéal, avec une salle pour cette mesure qui est bien traitée acoustiquement, sans paroi réfléchissante de manière franche).

Grâce à cette analyse statistique fine, on peut alors confirmer que la méthode BeamLearning offre globalement de meilleures performances, à l'exception ponctuelle de sources pour lesquels la modification de la salle de mesure a fait perdre en précision l'inférence de la localisation de sources.

Pour offrir une vision plus globale de la comparaison entre les deux méthodes que celle proposée sur la figure 5.24, une analyse statistique similaire a été menée, cette fois sur l'ensemble des 300x50 (resp. 295x50) estimations de position, de manière à voir si les différences de performances récapitulées dans le tableau 5.12 sont statistiquement représentatives. Un test de Wilconxon-Mann-Whitney a donc été mené sur ces deux populations avec les mêmes paramètres que pour les 50 haut-parleurs individuels, permettant de conclure avec une forte confiance ($p = 2.2 \times 10^{-19}$) que l'amélioration de $0,5$ degrés environ offerte par la méthode BeamLearning par rapport à SH-MUSIC est statistiquement représentative.

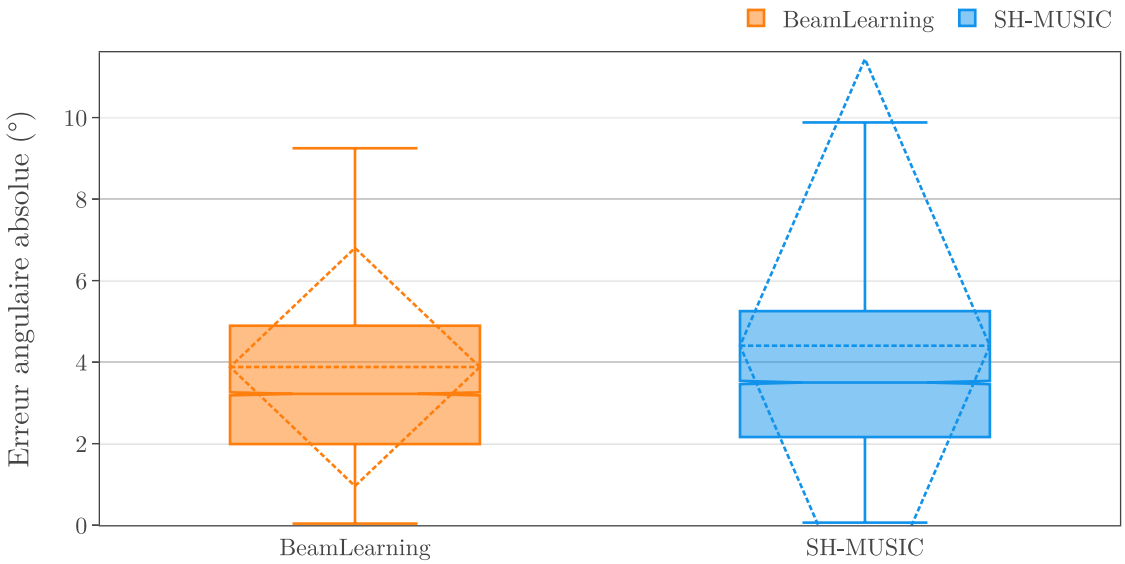


FIGURE 5.25 – Comparaison statistique des erreurs commises pour l'ensemble des trames et des haut-parleurs pour la méthode BeamLearning (1 5000 estimations de position) et la méthode SH-MUSIC (1 4750 estimations de position) par représentation de *type box and whisker*

La figure 5.25 représente graphiquement les données sous-jacentes à l'aide d'un graphique *box-and-whisker* du même type que celle utilisée à la figure 5.24 pour chacun des haut-parleurs. L'analyse de la figure 5.25 met à nouveau en lumière le fait que la répartition des erreurs offertes par BeamLearning est beaucoup plus homogène autour de la valeur médiane, et que la méthode SH-MUSIC, même si elle offre de très bonnes performances dans l'ensemble, est impactée par un nombre non négligeable d'estimations de positions à fortes valeurs d'erreur angulaire, menant à un écart-type plus élevé.

5.3.7 Synthèse des principaux résultats obtenus grâce à l'approche par BeamLearning

L'approche de localisation de sources proposée, le BeamLearning, est testée dans ce chapitre face à différentes situations issues de simulations numériques et d'expériences, pour une tâche de localisation en 2D ou en 3D, avec des approches de classification ou de régression pour le paradigme d'apprentissage. Les signaux émis par les sources à localiser sont variés et vont de simples signaux monochromatiques sans aucun bruits de fond, jusqu'à la localisation de voix humaines ou de sources

émettant des signaux musicaux dans le bruit. Enfin, les environnements proposés sont eux aussi variés puisque la localisation de sources acoustiques est proposée aussi bien en champ libre qu'en environnement réverbérant.

Un des avantages de cette approche par apprentissage supervisé est son caractère modulaire. Quelque soit l'environnement, la source ou l'antenne, l'approche et la topologie du réseau de neurone profond exploité est strictement identique. L'algorithme de Deep Learning proposé est utilisable pour toutes ces situations, et aucune précaution particulière ne doit être prise lors d'un changement de situation.

Un autre avantage de l'approche BeamLearning est sa robustesse. Cette robustesse s'acquiert à partir de la bonne constitution d'un jeu de données d'apprentissage : pour être robuste aux variations de contenu spectral des sources à localiser, le jeux de données d'apprentissage doivent être variés d'un point de vue fréquentiel ; pour être robuste au bruit de fond, l'optimisation des variables d'apprentissage du réseau de neurones doit être faite avec des données bruitées ; enfin, malgré un changement de traitement acoustique de la salle entre l'apprentissage et la validation, la localisation est suffisamment robuste pour obtenir de meilleures performances globales de localisation que l'algorithme SH-MUSIC, pourtant optimisé pour son utilisation avec l'antenne ambisonique Zylia testée ici.

Comme nos développements reposent sur la généralisation par le réseau de neurones profond de caractéristiques communes présentes dans les données brutes captées par des antennes microphoniques, l'approche BeamLearning permet aussi un étalonnage implicite de l'antenne, lorsque les données utilisées pour la phase d'apprentissage sont des données mesurées. De plus, toutes les caractéristiques physiques de l'antenne, depuis la position de chacun de ses microphones jusqu'à la diffraction de l'antenne et de son support sont inclus lors de la captation des données. Toutes ces informations, généralement négligées dans les approches *modèles*, sont au contraire autant d'informations pertinentes pour le réseau de neurones, et l'aident dans la tâche de localisation de sources acoustiques.

L'analyse des résultats obtenus grâce à la méthode proposée a prouvé que la précision de localisation obtenue lors des différents tests est, dans la grande majorité des cas, meilleure que les approches

modèles testées (MUSIC et SRP-PHAT). Par exemple, dans le cas de DOA à deux dimensions dans un environnement réverbérant simulé, avec un RSB de 5 dB, l'approche BeamLearning obtient une précision angulaire absolue de $10,8°$, soit $4,8°$ de mieux que MUSIC et $7,6°$ de mieux que SRP-PHAT. Dans le cas de la DOA en trois dimensions avec des captations expérimentales sans bruit de fond rajouté numériquement, l'approche BeamLearning fait légèrement mieux que SH-MUSIC ($3,88°$ contre $4,40°$).

Enfin, le temps de calcul nécessaire pour déterminer des positions de sources à partir de trames de 1 024 échantillons captées sur une antenne microphonique dans le cas à 3 dimensions est en moyenne 73 fois plus rapide avec l'approche proposée qu'avec l'approche SH-MUSIC, ce qui ouvre la voie à une localisation en temps réel et au suivi de sources mobiles, mais également à un couplage avec des méthodes de reconnaissance de signatures acoustiques par exemple.

Conclusions et perspectives

Synthèse sur l'approche de localisation de sources par BeamLearning

La localisation de sources acoustiques est un défi, qui trouve des applications dans un grand nombre de domaines. Depuis l'industrie, pour la détection de défauts dans un mécanisme, jusqu'à la vie quotidienne, où les assistants vocaux commencent à exploiter la position du locuteur pour filtrer spatialement le signal, et en extraire nos requêtes. Jusqu'à présent, de nombreuses méthodes, qui s'appuient sur des modèles de propagation acoustique ou des modèles de signaux, sont disponibles dans la littérature. Mais il est parfois difficile de trouver l'algorithme approprié à la situation de mesure, tant leur variété est importante. Il est donc nécessaire d'avoir une bonne connaissance de l'ensemble de ces algorithmes, pour choisir, en fonction des hypothèses faites, et lequel sera le plus approprié suivant la situation de mesure.

L'approche BeamLearning développée lors de cette thèse de doctorat n'a pas vocation à remplacer les méthodes conventionnelles de localisation de sources acoustiques, qui offrent des performances de qualité, mais elle vise plutôt à explorer un nouveau paradigme, celui de l'apprentissage supervisé , qui pourrait se révéler être porteur d'innovations dans ce domaine dans les années à venir. L'utilisation de ce genre de méthodes est très récente pour la localisation de sources, mais depuis quelques années, l'essor des techniques de Deep Learning ont permis d'énormes avancées dans d'autres domaines, comme celui de la vision assistée par ordinateur, ou de la reconnaissance vocale. Cette thèse de doctorat propose ainsi de redéfinir le problème de localisation de sources acoustiques, et de l'aborder à travers une approche centrée sur les *données*.

La manière d'appréhender ce problème d'apprentissage supervisé, lors de cette thèse de doctorat, a été de travailler sur trois axes interconnectés. Tout d'abord, le cahier des charges doit être défini, pour savoir ce qui est attendu dans le cadre de la localisation de sources, tant au niveau de l'environnement, que de l'antenne utilisée ou de la précision nécessaire de l'estimation de la position de la source. Ensuite, une base de données doit être constituée, et doit répondre à différents critères afin de pouvoir être utilisée à des fins d'apprentissage : elle doit contenir un nombre important d'exemples, et ces exemples doivent être suffisamment représentatifs de la situation définie par le cahier des charges. Enfin, une architecture de réseau de neurones doit être développée pour être à même de généraliser les caractéristiques communes aux données, et de résoudre le problème de localisation de sources acoustiques.

Le réseau de neurones utilisé pour l'approche BeamLearning

Le réseau proposé pour l'approche du BeamLearning est un réseau de neurones profond dont la topologie s'articule autour de quatre sous parties, qui ont chacune été conçues au cours de cette thèse pour leur bonne adéquation entre la physique du problème à résoudre, les grandeurs calculées par les approches *modèles*, et les couches neuronales communément utilisées dans le domaine du Deep Learning. Ces quatre sous parties du réseau sont : l'entrée du réseau, une cascade de bancs de filtres, un calcul d'énergie et la sortie du réseau.

L'entrée du réseau de neurones est constituée des données brutes provenant de la pression mesurée par les capteurs de l'antenne microphonique. Aucun traitement particulier n'est fait sur ces données, si ce n'est un filtrage passe bas, pour des raisons expérimentales. Le réseau de neurones doit donc trouver quelles caractéristiques du signal sont pertinentes avant de les extraire. C'est ce qu'on appelle le *joint feature learning*.

Viennent ensuite les bancs de filtres mis en cascade, qui sont obtenus à partir de convolutions résiduelles à trous séparables en profondeur. Le caractère multi-échelle de ces convolutions présente l'intérêt de pouvoir extraire de l'information à différentes fréquence, tout en gardant un nombre limité de coefficients à optimiser. Ces convolutions sont entrecoupées de non linéarités obtenues grâce à la fonction tangente hyperbolique, ce qui permet de conserver des données centrées sur 0, comme l'est la pression acoustique. Enfin, pour assurer la stabilité des performances de localisation face à une variabilité d'amplitude des signaux d'entrée, les variables d'apprentissage sont normalisées.

Puis un calcul déterministe pseudo-énergétique est effectué. La fonction de cette partie du réseau de neurones est de convertir les données, jusque là centrées sur 0, en valeurs quadratiques moyennes. Cette transformation est usuelle dans les algorithmes classiques de localisation de sources, qui cherchent souvent à maximiser ou minimiser l'énergie d'un signal provenant d'une direction donnée.

Enfin, l'approche BeamLearning peut être utilisée de manière équivalente pour une représentation continue ou par secteurs angulaires la position de la source, grâce à une méthode de classification ou de régression. Dans le premier cas, l'espace dans lequel la source est susceptible de se trouver est subdivisé en plusieurs sous-espaces. Cette méthode d'estimation offre l'avantage de faire converger l'approche BeamLearning très rapidement. En revanche, la précision angulaire est directement liée au nombre de zones qui divisent l'espace où les sources sont recherchées. Au contraire, la régression permet d'obtenir

une valeur chiffrée du ou des angles estimés. La régression, en contrepartie d'un temps d'apprentissage plus long, permet de d'affiner énormément la précision de localisation.

Le réseau de neurones ainsi constitué contient plusieurs millions de variables à optimiser au cours de la phase d'entraînement. Pour cela, un grand nombre d'exemples de données de pression multicanales mesurées par l'antenne à laquelle on adjoint l'intelligence artificielle sont présentés au réseau, qui estime à partir de ces données d'entrée, la position de la source qui a émis ce signal. L'erreur d'estimation commise permet alors d'ajuster la valeurs de ces variables d'apprentissage, jusqu'à minimiser statistiquement l'erreur d'estimation de position faite. La qualité de ces données est gage de la bonne estimation, par le réseau de neurones, dans de nouvelles situations.

Base de données simulées ou expérimentales, deux approches complémentaires

Contrairement à un grand nombre de domaines où l'apprentissage supervisé est appliqué, dans le cas de la localisation de sources acoustiques, le problème physique de propagation de sources depuis une position jusqu'à une antenne quelconque, dans un environnement réverbérant, est calculable sans altérer trop fortement le réalisme des données. Cette caractéristique permet d'envisager la création de jeux de données simulées pour n'importe quelle situation donnée. De plus, le laboratoire possède un spatialisateur 3D développé lors d'une précédente thèse de doctorat, qui permet de synthétiser physiquement des fronts d'ondes quelconques pour simuler expérimentalement de manière réaliste la propagation du champ émanant d'une source, depuis n'importe quelle position théorique jusqu'au centre du spatialisateur. Ainsi, la création automatisée d'un jeux de données est aussi possible, pour des données mesurées depuis une antenne microphonique.

Dans le cadre de cette thèse, nous avons proposé des solutions pour constituer des jeux de données simulées grâce à des réponses impulsionnelles de salles. Ces réponses impulsionnelles sont calculées grâce à la méthode des sources images, qui a bénéficié de deux développements particuliers lors de cette thèse de doctorat. Tout d'abord, pour gagner en précision temporelle, les signaux peuvent être retardés d'un nombre non entiers d'échantillons au moyen de retards fractionnaires, obtenus par interpolation grâce aux polynômes de Lagrange que nous avons implémenté sur GPU pour accélérer le temps de création de centaines de milliers de réponses impulsionnelles. Puis, pour palier les dérives temporelles des durées de réverbération obtenues par la méthode des sources images, un dénombrement des sources images est proposé afin d'obtenir rapidement un coefficient d'absorption modifié, plutôt que de l'obtenir

par l'inversion de formules issues de l'acoustique statistique. Ce coefficient d'absorption modifié permet ainsi d'obtenir grâce aux sources images une durée de réverbération cible, et donc de simuler au plus proche une pièce en particulier.

Des jeux de données expérimentales ont aussi été constitués à partir de la sphère de spatialisation du LMSSC. Grâce à cette sphère de spatialisation, utilisant le formalisme ambisonique jusqu'à l'ordre 5, la pression mesurée au centre de la sphère de spatialisation est assurée d'être physiquement conforme à la pression cible sur une gamme fréquentielle étendue. Les données ainsi captées par des antennes de microphones intègrent donc naturellement toutes les caractéristiques physiques de l'antenne, depuis les réponses en fréquences propres à chaque microphone, jusqu'à la diffraction du corps et du pied de l'antenne. Ces données ainsi enregistrées permettent un étalonnage implicite de l'antenne lors de la phase d'apprentissage. Cette approche de spatialisation grâce à la *SpherBedev* permet en outre une automatisation complète des mesures, ainsi que la sauvegarde de toutes les caractéristiques de la mesure.

Toutes les caractéristiques de ces exemples sont sauvegardées dans une base de données au format JSON, ce qui permet une navigation rapide et efficace entre ces centaines de milliers d'exemples. Ainsi, plusieurs situations ont pu être testées afin d'apprécier les performances de l'approche BeamLearning et de les confronter à des méthodes conventionnelles reconnues pour leur efficacité.

Validation dans différentes situations expérimentales et numériques

Pour accompagner le développement de l'approche BeamLearning, plusieurs *scenari* de localisation ont été développés. Le cahier des charges est devenu de plus en plus exigeant, pour passer de la classification de la position de signaux mono-fréquentiels en champ libre, à la localisation, par régression, de sources au contenu spectral varié, en trois dimensions, dans un environnement réverbérant. Plusieurs tendances globales peuvent alors être tirées sur cette approche à partir des résultats obtenus au cours de cette thèse.

Le choix de la représentation de la position de la source peut être fait, soit par classification, soit par régression. Lorsque le choix se porte sur la classification, l'espace de recherche de la source est divisé en plusieurs sous-espaces. Une probabilité d'appartenance à chaque sous-espace est alors attribué par le réseau de neurones, et permet de déterminer la position de la source. Dans ce cas, la précision angulaire dépend exclusivement du nombre de classes (donc de sous-espaces) défini par

l'utilisateur. Or l'espace dans lequel évolue la source à localiser est continu. Ainsi, le découper en sous-espaces introduit des ambiguïtés au niveau des frontières de ces sous-espaces. Il n'est donc pas pertinent de faire, pour une DOA en 2D, plus d'une centaine de classes. En revanche, si la précision angulaire attendue n'est pas très importante, comme dans le cas de la localisation de locuteur, cette approche par classification ne nécessite pas un temps d'apprentissage trop long (quelques minutes, pour de la DOA en 2D en champ libre). Au contraire, l'approche par régression nécessite un temps d'apprentissage plus long (quasiment 50 fois plus d'itérations, et environ 2h pour de la DOA en 2D en champ libre), mais elle offre bien entendu une précision angulaire beaucoup plus fine.

La précision angulaire de l'approche BeamLearning dans le cas de la régression est *a minima* équivalente et surpasse dans la plupart des cas celle des algorithmes MUSIC ou SRP-PHAT, dans les situations testées. Par exemple, dans le cas de la DOA à 2D de bruit de *cocktail party* dans une salle à la durée de réverbération de 0,5 s, les performances de l'approche BeamLearning dépassent assez largement celles des approches modèles ($10,8°$ contre $15,6°$ pour MUSIC et $18,4°$ pour SRP-PHAT à $RSB = 5\ dB$). De même, dans le cas de DOA à 3D, malgré un changement de disposition de la salle et de son traitement acoustique entre la phase d'apprentissage et la phase de validation, l'approche BeamLearning offre une précision légèrement meilleure que l'algorithme SH-MUSIC, et les estimations sont beaucoup plus concordantes pour une même source lorsqu'elles sont effectuées sur des trames de 1024 échantillons (l'écart type de l'erreur valant $2,92°$ pour l'approche BeamLearning et $7,03°$ pour l'algorithme SH-MUSIC).

Enfin, le temps de calcul nécessaire à l'estimation des positions de sources sont très largement en faveur de notre approche par rapport aux méthodes de localisation de sources conventionnelles testées, en partie grâce à la puissance de calcul offerte par le calcul sur GPU pour l'inférence. Alors qu'il faut 1min35 avec l'approche BeamLearning pour estimer la position de 15 000 sources, il faut près de 2h à l'algorithme SH-MUSIC pour le même nombre de sources. Ainsi, l'approche BeamLearning permet une localisation en temps réel des sources, ce qui permet d'envisager le couplage de cette approche avec d'autres fonctionnalités, comme le suivi de trajectoire ou la reconnaissance de signature acoustique.

Perspectives de développements complémentaires

Les cas de figures proposés dans ce manuscrit ont mis en exergue que la qualité de généralisation du réseau de neurones était intimement liée à la diversité des exemples présentés lors de la phase d'optimisation des variables d'apprentissages : pour être plus robuste face aux fréquences du signal, il faut une diversité fréquentielle de signaux ; pour être robuste au bruit de mesure, il faut que les exemples d'apprentissages soient bruités. Ainsi, pour améliorer la robustesse à la géométrie de la salle, augmenter le nombre d'exemples de la base de données de salle serait pertinent, et pourrait faire l'objet de développements futurs.

Un des aspects de la localisation de sources, qui n'a pas été du tout abordé durant cette thèse de doctorat, et qui serait une plus value significative à l'approche BeamLearning, est la localisation simultanée de sources multiples. Dans la littérature, certains algorithmes, tant basés sur des modèles de propagation que sur l'apprentissage supervisé, offrent la possibilité de localiser un nombre connu de sources. Or dans la plupart des cas, le nombre de sources n'est pas connu *a priori* lors de la localisation de ces sources. Ainsi, un axe de développement de l'approche BeamLearning, serait l'estimation jointe du nombre de sources présentes dans la scène sonore, ainsi que de leurs positions respectives. Pour ce faire, il serait pertinent, à l'instar de certaines méthodes d'apprentissages supervisées, de faire travailler conjointement deux algorithmes, et que le résultat du premier, estimant le nombre de sources présentes, serve d'entrée pour l'algorithme servant à la localisation de ces sources.

Une autre extension applicative de l'approche, rendue possible par la vitesse de traitement des informations par le réseau de neurones profond, est le suivi de sources mobiles. Pour ce faire, plusieurs pistes peuvent être suivies, à commencer par une estimation statistique de la position de la source, à partir de plusieurs trames de signal. Cette estimation peut être faite au moyen d'une simple moyenne, ou d'estimateurs plus évolués, comme le filtre de Kalman. Ce type d'estimation permet d'ajouter une *mémoire* au réseau de neurones, et ainsi de stabiliser l'estimation d'une nouvelle positions en fonction des positions précédentes.

Enfin, la localisation de sources acoustiques pourrait être couplée à un algorithme de reconnaissance de signature acoustique, pour estimer conjointement la position et la nature de la source. Cette association a déjà été proposée dans la littérature, avec des approches différentes en terme de localisation [63, 68]. Mais grâce à la rapidité de localisation de sources, il est envisageable soit de localiser

et de reconnaître conjointement la source, soit d'aborder ces problèmes séquentiellement, en estimant dans un premier temps la position de la source acoustique, pour ensuite filtrer le signal provenant de la direction estimée, et ainsi améliorer la qualité du signal utilisé pour la reconnaissance acoustique.

Certains de ces développements seront étudiés prochainement dans le cadre de l'ANR Astrid Deeplomatics [6] , traîtant de la fusion multimodale de données pour la protection de zones sensibles aux attaques d'engins volants à faible signature acoustique. Ce projet scientifique, regroupant les laboratoires LMSSC et Cedric du Cnam, le groupe Acoustique et Protection du Combattant et le groupe Visionique Avancée et Processing de l'Institut Saint-Louis ainsi que l'industriel ROBOOST Security Defense Health, a pour objectif de développer un système de défense couplant des informations acoustiques et optiques. Dans un premier temps, des antennes microphoniques sont disposées autour d'une zone à protéger. Dès qu'une des antennes détecte l'approche d'un drone, une caméra utilisant de l'imagerie enregistre une série d'images de l'objet détecté, et si un drone est reconnu grâce à de la vision assistée par ordinateur dans cette image, alors la caméra reçoit pour consigne de suivre visuellement les déplacements du drone afin d'aider à sa neutralisation. L'approche BeamLearning proposée dans le cadre de cette thèse sera donc chargée de la partie localisation de sources acoustiques de ce projet.

6. https ://deeplomatics.gitlab.io/

Bibliographie

[1] Jont B Allen et David A Berkley: *Image method for efficiently simulating small-room acoustics.* The Journal of the Acoustical Society of America, Vol. 65(No. 4) :pp.943–950, Avr. 1979.

[2] Marvin E Goldstein: *Aeroacoustics.* Mc Graw Hill, 1976.

[3] Michael Brandstein et Darren Ward: *Microphone arrays : signal processing techniques and applications.* Springer-Verlag, Berlin, Germany, 2001.

[4] Jacob Benesty, Jingdong Chen et Yiteng Huang: *Microphone array signal processing*, tome Vol. 1. Springer-Verlag, Berlin, Germany, 2008.

[5] Ralph Schmidt: *Multiple emitter location and signal parameter estimation.* IEEE transactions on antennas and propagation, Vol. 34(No. 3) :pp.276–280, Avr. 1986.

[6] Richard Roy et Thomas Kailath: *ESPRIT-estimation of signal parameters via rotational invariance techniques.* IEEE Transactions on acoustics, speech, and signal processing, Vol. 37(No. 7) :pp.984–995, Mars 1986.

[7] Charles Vanwynsberghe, P Challenge, Raphael Leiba, Jacques Marchal, Régis Marchiano, F Oliver, Gilles Puy et Pierre Vandergheynst: *Localisation et identification spectrale conjointe de sources large bande par parcimonie groupée.* Dans *Congrès Français d'Acoustique*, numéro No. 558, 2016.

[8] Antoine Peillot: *Imagerie acoustique par approximations parcimonieuses des sources.* Thèse de doctorat, Université Pierre et Marie Curie-Paris VI, 2012.

[9] Angeliki Xenaki, Peter Gerstoft et Klaus Mosegaard: *Compressive beamforming.* The Journal of the Acoustical Society of America, Vol. 136(No. 1) :pp.260–271, 2014.

[10] Georges Bienvenu et Laurent Kopp: *Optimality of high resolution array processing using the eigensystem approach.* IEEE Transactions on acoustics, speech, and signal processing, Vol. 31(No. 5) :pp.1235–1248, 1983.

[11] Joseph H DiBiase, Harvey F Silverman et Michael S Brandstein: *Robust localization in reverberant rooms.* Dans *Microphone Arrays*, chapitre Ch. 8, pages 157–180.

[12] Charles Knapp et Glifford Carter: *The generalized correlation method for estimation of time delay.* IEEE transactions on acoustics, speech, and signal processing, Vol. 24(No. 4) :pp.320–327, Août 1976.

[13] Thomas F Brooks et William M Humphreys: *A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays.* Journal of Sound and Vibration, Vol. 294(No. 4-5) :pp.856–879, 2006.

[14] Aro Ramamonjy: *Développement de nouvelles méthodes de classification/localisation de signaux acoustiques appliquées aux véhicules aériens.* Thèse de doctorat, Conservatoire national des arts et metiers, Paris, 2019.

[15] Aro Ramamonjy, Eric Bavu, Alexandre Garcia et Sébastien Hengy: *Source localization and identification with a compact array of digital mems microphones.* Dans *25th International Congress on Sound and Vibration (ICSV25)*, 2018.

[16] Darrell R Jackson et David R Dowling: *Phase conjugation in underwater acoustics.* The Journal of the Acoustical Society of America, Vol. 89(No. 1) :pp.171–181, 1991.

[17] Nadia Aloui: *Localisation sonore par retournement temporel.* Thèse de doctorat, École doctorale électronique, électrotechnique, automatique, traitement du signal (Grenoble), 2014.

[18] Eric Bavu: *Le puits à retournement temporel dans le domaine audible : un outil de focalisation et d'imagerie à haute résolution de sources sonores et vibratoires.* Thèse de doctorat, 2008.

[19] Stéphanie Lobréau: *Imagerie et caractérisation instationnaire de sources acoustiques en milieu réverbérant et bruité par renversement temporel et séparation de champs sur antenne hémisphérique double couche.* Thèse de doctorat, Conservatoire national des arts et metiers, Paris, 2015.

[20] Allen William Mills: *On the minimum audible angle.* The Journal of the Acoustical Society of America, Vol. 30(No. 4) :pp.237–246, 1958.

[21] Jens Blauert: *Spatial hearing : the psychophysics of human sound localization.* The MIT Press, 1997.

[22] Hélène Bahu: *Localisation auditive en contexte de synthèse binaurale non-individuelle.* Thèse de doctorat, Université Paris VI, 2016.

[23] Stanley Smith Stevens et Edwin B Newman: *The localization of actual sources of sound.* The American journal of psychology, Vol. 48(No. 2) :pp.297–306, 1936.

[24] Jack Hebrank et Donald Wright: *Spectral cues used in the localization of sound sources on the median plane.* The Journal of the Acoustical Society of America, Vol. 56(No. 6) :pp.1829–1834, 1974.

[25] Adelbert W Bronkhorst: *Localization of real and virtual sound sources.* The Journal of the Acoustical Society of America, Vol. 98(No. 5) :pp.2542–2553, 1995.

[26] Hans Wallach: *The role of head movements and vestibular and visual cues in sound localization.* Journal of Experimental Psychology, Vol. 27(No. 4) :pp.339, 1940.

[27] Robert Geirhos, David HJ Janssen, Heiko H Schütt, Jonas Rauber, Matthias Bethge et Felix A Wichmann: *Comparing deep neural networks against humans : object recognition when the signal gets weaker.* arXiv preprint arXiv :1706.06969, 2017.

[28] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár et C Lawrence Zitnick: *Microsoft COCO : Common objects in context.* Dans *European conference on computer vision*, pages pp.740–755. Springer, 2014.

[29] Alex Krizhevsky, Geoffrey Hinton *et al.*: *Learning multiple layers of features from tiny images.* 2009.

[30] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li et Li Fei-Fei: *Imagenet : A large-scale hierarchical image database.* Dans *2009 IEEE conference on computer vision and pattern recognition*, pages pp.248–255, Miami, Floride, 2009. Ieee.

[31] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal et Marvin Ritter: *Audio set : An ontology and human-labeled dataset for audio events.* Dans *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages pp.776–780, Los Angeles, New Orleans, 2017. IEEE.

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey et Sanjeev Khudanpur: *Librispeech : an ASR corpus based on public domain audio books*. Dans *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, Brisbane, QLD, 2015.

[33] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj et Tuomas Virtanen: *DCASE 2017 challenge setup : Tasks, datasets and baseline system*. Dans *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[34] J. Salamon, C. Jacoby et J. P. Bello: *A Dataset and Taxonomy for Urban Sound Research*. Dans *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov.

[35] Kevin P Murphy: *Machine learning : a probabilistic perspective*. The MIT Press, 2012.

[36] Lior Rokach et Oded Z Maimon: *Data mining with decision trees : theory and applications*, tome 69 2e édition. World scientific, 2008.

[37] Nicolas Tome Michel Crucianu, Marin Ferecatu: *Apprentissage, réseaux de neurones et modèles graphiques*. 2017.

[38] Ingo Steinwart et Andreas Christmann: *Support vector machines*. Springer Verlag, 2008.

[39] Ian Goodfellow, Yoshua Bengio et Aaron Courville: *Deep Learning*. The MIT Press, 2016. http://www.deeplearningbook.org.

[40] Pierre Lecomte et Philippe Aubert Gauthier: *Real-time 3D ambisonics using Faust, processing, pure data, and OSC*. Dans *15th International Conference on Digital Audio Effects (DAFx-15), Trondheim*, 2015.

[41] P Lecomte, PA Gauthier, C Langrenne, A Berry et A Garcia: *Filtrage directionnel dans un scène sonore 3D par une utilisation conjointe de Beamforming et d'Ambisonie d'ordre elevés*. Proc. of Congrès Français d'Acoustique, pages pp.169–175, 2016.

[42] Pierre Lecomte: *Ambisonie d'ordre élevé en trois dimensions : captation, transformations et décodage adaptatifs de champs sonores*. Thèse de doctorat, Conservatoire national des arts et metiers-CNAM, 2016.

[43] Pierre Lecomte: *AMBITOOLS : TOOLS FOR SOUND FIELD SYNTHESIS WITH HIGHER ORDER AMBISONICS-V1. 0*.

[44] Pierre LECOMTE: *Ambitools*. https://github.com/sekisushai/ambitools, mai 2020.

[45] Hadrien PUJOL, Éric BAVU et Alexandre GARCIA: *Antennes microphoniques intelligentes : Localisation de sources par Deep Learning*. Dans *Congrès français d'acoustique'18, Le Havre*, 2018.

[46] Hadrien PUJOL, Éric BAVU et Alexandre GARCIA: *Constitution d'une base de données physiquement valide pour la localisation de sources par Deep-Learning*. Dans *Congrès français d'acoustique'18, Le Havre*, 2018.

[47] Hendrik PURWINS, Bo LI, Tuomas VIRTANEN, Jan SCHLÜTER, Shuo Yiin CHANG et Tara SAINATH: *Deep learning for audio signal processing*. IEEE Journal of Selected Topics in Signal Processing, Vol. 13(No. 2) :pp.206–219, 2019.

[48] Michael J BIANCO, Peter GERSTOFT, James TRAER, Emma OZANICH, Marie A ROCH, Sharon GANNOT et Charles Alban DELEDALLE: *Machine learning in acoustics : Theory and applications*. The Journal of the Acoustical Society of America, Vol. 146(No. 5) :pp.3590–3628, 2019.

[49] Aditya Arie NUGRAHA, Antoine LIUTKUS et Emmanuel VINCENT: *Multichannel audio source separation with deep neural networks*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 24(No. 9) :pp.1652–1664, 2016.

[50] Aaron van den OORD, Sander DIELEMAN, Heiga ZEN, Karen SIMONYAN, Oriol VINYALS, Alex GRAVES, Nal KALCHBRENNER, Andrew SENIOR et Koray KAVUKCUOGLU: *Wavenet : A generative model for raw audio*. arXiv preprint arXiv :1609.03499, 2016.

[51] Éric BAVU, Aro RAMAMONJY, Hadrien PUJOL et Alexandre GARCIA: *TimeScaleNet : a Multiresolution Approach for Raw Audio Recognition using Learnable Biquadratic IIR Filters and Residual Networks of Depthwise-Separable One-Dimensional Atrous Convolutions*. IEEE Journ. of Selected Topics in Signal Processing, Vol. 13(No. 2) :pp.220–235, 2019.

[52] Ben Zion STEINBERG, Mark J BERAN, Steven H CHIN et James H HOWARD JR: *A neural network approach to source localization*. The Journal of the Acoustical Society of America, 90(4) :2081–2090, 1991.

[53] Kamen GUENTCHEV et John WENG: *Learning-based three dimensional sound localization using a compact non-coplanar array of microphones*. Dans *Proc. AAAI Spring Symposium on Intelligent Environments*, 1998.

[54] Juyang Weng et Kamen Y Guentchev: *Three-dimensional sound localization from a compact non-coplanar array of microphones using tree-based learning.* The Journal of the Acoustical Society of America, 110(1) :310–323, 2001.

[55] Ronen Talmon, Israel Cohen et Sharon Gannot: *Supervised source localization using diffusion kernels.* Dans *2011 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, pages 245–248. IEEE, 2011.

[56] Antoine Deleforge, Florence Forbes et Radu Horaud: *Acoustic space learning for sound-source separation and localization on binaural manifolds.* International journal of neural systems, Vol. 25(No. 01) :pp.1440003, 2015.

[57] Wenxu Liu, Yixin Yang, Mengqian Xu, Liangang Lü, Zongwei Liu et Yang Shi: *Source localization in the deep ocean using a convolutional neural network.* The Journal of the Acoustical Society of America, 147(4) :EL314–EL319, 2020.

[58] Junhyeong Pak et Jong Won Shin: *Sound localization based on phase difference enhancement using deep neural networks.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(8) :1335–1345, 2019.

[59] Luca Comanducci, Federico Borra, Paolo Bestagini, Fabio Antonacci, Stefano Tubaro et Augusto Sarti: *Source Localization Using Distributed Microphones in Reverberant Environments Based on Deep Learning and Ray Space Transform.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28 :2238–2251, 2020.

[60] Masahiro Yasuda, Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu et Keisuke Imoto: *Sound Event Localization Based on Sound Intensity Vector Refined by Dnn-Based Denoising and Source Separation.* Dans *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 651–655. IEEE, 2020.

[61] Reza Varzandeh, Kamil Adiloğlu, Simon Doclo et Volker Hohmann: *Exploiting Periodicity Features for Joint Detection and DOA Estimation of Speech Sources Using Convolutional Neural Networks.* Dans *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 566–570. IEEE, 2020.

[62] Soumitro Chakrabarty et Emanuel A.P. Habets: *Multi-speaker DOA estimation using deep convolutional networks trained with noise signals.* IEEE Journal of Selected Topics in Signal Processing, Vol. 13(No. 1) :pp.8–21, 2019.

[63] Laureline Perotin, Romain Serizel, Emmanuel Vincent et Alexandre Guerin: *CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings*. IEEE Journal of Selected Topics in Signal Processing, Vol. 13(No. 1) :pp.22–33, 2019.

[64] Andreas Brendel et Walter Kellermann: *Distributed source localization in acoustic sensor networks using the coherent-to-diffuse power ratio*. IEEE Journal of Selected Topics in Signal Processing, 13(1) :61–75, 2019.

[65] Kay L Gemba, Santosh Nannuru et Peter Gerstoft: *Robust ocean acoustic localization with sparse Bayesian learning*. IEEE Journal of Selected Topics in Signal Processing, 13(1) :49–60, 2019.

[66] Yining Liu, Haiqiang Niu et Zhenglin Li: *A multi-task learning convolutional neural network for source localization in deep ocean*. The Journal of the Acoustical Society of America, 148(2) :873–883, 2020.

[67] Soumitro Chakrabarty et Emanuël AP Habets: *Broadband DOA estimation using convolutional neural networks trained with noise signals*. Dans *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on applications*, pages pp.136–140, 2017.

[68] Sharath Adavanne, Archontis Politis, Joonas Nikunen et Tuomas Virtanen: *Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks*. IEEE Journ. of Selected Topics in Signal Processing, Vol. 13(No. 1) :pp.34–48, 2019.

[69] Sharath Adavanne, Archontis Politis et Tuomas Virtanen: *A Multi-room Reverberant Dataset for Sound Event Localization and Detection*. Dans *Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, Munich, Germany, 2019. https://arxiv.org/abs/1905.08546.

[70] Toni Hirvonen: *Classification of spatial audio location and content using convolutional neural networks*. Dans *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[71] Weipeng He, Petr Motlicek et Jean Marc Odobez: *Deep neural networks for multiple speaker detection and localization*. Dans *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 74–79. IEEE, 2018.

[72] Harshavardhan Sundar, Weiran Wang, Ming Sun et Chao Wang: *Raw Waveform Based End-to-end Deep Convolutional Network for Spatial Localization of Multiple Acoustic Sources*. Dans

ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4642–4646. IEEE, 2020.

[73] Nelson YALTA, Kazuhiro NAKADAI et Tetsuya OGATA: *Sound source localization using deep learning models.* Journal of Robotics and Mechatronics, Vol. 29(No. 1) :pp.37–48, 2017.

[74] Ning MA, Tobias MAY et Guy J BROWN: *Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12) :2444–2453, 2017.

[75] Quan NGUYEN, Laurent GIRIN, Gérard BAILLY, Frédéric ELISEI et Duc Canh NGUYEN: *Autonomous sensorimotor learning for sound source localization by a humanoid robot.* 2018.

[76] Sunit SIVASANKARAN, Emmanuel VINCENT et Dominique FOHR: *Keyword-based speaker localization : Localizing a target speaker in a multi-speaker environment.* 2018.

[77] Fabio VESPERINI, Paolo VECCHIOTTI, Emanuele PRINCIPI, Stefano SQUARTINI et Francesco PIAZZA: *A neural network based algorithm for speaker localization in a multi-room environment.* Dans *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.

[78] Eric L FERGUSON, Stefan B WILLIAMS et Craig T JIN: *Sound source localization in a multipath environment using convolutional neural networks.* Dans *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages pp.2386–2390. IEEE, 2018.

[79] Zhenyu TANG, John D KANU, Kevin HOGAN et Dinesh MANOCHA: *Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks.* arXiv preprint arXiv :1904.08452, 2019.

[80] Daniele SALVATI, Carlo DRIOLI et Gian Luca FORESTI: *Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions.* IEEE Transactions on Emerging Topics in Computational Intelligence, 2(2) :103–116, 2018.

[81] Ryu TAKEDA et Kazunori KOMATANI: *Sound source localization based on deep neural networks with directional activate function exploiting phase information.* Dans *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on Acoustics, speech and Signal Processing (ICASSP)*, pages pp.405–409. IEEE, 2016.

[82] Mariam Yiwere et Eun Joo Rhee: *Distance estimation and localization of sound sources in reverberant conditions using deep neural networks.* International Journal of Applied Engineering Research, Vol. 12(No. 22) :pp.12384–12389, 2017.

[83] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L Jones, Eng Siong Chng et Haizhou Li: *A learning-based approach to direction of arrival estimation in noisy and reverberant environments.* Dans *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages pp.2814–2818, 2015.

[84] Dmitry Suvorov, Ge Dong et Roman Zhukov: *Deep residual network for sound source localization in the time domain.* arXiv preprint arXiv :1808.06429, 2018.

[85] Juan Vera-Diaz, Daniel Pizarro et Javier Macias-Guarasa: *Towards End-to-End Acoustic Localization Using Deep Learning : From Audio Signals to Source Position Coordinates.* Sensors, Vol. 18(No. 10) :pp.3418, 2018.

[86] Yankun Huang, Xihong Wu et Tianshu Qu: *A Time-domain Unsupervised Learning Based Sound Source Localization Method.* Dans *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pages 26–32. IEEE, 2020.

[87] Danilo Comminiello, Marco Lella, Simone Scardapane et Aurelio Uncini: *Quaternion convolutional neural networks for detection and localization of 3d sound events.* Dans *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8533–8537. IEEE, 2019.

[88] Y Hu, PN Samarasinghe, S Gannot et TD Abhayapala: *Semi-supervised multiple source localization using relative harmonic coefficients under noisy and reverberant environments.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020.

[89] Yonggang Hu, Prasanga N Samarasinghe, Thushara D Abhayapala et Sharon Gannot: *Unsupervised Multiple Source Localization Using Relative Harmonic Coefficients.* Dans *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 571–575. IEEE, 2020.

[90] Heinrich W Löllmann, Christine Evers, Alexander Schmidt, Heinrich Mellmann, Hendrik Barfuss, Patrick A Naylor et Walter Kellermann: *The LOCATA challenge data corpus for acoustic source localization and tracking.* Dans *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 410–414. IEEE, 2018.

[91] Junhyeong Pak et Jong Won Shin: *LOCATA Challenge : A Deep Neural Networks-Based Regression Approach for Direction-Of-Arrival Estimation.* Dans *Proc. of LOCATA Challenge Workshop-a satellite event of IWAENC*, 2018.

[92] Sharon Gannot, Martin Haardt, Walter Kellermann et Peter Willett: *Introduction to the Issue on Acoustic Source Localization and Tracking in Dynamic Real-Life Scenes.* IEEE Journal of Selected Topics in Signal Processing, 13(1) :3–7, 2019.

[93] Julius O. Smith: *Physical Audio Signal Processing.* W3K Publishing, ISBN 978-0-9745607-2-4.

[94] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho et Yoshua Bengio: *Gated feedback recurrent neural networks.* Dans *ICML'15 : Procedings of the 32nd International Conference on International conference on machine learning*, tome Vol. 37, pages pp.2067–2075, July 2015.

[95] Sepp Hochreiter et Jürgen Schmidhuber: *Long short-term memory.* Neural computation, Vol. 9(No. 8) :pp.1735–1780, 1997.

[96] Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta et Deva Ramanan: *Pixelnet : Towards a general pixel-level architecture.* arXiv preprint arXiv :1609.06694, 2016.

[97] Jean Michel Combes, Alexander Grossmann et Philippe Tchamitchian: *Wavelets : Time-Frequency Methods and Phase Space Proceedings of the International Conference, Marseille, France, December 14–18, 1987.* Springer Verlag, 1991.

[98] Lukasz Kaiser, Aidan N. Gomez et Francois Chollet: *Depthwise Separable Convolutions for Neural Machine Translation.* Dans *International Conference on Learning Representations*, pages pp.1–10, 2018.

[99] Dario Rethage, Jordi Pons et Xavier Serra: *A wavenet for speech denoising.* Dans *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages pp.5069–5073. IEEE, 2018.

[100] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun: *Identity mappings in deep residual networks.* Dans *European conference on computer vision ECCV'16*, tome Vol. 9908, pages pp.630–645. Springer, 2016.

[101] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun: *Deep residual learning for image recognition.* Dans *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages pp.770–778, 2016.

[102] Rupesh K Srivastava, Klaus Greff et Jürgen Schmidhuber: *Training very deep networks.* Dans *NIPS'15 : Proceedings of the 28th Internatinal Conference on Neuronal information processing systems*, pages pp.2377–2385, Dec. 2015.

[103] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke et Andrew Rabinovich: *Going deeper with convolutions.* Dans *IEEE conference on computer vision and pattern recognition (CVPR)*, pages pp.1–9, 2015.

[104] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke et Alexander A Alemi: *Inception-v4, inception-resnet and the impact of residual connections on learning.* Dans *Thirty-first AAAI conference on artificial intelligence*, Fev. 2017.

[105] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu et Kaiming He: *Aggregated residual transformations for deep neural networks.* Dans *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages pp.1492–1500, 2017.

[106] Frank Rosenblatt: *The perceptron : a probabilistic model for information storage and organization in the brain.* Psychological review, Vol. 65(No. 6) :pp.386–408, 1958.

[107] Bekir Karlik et A Vehbi Olgac: *Performance analysis of various activation functions in generalized MLP architectures of neural networks.* International Journal of Artificial Intelligence and Expert Systems (IJAE), Vol. 1(No. 4) :pp.111–122, 2011.

[108] Sergey Ioffe et Christian Szegedy: *Batch normalization : Accelerating deep network training by reducing internal covariate shift.* arXiv preprint arXiv :1502.03167, 2015.

[109] Jimmy Lei Ba, Jamie Ryan Kiros et Geoffrey E Hinton: *Layer normalization.* arXiv preprint arXiv :1607.06450, 2016.

[110] Tim Salimans et Durk P Kingma: *Weight normalization : A simple reparameterization to accelerate training of deep neural networks.* Dans *NIPS'16 : Proceedings of the 30th Internatinal Conference on Neuronal information processing systems*, pages pp.901–909, Dec. 2016.

[111] Günter Klambauer, Thomas Unterthiner, Andreas Mayr et Sepp Hochreiter: *Self-normalizing neural networks.* Dans *Advances in Neural Information Processing Systems 30*, pages pp.972–981, 2017.

[112] Yann LeCun, Yoshua Bengio et Geoffrey Hinton: *Deep learning.* Nature, Vol. 521 :pp.436–444, 2015.

[113] Andrew L Maas, Awni Y Hannun et Andrew Y Ng: *Rectifier nonlinearities improve neural network acoustic models.* Dans *Proc. ICML*, tome Vol. 30, page pp.3, 2013.

[114] Djork Arné Clevert, Thomas Unterthiner et Sepp Hochreiter: *Fast and accurate deep network learning by exponential linear units (elus).* arXiv preprint arXiv :1511.07289, 2015.

[115] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille et Quoc V Le: *Smooth Adversarial Training.* arXiv preprint arXiv :2006.14536, 2020.

[116] Martín Abadi, Ashish Agarwal *et al.*: *TensorFlow : Large-Scale Machine Learning on Heterogeneous Systems*, 2015. http://tensorflow.org/, Software available from tensorflow.org.

[117] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun: *Delving deep into rectifiers : Surpassing human-level performance on imagenet classification.* Dans *Proceedings of the 2015 IEEE international conference on computer vision (ICSV)*, pages pp.1026–1034, 2015.

[118] Diederik P Kingma et Jimmy Ba: *Adam : A method for stochastic optimization.* arXiv preprint arXiv :1412.6980, 2014.

[119] Sebastian Ruder: *An overview of gradient descent optimization algorithms.* arXiv preprint arXiv :1609.04747, 2016.

[120] Michael Vorländer: *Computer simulations in room acoustics : Concepts and uncertainties.* The Journal of the Acoustical Society of America, Vol. 133(No. 3) :pp.1203–1213, 2013.

[121] Carlos Spa Carvajal: *Time-domain numerical methods in room acoustics simulations.* Thèse de doctorat, Universitat Pompeu Fabra, 2009.

[122] Lauri Savioja et U Peter Svensson: *Overview of geometrical room acoustic modeling techniques.* The Journal of the Acoustical Society of America, Vol. 138(No. 2) :pp.708–730, 2015.

[123] J D Polack: *Modifying chambers to play billiards : the foundations of reverberation theory.* Acta Acustica united with Acustica, Vol. 76(No. 6) :pp.256–272, 1992.

[124] Philip M Morse et Richard H Bolt: *Sound waves in rooms.* Reviews of modern physics, Vol. 16(No. 2) :pp.69, 1944.

[125] Hugo Dujourdy: *Acoustical diffusion in workspaces.* Theses de doctorat, Université Pierre et Marie Curie - Paris VI, avril 2016. https://tel.archives-ouvertes.fr/tel-01392577.

[126] Chaima Soussi, Walid Larbi et Jean François Deü: *Experimental and numerical analysis of sound transmission loss through double glazing windows.* Dans *Proceeding of the 2nd International Conference on Acoustics and Vibration (ICAV)*, pages pp.195–203, 2018.

[127] Lauri Savioja, Timo J. Rinne et Tapio Takala: *Simulation of Room Acoustics with a 3-D Finite Difference Mesh.* Dans *Proceedings of the 1994 International Computer Music Conference, ICMC 1994, Aarhus, Denmark, September 12-17, 1994*, 1994.

[128] Konrad Kowalczyk: *Boundary and medium modelling using compact finite difference schemes in simulations of room acoustics for audio and architectural design applications.* Thèse de doctorat, Queen's University Belfast, 2010.

[129] Philip McCord Morse et K Uno Ingard: *Theoretical acoustics.* Princeton university press, 1986.

[130] Heinrich Kuttruff: *Room acoustics.* CRC Press, 2016.

[131] Fridolin Mechel: *Room acoustical fields.* Springer Verlag, 2012.

[132] Michael Barron: *Auditorium acoustics and architectural design.* Spon Press, 2009.

[133] Massimo Garai: *Measurement of the sound-absorption coefficient in situ : the reflection method using periodic pseudo-random sequences of maximum length.* Applied acoustics, Vol. 39(No. 1-2) :pp.119–139, 1993.

[134] Mutsushige Yuzawa: *A method of obtaining the oblique incident sound absorption coefficient through an on-the-spot measurement.* Applied Acoustics, Vol. 8(No. 1) :pp.27–41, 1975.

[135] U Ingård et RH Bolt: *A free field method of measuring the absorption coefficient of acoustic materials.* The Journal of the Acoustical Society of America, Vol. 23(No. 5) :pp.509–516, 1951.

[136] BRITISH STANDARD et BSEN ISO: *Acoustics—Measurement of sound absorption in a reverberation room*, 2003.

[137] Giuliana Benedetto et Renato Spagnolo: *Reverberation time in enclosures : The surface reflection law and the dependence of the absorption coefficient on the angle of incidence.* The Journal of the Acoustical Society of America, Vol. 77(No. 4) :pp.1447–1451, 1985.

[138] Federico Cruz Barney: *Evaluation des performances d'un environnement informatique d'acoustique prévisionnelle.* Thèse de doctorat, Université du Maine, Vol. 6, 1999.

[139] Eric A Lehmann et Anders M Johansson: *Prediction of energy decay in room impulse responses simulated with an image-source model.* The Journal of the Acoustical Society of America, Vol. 124(No. 1) :pp.269–277, 2008.

[140] Marc Aretz, Pascal Dietrich et Michael Vorländer: *Application of the mirror source method for low frequency sound prediction in rectangular rooms.* Acta Acustica united with Acustica, Vol. 100(No. 2) :pp.306–319, 2014.

[141] Stefan Drechsler et Uwe M Stephenson: *The effect of edge caused diffusion on the reverberation time-A semi analytical approach.* Dans *Proceedings of Meetings on Acoustics ICA2013*, tome Vol. 19, page pp.015118. Acoustical Society of America, 2013.

[142] Mike Barron: *Non-linear decays in simple spaces and their possible exploitation.* Dans *Proc. Institute of Acoustics, 8th International Conference on Auditorium Acoustics, Dublin, 20-22 May*, 2011.

[143] Standard Ecma: *ECMA-404 The JSON Data Interchange Format*, 2015.

[144] Éric Bavu, Hadrien Pujol et Alexandre Garcia: *Antennes non calibrées, suivi métrologique et problèmes inverses : une approche par Deep Learning.* Dans *Congrès fracçais d'acoustique '18, Le Havre*, 2018.

[145] Gordon S Kino: *Acoustic waves : devices, imaging, and analog signal processing*, tome Vol. 107. Prentice-hall Englewood Cliffs, NJ, 1987.

[146] Bart GM Husslage, Gijs Rennen, Edwin R Van Dam et Dick Den Hertog: *Space-filling Latin hypercube designs for computer experiments.* Optimization and Engineering, Vol. 12(No. 4) :pp.611–630, 2011.

[147] Justin Salamon et Juan Pablo Bello: *Deep convolutional neural networks and data augmentation for environmental sound classification.* IEEE Signal Processing Letters, Vol. 24(No. 3) :pp.279–283, 2017.

[148] Odeon: *Women cocktail party backgroundspeech.*

[149] Lokki Tapio, Pätynen Jukka et Pulkki Ville: *Anechoic recordings of symphonic music*, 2008.

[150] Robin Scheibler, Eric Bezzam et Ivan Dokmanić: *Pyroomacoustics : A Python Package for Audio Room Simulation and Array Processing Algorithms.* Dans *2018 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages pp.351–355. IEEE, 2018.

[151] Saul R Dooley et Asoke K Nandi: *Adaptive subsample time delay estimation using Lagrange interpolators.* IEEE Signal Processing Letters, Vol. 6(No. 3) :pp.65–67, 1999.

[152] Vesa Valimaki et Timo I Laakso: *Principles of fractional delay filters.* Dans *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, tome Vol. 6, pages pp.3870–3873. IEEE, 2000.

[153] Vesa Valimaki et Azadeh Haghparast: *Fractional delay filter design based on truncated Lagrange interpolation.* IEEE Signal Processing Letters, Vol. 14(No. 11) :pp.816–819, 2007.

[154] Andreas Franck: *Efficient algorithms and structures for fractional delay filtering based on Lagrange interpolation.* Journal of the Audio Engineering Society, Vol. 56(No. 12) :pp.1036–1056, 2009.

[155] Tian Bo Deng: *Robust structure transformation for causal Lagrange-type variable fractional-delay filters.* IEEE Transactions on Circuits and Systems I : Regular Papers, Vol. 56(No. 8) :pp.1681–1688, 2008.

[156] Parinya Soontornwong, Sorawat Chivapreecha et Chusit Pradabpet: *A transient-free structure for Lagrange-type variable fractional-delay digital filter.* Dans *2014 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages pp.1–5. IEEE, 2014.

[157] Timo I Laakso, Vesa Valimaki, Matti Karjalainen et Unto K Laine: *Splitting the unit delay [FIR/all pass filters design].* IEEE Signal Processing Magazine, Vol. 13(No. 1) :pp.30–60, 1996.

[158] VP Valimaki: *Discrete-time modeling of acoustic tubes using fractional delay filters.* Thèse de doctorat, 1998.

[159] V Välimäki et TI Laakso: *Fractional delay filters—design and applications.* Dans *Nonuniform Sampling*, pages pp.835–895. Springer, 2001.

[160] Dutta Roy: *Maximally Flat FD FIR Filter : Lagrange Interpolation.* Thèse de doctorat, PhD thesis, University of Calcutta, 1995.

[161] Alan V Oppenheim et Ronald W Schafer: *Digital Signal Processing.* Prentice-Hall, 1975.

[162] Patrick M Peterson: *Simulating the response of multiple microphones to a single acoustic source in a reverberant room.* The Journal of the Acoustical Society of America, Vol. 80(No. 5) :pp.1527–1529, 1986.

[163] Lothar Cremer et Helmut A Müller: *Principles and applications of room acoustics*, tome Vol. 1. Chapman & Hall, 1982.

[164] Eric A Lehmann, Anders M Johansson et Sven Nordholm: *Reverberation-time prediction method for room impulse responses simulated with the image-source model.* Dans *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages pp.159–162. IEEE, 2007.

[165] Manfred R Schroeder et KH Kuttruff: *On frequency response curves in rooms. Comparison of experimental, theoretical, and Monte Carlo results for the average frequency spacing between maxima.* The Journal of the Acoustical Society of America, Vol. 34(No. 1) :pp.76–80, 1962.

[166] Jin Sung Suh et PA Nelson: *Measurement of transient response of rooms and comparison with geometrical acoustic models.* The Journal of the Acoustical Society of America, Vol. 105(No. 4) :pp.2304–2317, 1999.

[167] Augustinus J Berkhout: *A holographic approach to acoustic control.* Journal of the audio engineering society, Vol. 36(No. 12) :pp.977–995, 1988.

[168] Michael A Gerzon: *Periphony : With-height sound reproduction.* Journal of the audio engineering society, Vol. 21(No. 1) :pp.2–10, 1973.

[169] Jeffrey Stephen Bamford: *An analysis of ambisonic sound systems of first and second order.* Thèse de doctorat, University of Waterloo, 1995.

[170] Dave G Malham: *Experience with a large area 3d ambisonic sound systems.* Proceedings-Institute of Acoustics, Vol. 14 :pp.209–209, 1992.

[171] Jérôme Daniel: *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia.* Thèse de doctorat, Thèse de doctorat de l'Université Paris VI, France, 2000.

[172] George B Arfken et Hans J Weber: *Mathematical methods for physicists.* 1999.

[173] Christian Nachbar, Franz Zotter, Etienne Deleflie et Alois Sontacchi: *Ambix-a suggested ambisonics format.* Dans *Ambisonics Symposium, Lexington*, page pp.11, 2011.

[174] Michel Bruneau: *Manuel d'acoustique fondamentale*. 1998.

[175] Earl G Williams: *Fourier acoustics : sound radiation and nearfield acoustical holography*. Elsevier, 1999.

[176] Jérôme Daniel: *Spatial sound encoding including near field effect : Introducing distance coding filters and a viable, new ambisonic format*. Dans *Audio Engineering Society Conference : 23rd International Conference : Signal Processing in Audio Recording and Reproduction*. Audio Engineering Society, 2003.

[177] Boaz Rafaely: *Fundamentals of spherical array processing*, tome Vol. 8. Springer, 2015.

[178] Boaz Rafaely, Barak Weiss et Eitan Bachmat: *Spatial aliasing in spherical microphone arrays*. IEEE Transactions on Signal Processing, Vol. 55(No. 3) :pp.1003–1010, 2007.

[179] Mark A Poletti: *Three-dimensional surround sound systems based on spherical harmonics*. Journal of the Audio Engineering Society, Vol. 53(No. 11) :pp.1004–1025, 2005.

[180] Franz Zotter, Hannes Pomberger et Matthias Frank: *An alternative ambisonics formulation : Modal source strength matching and the effect of spatial aliasing*. Dans *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[181] Christophe Langrenne, Eric Bavu et Alexandre Garcia: *A linear phase IIR filterbank for the radial filters of ambisonic recordings*. Dans *EAA Spatial Audio Signal Processing Symposium*, 2019.

[182] Pierre Amaury Grumiaux, Srdjan Kitic, Laurent Girin et Alexandre Guérin: *High-Resolution Speaker Counting In Reverberant Rooms Using CRNN With Ambisonics Features*. arXiv preprint arXiv :2003.07839, 2020.

[183] Daniel P Jarrett, Emanuël AP Habets, Mark RP Thomas et Patrick A Naylor: *Simulating room impulse responses for spherical microphone arrays*. Dans *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages pp.129–132. IEEE, 2011.

[184] Pierre Lecomte, Philippe Aubert Gauthier, Christophe Langrenne, Alain Berry et Alexandre Garcia: *Cancellation of room reflections over an extended area using Ambisonics*. The Journal of the Acoustical Society of America, Vol. 143(No. 2) :pp.811–828, 2018.

[185] Adrian Freed: *Open sound control : A new protocol for communicating with sound synthesizers*. Dans *International Computer Music Conference (ICMC)*, 1997.

[186] Stéphane Letz, Yann Orlarey et Dominique Fober: *Jack audio server for multi-processor machines*. 2005.

[187] Juan Pablo Cáceres et Chris Chafe: *JackTrip : Under the hood of an engine for network audio*. Journal of New Music Research, Vol. 39(No. 3) :pp.183–187, 2010.

[188] E.F.F Codd: *Is Your DBMS Really Relational ?* Computerworld, Vol. 19(No. 41) :pp. 1–2, 1985-10-14, ISSN 0010-4841.

[189] Yakov Shafranovich: *Common format and MIME type for comma-separated values (CSV) files*. 2005.

[190] Eric Van der Vlist: *XML schema*. O'Reilly Media, Inc., 2002.

[191] Wes McKinney: *Data Structures for Statistical Computing in Python*. Dans Stéfan van der Walt et Jarrod Millman (rédacteurs) : *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[192] Mark C Sullivan: *Practical array processing*. McGraw-Hill, 2009.

[193] Harry L Van Trees: *Optimum array processing : Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[194] Max Virepinte: *Étalonnage champ libre en amplitude et en phase des microphones*. Mémoire d'ingénieur, Conservatoire national des arts et metiers-CNAM, 2020.

[195] L Lamotte, T Le Magueresse et C Picard: *Mesurer et corriger la réponse en fréquence d'un MEMS*.

[196] Charles Vanwynsberghe: *Réseaux à grand nombre de microphones : applicabilité et mise en œuvre*. Thèse de doctorat, Paris 6, 2016.

[197] Patrice Y Simard, David Steinkraus, John C Platt *et al.*: *Best practices for convolutional neural networks applied to visual document analysis.* Dans *Icdar*, tome Vol. 3, 2003.

[198] Boaz Rafaely, Yotam Peled, Morag Agmon, Dima Khaykin et Etan Fisher: *Spherical microphone array beamforming*. Dans *Speech Processing in Modern Communication*, pages pp.281–305. Springer, 2010.

[199] Lutz Prechelt: *Early stopping-but when ?* Dans *Neural Networks : Tricks of the trade*, pages pp.55–69. Springer, 1998.

[200] Dalia El Badawy et Ivan Dokmanić: *Direction of Arrival With One Microphone, a Few LEGOs, and Non-Negative Matrix Factorization*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 26(No. 12) :pp.2436–2446, 2018.

[201] Daniel P Jarrett, Emanuël AP Habets et Patrick A Naylor: *Theory and applications of spherical microphone array processing*, tome Vol. 9. Springer, 2017.

[202] Xuan Li, Shefeng Yan, Xiaochuan Ma et Chaohuan Hou: *Spherical harmonics MUSIC versus conventional MUSIC*. Applied Acoustics, Vol. 72(No. 9) :pp.646–652, 2011.

[203] Wenxia Wang, Shefeng Yan et Linlin Mao: *On the Statistical Performance of Spherical Harmonics MUSIC*. Signal Processing, page pp.107622, 2020.

[204] Miodrag Lovric (rédacteur): *International Encyclopedia of Statistical Science*. Springer, 2011, ISBN 978-3-642-04897-5. https://doi.org/10.1007/978-3-642-04898-2.

[205] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä et Ludolf Erwin Meester: *A Modern Introduction to Probability and Statistics : Understanding why and how*.

Annexe A

Présentation de l'algorithme SRP-PHAT

On définit la matrice de corrélation généralisée R^{GCC}, qui correspond à la transformée de Fourier inverse de la corrélation spectrale croisée, entre les signaux $x_1(t)$ et $x_2(t)$ captés par deux microphones comme :

$$R^{GCC}_{x_1,x_2}(k) = (1/L)\sum_{f=0}^{L-1} \Phi(f) E[X_1(f)X_2^*(f)] e^{j\omega k/L}$$

avec $\Phi(f)$ une fonction de pondération quelconque dans le domaine fréquentiel. L'algorithme SRP-PHAT utilise donc les information de cette matrice de corrélation généralisée avec la pondération PHAT (*PHAse array Transform*) qui est définie comme :

$$\Phi_{PHAT}(f) = |E[X_1(f)X_2^*(f)]|^{-1}$$

On obtient alors une fonctino qui ne dépend plus de la norme des vecteurs mais uniquement de leur phase :

$$R^{PHAT}_{x_1,x_2}(k) = (1/L)\sum_{f=0}^{L-1} arg(E[X_1(f)X_2^*(f)]) e^{j\omega k/L}$$

En pratique pour déterminer la position d'une source par cette méthode il faut tout d'abord appairer les microphones, puis calculer sur l'ensemble des points d'intérêts la matrice de corrélation :

$$\mathcal{S}(s, R^{GCC}) = \sum_{m=1}^{M} R^{GCC}_{m,s}$$

La position estimée sera donc la position qui maximise sur $\mathcal{S}$ la somme des matrices R^{GCC} pour tous les points d'intérêts.

Annexe B

Présentation de l'algorithme MUSIC

Soient M microphones ($\mathbf{X}$) captant N source distinctes ($\mathbf{S}$), alors chaque microphone capte les N sources, mais aussi un bruit de fond noté $\mathbf{V}$:

$$\mathbf{X} = \mathbf{A}_{m,n}\mathbf{S} + \mathbf{V}$$

Avec $\mathbf{A}_{m,n}$ la matrice des retards du aux distances entre les sources et les microphones :

$$\mathbf{A}_n = (e^{-j\omega(k_n+\tau_{1,n})}, e^{-j\omega(k_n+\tau_{2,n})}, ...e^{-j\omega(k_n+\tau_{M,n})})^T$$

On définit ensuite la matrice de densité spectrale croisée :

$$\mathbf{\Phi} = E[\mathbf{X}\mathbf{X}^H] = \mathbf{A}_m\mathbf{\Phi}\mathbf{A}_m^H + \sigma_v^2$$

L'algorithme MUSIC cherche alors à sélectionner les vecteurs propres associés aux valeurs propres jugés suffisamment grandes pour correspondre à une source ponctuelle et non à du bruit. Pour ce faire, la matrice $\mathbf{\Phi}$ est décomposé en un sous espace correspondant uniquement à du bruit $\mathbf{G}$, et à un sous espace constitué de sources et de bruit $\mathbf{Q}$.

En pratique, les signaux de pressions $\mathbf{X}$ sont mesurés sur les capteurs, à partir desquels la matrice $\mathbf{\Phi}$, ainsi que les sous espaces $\mathbf{Q}$ et $\mathbf{G}$ sont calculés. Enfin, le calcul d'un pseudo-spectre sur tous les points d'intérêts. Plutôt que de vérifier si le point d'intérêt appartient au sous-ensemble constitué de sources et de bruit $\mathbf{Q}$, on vérifie si ce point d'intérêt est orthogonal au sous espace correspondant uniquement à du bruit ($\mathbf{G}$), en cherchant les points maximisant :

$$J = 1/(\mathbf{A}^H\mathbf{G}\mathbf{G}^H\mathbf{A})$$

Une variante de l'algorithme MUSIC est aussi utilisé dans ce manuscrit de thèse de doctorat : SH-MUSIC. Cette variante propose de travailler dans le domaine ambisonique plutôt que dans le domaine fréquentiel pour estimer la position des sources. Mais les démarches sont similaires, et ne seront donc pas exposées ici.

Annexe C

Schéma JSON

```
{
"$schema": "http://json-schema.org/draft-04/schema#",
"Version": 0,
"type": "object",
"properties":{
"Nom": {"type": "string",
        "Details": "[requis] uuid.uuid4().hex"},

"Chemin": {"type": "string",
        "Details": "[requis] Chemin vers le fichier .wav"},

"Commentaire": {"type": "string",
        "Details": "[option] Commentaire "},

"Date": {"type": "string",
        "Details": "[option] JJ/MM/AAAA" },

"Duree": {"type": "number",
        "Details": "[option] Duree totale de l enregistrement (s)"},

"Emission": {"type": "number",
        "Details": "[option] Duree d emission du signal (s)"},

"Debut": {"type": "number",
        "Details": "[option] Debut de l information dans le fichier .wav (s)"},

"Rapport_signal_bruit": {"anyOf": [{"type": "string"}, {"type": "number"}],
        "Details": "[option] Inf ou valeur du rapport"},

"Forme": {"type": "string",
```

```
        "enum":["Sinus", "Cosinus", "Dirac", "Bruit_Blanc", "Bruit_Rose", "RIR",
        "Parole", "Son_Environnementaux"],
        "Details": "[requis] Forme mathematique du signal"},

"Frequence": {"type": "number",
        "Details": "[option] Frequence de la source (Hz)"},

"Frequence_echantillonnage": {"type": "number",
        "Details": "[requis] Frequence d echantillonage du fichier .wav (Hz)"},

"Enregistrement": {"type": "string",
        "enum":["Simulation", "Experience"],
        "Details": "[requis] Indique si le .wav est un enregistrement reel ou d une simulation"},

"Antenne": {"type": "string",
        "Details": "[requis] Presentation succinte de l'antenne ayant servi pour enregistrer les donnees"},

"Canaux": {"type": "number",
        "Details": "[requis] Nombre de canaux dans le fichier .wav"},

"Environement":{"type": "string",
        "enum": ["Champ libre", "Reflexion au sol", "Salle"],
        "Details": "[option] Environement de propagation de la source"},

"Type_salle":{"type": "array",
        "items": [{"type": "number",
                "enum": [10.0, 9.0, 8.0]},
        {"type": "number",
                "enum": [9.0, 8.0]},
        {"type": "number",
                "enum": [4.0]},
        {"type": "number",
                "enum": [0.1, 0.2, 0.4, 0.5, 0.6, 0.7]}],
        "Details": "[option] Caracteristique de la salle [L, P, H, alpha] "},

"Ordre_reflexion":{"type": "number",
        "Details": "[option] Ordre maximal de reflexion de la source image pour pyroomacoustics"},

"X_reel": {"type": "number",
        "Details": "[requis] Coordonnee reelle de la source en X (m)"},

"Y_reel": {"type": "number",
        "Details": "[requis] Coordonnee reelle de la source en Y (m)"},

"Z_reel": {"type": "number",
```

```
        "Details": "[requis] Coordonnee reelle de la source en Z (m)"},

"R_reel": {"type": "number",
        "Details": "[option] Distance reelle de la source a l antenne (m)"},

"Theta_reel": {"type": "number",
        "Details": "[option] Azimut reel de la source (deg)"},

"Phi_reel": {"type": "number",
        "Details": "[option] Elevation reel de la source (deg)"},

"Variance": {"type": "number",
        "Details": "[option] Variance de position utilisee pour la generation de la source"},

"X_theorique": {"type": "number",
        "Details": "[requis] Position theorique de la source en X (m)"},

"Y_theorique": {"type": "number",
        "Details": "[requis] Position theorique de la source en Y (m)"},

"Z_theorique": {"type": "number",
        "Details": "[requis] Position theorique de la source en Z (m)"},

"R_theorique": {"type": "number",
        "Details": "[option] Distance theorique de la source a l antenne (m)"},

"Theta_theorique": {"type": "number",
        "Details": "[option] Azimut theorique de la source (deg)"},

"Phi_theorique": {"type": "number",
        "Details": "[option] Elevation theorique de la source (deg)"},

"Version": {"type": "number",
        "Details": "[requis] Version de la base de donnee ayant servi a sauvegarder les informations"}

},
"additionalProperties": false,
"required": ["Nom", "Chemin", "Forme", "Enregistrement",
"Antenne", "X_reel", "Y_reel", "Z_reel", "X_theorique", "Y_theorique",
"Z_theorique", "Version","Frequence_echantillonnage","Canaux"]

}
```

Annexe D

Antenne MINI DSP

Features
- Multichannel USB mic array
- Onboard DSP for beamforming/ noise reduction / echo cancellation / de-reverb

Technical
- XMOS XVF3000 series
- USB 2.0 audio streaming
- Knowles SPH1668LM4H MEMS (7)
- Flexible I2S in/out
- PDM to I2S conversion on header
- Stackable add-on board
- 12 x RGB led

OS compatibility
- UAC2.0 with Windows ASIO driver
- OSx, Linux Alsa 2.0 compatible
- RPi step by step application notes

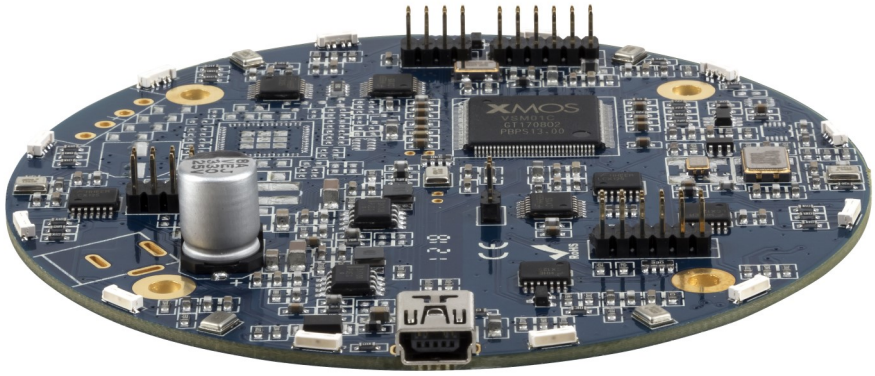
Power
- USB Bus powered
- DC power input option

Applications
- Voice activated projects
- Far field microphone application
- DIY mic array for Alexa/Cortana..
- Recording/conferencing
- Robotics/IoT/Smart home..

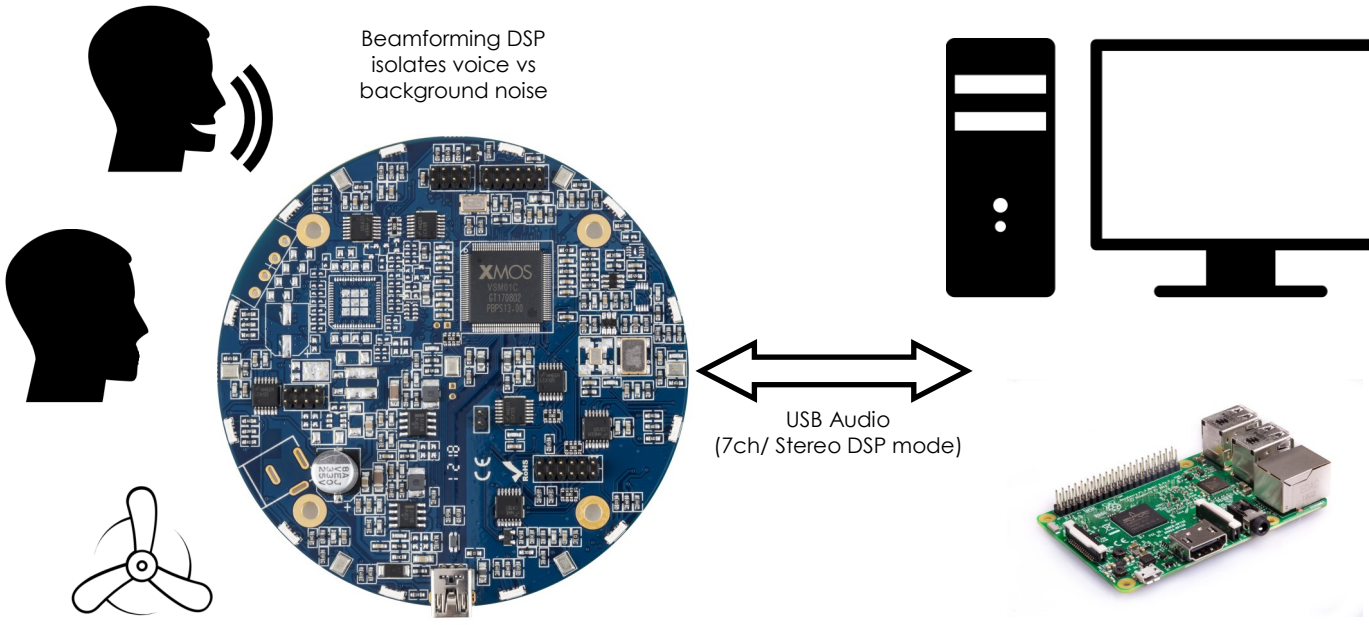
The **UMA-8** is a high-performance yet low cost multichannel USB microphone array built around XMOS multicore technology. Seven high-performance MEMS microphones are configured in a circular arrangement to provide high-quality voice capture for a wide range of applications.

Leveraging the onboard DSP processing from XMOS latest vocal fusion chip-sets, the **UMA-8** supports voice algorithms including beamforming, noise reduction, acoustic echo cancellation and de-reverb. The UMA-8 is a fully compliant UAC2 audio interface with driverless support for Mac/Linux and ASIO drivers for Windows.

From DIYers to OEM, this pocket-size platform is engineered for flexibility in firmware, software and hardware.



SYSTEM DIAGRAM



TECHNICAL SPECIFICATIONS

Item	Description
USB streaming engine	XMOS XVF3000 - Multicore USB audio processor with embedded DSP
USB audio capabilities	USB audio recording in 2 possible modes depending on firmware: - 8-channel mode (7 x MEMS installed + 1 x spare PDM port in the center) - Stereo recording with DSP processing enabled USB audio playback: Mono Audio on I2S out (e.g. external amplifier/DAC board.)
DSP processing (prebuilt firmware)	• Beamforming with configurable beam width (up to 20dB attenuation) • Perceptual acoustic echo cancellation (up to 80dB attenuation) • Noise suppression (up to 20dB attenuation) • De-reverb (up to 20dB attenuation) • Manual mode for control of beam forming
UAC2.0 drivers	Driverless interface for Mac OS X v10.6.4 and up Thesycon Windows ASIO driver (All versions) Linux Alsa 2.0 compliant
Resolution / Sample rate	24bit @ 11/16/32/44.1/48 kHz
I2S port	Output port for PDM to I2S conversion (upcoming firmware update required)
MEMS microphones	7 x Knowles SPH1668LM4H with low noise buffer and high performance modulator • Low distortion: 1.6% @ 120 dB SPL • High SNR: 65 dB and flat frequency response • RF shielded against mobile interference • Ominidirectional pick-up pattern
LED	12 x RGB LED / Bottom mounted - Circular light guide included
Expansion connector	2 x 12-pin, 2 mm pitch expansion connector for connectivity to hardware. XMOS JTAG connector for custom code.
Power supply	USB powered
Dimensions (diameter) mm	90 mm diameter / 20mm height with LED ring, 14mm height without LED ring

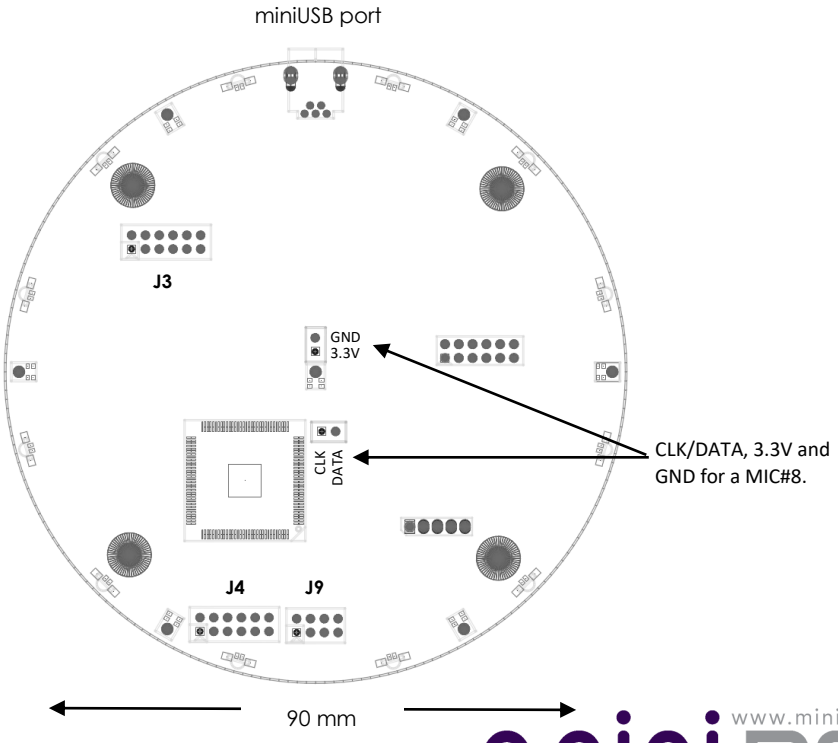
MECHANICAL DRAWINGS

J3 / Audio data & clocks

J3.1 - I2S_OUT_0	J3.2 - I2S_IN_0
J3.3 - I2S_OUT_1	J3.4 - I2S_IN_1
J3.5 - I2S_OUT_2	J3.6 - I2S_IN_2
J3.7 - I2S_OUT_3	J3.8 - I2S_OUT_4
J3.9 - MCLK	J3.10 - I2S_BCLK
J3.11 - GND	J3.12 - I2S_LRCLK

J4 / XMOS JTAG connector

J2.1 - GND	J2.2 - 3.3V
J2.3 - GND	J2.4 - 3.3V
J2.5 - N/A	J2.6 - UART_TX
J2.7 - UART_RX	J2.8 - XMOS_RST
J2.9 - I2C_SDATA	J2.10 - I2C_SCLK
J2.11 - N/A	J2.12 - N/A



Annexe E

Antenne CMA Cube

E.1 Définition géométrique

La position des microphones n'est pas facilement calculable. Une démonstration géométrique est donc nécessaire pour convertir les position des microphones sur les arrêtes du cubes à des positions dans le repère de la pièce. Pour aider la compréhension de la démonstration géométrique, on propose un schéma en figure E.1.

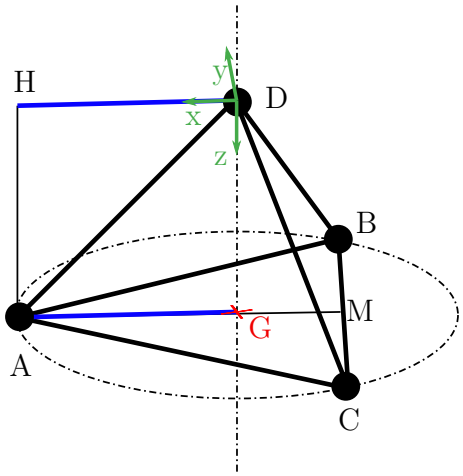


FIGURE E.1 – Schéma explicatif du calcul des positions des microphones

Contrairement à la configuration prévue initialement lors de son développement [14], l'axe $\overrightarrow{z}$ est orienté vers le bas. Ainsi, l'angle recherché est l'angle $\widehat{HDA}$, où $\overrightarrow{HD}$ correspond à la direction $\overrightarrow{y}$. Dans le traingle équilatéral (ABC), de coté unitaire, le centre de gravité G est situé au 2/3 de la hauteur. $AC = \sqrt{2}$.

Donc CM $= \frac{\sqrt{2}}{2}$, et AM $= \sqrt{\frac{3}{2}}$.
Or AG $= \frac{2}{3}\sqrt{\frac{3}{2}} = \sqrt{\frac{2}{3}}$.
Ainsi GM $= \frac{1}{3}\sqrt{\frac{3}{2}} = \frac{\sqrt{6}}{6}$.
Comme HD = AG $= \sqrt{\frac{2}{3}}$, HA $= \frac{\sqrt{3}}{3}$
Ainsi $\widehat{\text{HDA}} = \arccos(\sqrt{\frac{2}{3}})$.
Connaissant la longueur AD, la coordonnée en z est donc donnée par :$z_A = z_B = z_C = \frac{\sqrt{3}}{3}$DA
Comme l'axe $\overrightarrow{x}$ est // à (BC) et passe par G, les coordonnées des points sont :

$$A \begin{pmatrix} \sqrt{\frac{2}{3}} \\ 0 \\ \frac{\sqrt{3}}{3} \end{pmatrix}, \; B \begin{pmatrix} \frac{-\sqrt{6}}{6} \\ \frac{+\sqrt{2}}{2} \\ \frac{\sqrt{3}}{3} \end{pmatrix}, \; C \begin{pmatrix} \frac{-\sqrt{6}}{6} \\ \frac{-\sqrt{2}}{2} \\ \frac{\sqrt{3}}{3} \end{pmatrix}, \; D \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

On peut vérifier que $\left\|\overrightarrow{\text{OA}}\right\| = \left\|\overrightarrow{\text{OB}}\right\| = \left\|\overrightarrow{\text{OC}}\right\| = 1$

E.2 Définition des notations dans l'antenne

Ceci étant démontré, on peut utiliser ces résultats sur l'antenne CMA Cube. Les notations sont les suivantes : Les microphones sont notées de 0 à 6. Ces 7 microphones sont contenus dans 4 sondes numérotées de 13 à 16. Pour clarifier les correspondances de notations, les numéros de microphones et de sondes sont indiqués en figure E.2, et leurs coordonnées sont données directement dans le tableau E.1. La sonde 13 est enfoncée dans une mousse, de sorte à ce que le microphone 0 soit affleurant. Ce microphone correspond à l'origine du repère de l'antenne, et à l'un des angles du cube. Les sondes 14, 15 et 16 sont confondues avec les arêtes du cube. Ces 3 sondes contiennent 2 microphones, respectivement éloignés de 1,5 cm et 4,5 cm de l'angle du cube matérialisé par le microphone 0.

Mic.	Sonde	X (cm)	Y (cm)	Z (cm)
0	13	0	0	0
1	14	1,2	0	0,9
2	14	3,7	0	2,6
3	15	-0,6	1,1	0,9
4	15	-1,8	3,2	2,6
5	16	-0,6	-1,1	0,9
6	16	-1,8	-3,2	2,6

TABLE E.1 – Récapitulatif des coordonnées des microphones arrondies au mm

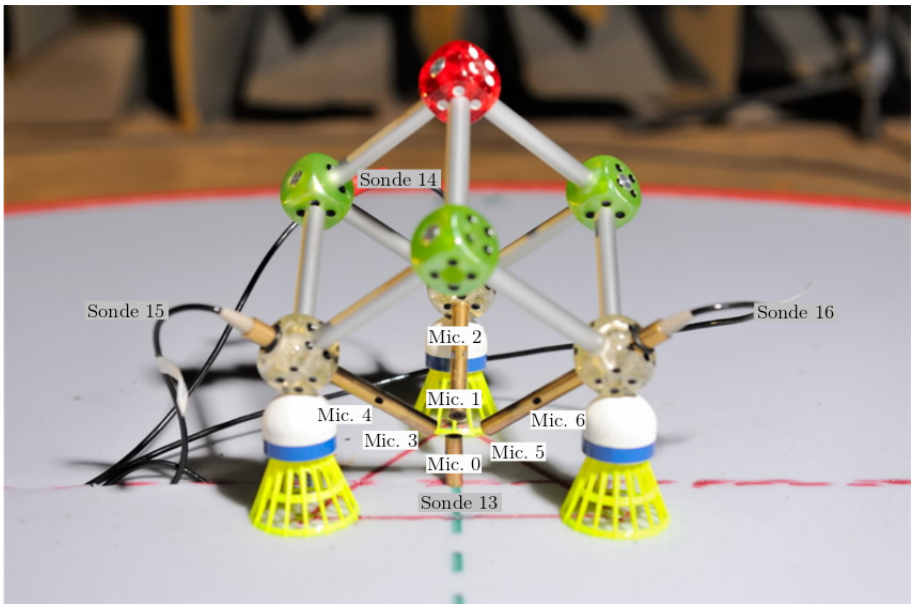


FIGURE E.2 – Photo de l'antenne CMA Cube avec numérotation des microphones

Annexe F

ZYLIA ZM-1 MICROPHONE

ZYLIA ZM-1 is a special type of microphone array that was designed for high quality multi-track audio recording. In order to capture selected sound sources on specified directions and distances the ZM-1 microphone uses 19 omnidirectional capsules (based on state-of-the-art MEMS technology).

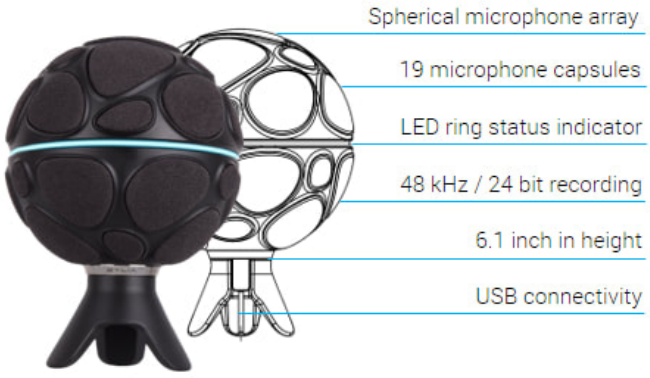


FIGURE F.1 – ZYLIA ZM-1 specification.

Physical Specifications :

— Size : height 155 mm (6.1 inches), length 103 mm (4 inches)

— Weight : 470 g (16.6 oz.)

— Stand threads : 1/4 inch and 5/8 inch

— Material of housing : ABS or anodized aluminum (limited edition only)

— Drivers for macOS (10.9 or later), Windows (7, 8.1, 10) and Linux (Debian)

ZM-1 Microphone Capsules ZM-1 is built using omnidirectional condenser microphone capsules based on Micro-Electro-Mechanical Systems (MEMS) technology. MEMS is driving the next evolvement in condenser microphones. These small-sized microphones take advantage of the enormous advances made in silicon technology over the past decades—including ultra-small fabrication geometries, excellent stability and repeatability of parameters, and low power consumption—all of which have become uncompromising requirements of the silicon industry. All capsules used in ZM-1 have very tight and time-constant tolerances, so that the sound is consistent from mic to mic and the highest performance of ZYLIA's DSP algorithms is guaranteed.

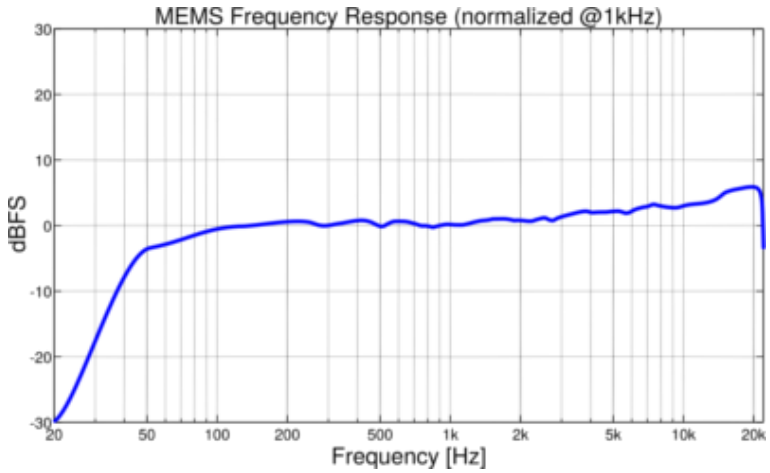


FIGURE F.2 – Frequency response of single capsule of the ZM-1.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19
X	0.0	32.7	-16.4	-16.3	6.3	36.6	36.5	6.2	-42.8	-42.7	-36.5	-6.2	42.8	42.7	-6.3	-36.6	-32.7	16.4	16.3
Y	0.0	0.1	28.3	-28.3	-45.8	-28.2	28.4	45.8	17.4	-17.6	-28.4	-45.8	-17.4	17.6	45.8	28.2	-0.1	-28.3	28.3
Z	49.0	36.5	36.5	36.5	16.3	16.3	16.3	16.3	16.3	16.3	-16.3	-16.3	-16.3	-16.3	-16.3	-16.3	-36.5	-36.5	-36.5

FIGURE F.3 – Microphone capsules placement - Cartesian coordinate system [mm]

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19
Radius [mm]	49	49	49	49	49	49	49	49	49	49	49	49	49	49	49	49	49	49	49
Elevation [deg]	90	48	48	48	19	19	19	19	19	19	-19	-19	-19	-19	-19	-19	-48	-48	-48
Azimuth [deg]	0	0	120	-120	-82	-38	38	82	158	-158	-142	-98	-22	22	98	142	-180	-60	60

FIGURE F.4 – Spherical coordinate system

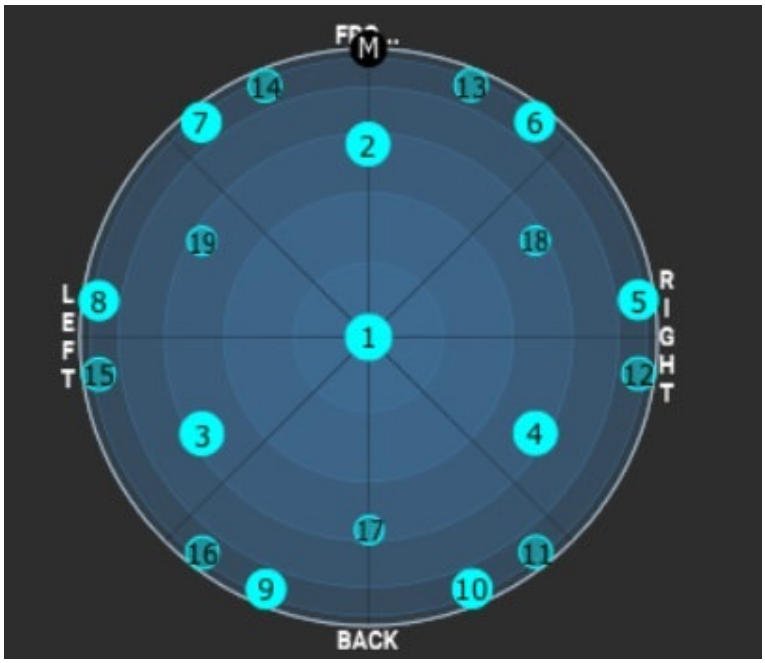


FIGURE F.5 – Visualization of the microphone capsules placement using IEM MultiEncoder.

le cnam

Hadrien PUJOL

Antennes microphoniques intelligentes : Localisation de sources acoustiques par Deep Learning

HESAM UNIVERSITÉ

Résumé : Pour ma thèse de doctorat, je propose d'explorer la piste de l'apprentissage supervisé, pour la tâche de localisation de sources acoustiques. Pour ce faire, j'ai développé une nouvelle architecture de réseau de neurones profonds. Mais, pour optimiser les millions de variables d'apprentissages de ce réseau, une base de données d'exemples conséquente est nécessaire. Ainsi, deux approches complémentaires sont proposées pour constituer ces exemples. La première est de réaliser des simulations numériques d'enregistrements microphoniques. La seconde, est de placer une antenne de microphones au centre d'une sphère de haut-parleurs qui permet de spatialiser les sons en 3D, et d'enregistrer directement sur l'antenne de microphones les signaux émis par ce simulateur expérimental d'ondes sonores 3D. Le réseau de neurones a ainsi pu être testé dans différentes conditions, et ses performances ont pu être comparées à celles des algorithmes conventionnels de localisation de sources acoustiques. Il en ressort que cette approche permet une localisation généralement plus précise, mais aussi beaucoup plus rapide que les algorithmes conventionnels de la littérature.

Mots clés : DOA 2D et 3D, Deep Learning, Convolution à trous, Base de données, Méthode des sources images, Retards fractionnaires, Spatialisation ambisonique, Antennes compactes

Abstract : For my PhD thesis, I propose to explore the path of supervised learning, for the task of locating acoustic sources. To do so, I have developed a new deep neural network architecture. But, to optimize the millions of learning variables of this network, a large database of examples is needed. Thus, two complementary approaches are proposed to constitute these examples. The first is to carry out numerical simulations of microphonic recordings. The second one is to place a microphone antenna in the center of a sphere of loudspeakers which allows to spatialize the sounds in 3D, and to record directly on the microphone antenna the signals emitted by this experimental 3D sound wave simulator. The neural network could thus be tested under different conditions, and its performances could be compared to those of conventional algorithms for locating acoustic sources. The results show that this approach allows a generally more precise localization, but also much faster than conventional algorithms in the literature.

Keywords : 2D and 3D DOA, Deep Learning, atrous convolutions, Sound source Datasets, Image source method, Fractional delays, Ambisonic spatialization, Compact microphone arrays