Localisation et rehaussement de sources de parole au format Ambisonique

Lauréline Perotin

UNIVERSITÉ DE LORRAINE

Thèse de Doctorat

Lauréline PEROTIN

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Lorraine
Mention Informatique

École doctorale : IAEM

Unité de recherche : Laboratoire Lorrain de Recherche en Informatique et ses Applications UMR 7503

Soutenue le 31 octobre 2019

Localisation et rehaussement de sources de parole au format Ambisonique

Analyse de scènes sonores pour faciliter la commande vocale

JURY

Rapportrice :	**Dorothea KOLOSSA**, Professeure, Ruhr-Universität Bochum, Allemagne
Rapporteur :	**Laurent GIRIN**, Professeur, Grenoble INP, France
Examinatrice :	**Christine EVERS**, EPSRC Research Fellow, Imperial College London, Royaume-Uni
Directeur de thèse :	**Emmanuel VINCENT**, Directeur de recherche, Inria Nancy - Grand Est, France
Co-directeur de thèse :	**Romain SERIZEL**, Maître de conférences, Université de Lorraine Loria, Nancy, France
Co-encadrant de thèse :	**Alexandre GUÉRIN**, Ingénieur de recherche, Orange Labs, Cesson Sévigné, France

Résumé

Cette thèse s'inscrit dans le contexte de l'essor des assistants vocaux mains libres. Dans un environnement domestique, l'appareil est généralement posé à un endroit fixe, tandis que le locuteur s'adresse à lui depuis diverses positions, sans nécessairement s'appliquer à être proche du dispositif, ni même à lui faire face. Cela ajoute des difficultés majeures par rapport au cas, plus simple, de la commande vocale en champ proche (pour les téléphones portables par exemple) : ici, la réverbération est plus importante ; des réflexions précoces sur les meubles entourant l'appareil peuvent brouiller le signal ; les bruits environnants sont également sources d'interférences. À ceci s'ajoutent de potentiels locuteurs concurrents qui rendent la compréhension du locuteur principal particulièrement difficile. Afin de faciliter la reconnaissance vocale dans ces conditions adverses, plusieurs prétraitements sont proposés ici. Nous utilisons un format audio spatialisé, le format Ambisonique, adapté à l'analyse de scènes sonores. Dans un premier temps, nous présentons une méthode de localisation des sources sonores basée sur un réseau de neurones convolutif et récurrent. Nous proposons des descripteurs inspirés du vecteur d'intensité acoustique qui améliorent la performance de localisation, notamment dans des situations réelles où plusieurs sources sont présentes et l'antenne de microphones est posée sur une table. La technique de visualisation appelée *layerwise relevance propagation* (LRP) met en valeur les zones temps-fréquence positivement corrélées avec la localisation prédite par le réseau dans un cas donné. En plus d'être méthodologiquement indispensable, cette analyse permet d'observer que le réseau de neurones exploite principalement les zones dans lesquelles le son direct domine la réverbération et le bruit ambiant.
Dans un second temps, nous proposons une méthode pour rehausser la parole du locuteur principal et faciliter sa reconnaissance. Nous nous plaçons dans le cadre de la formation de voies basée sur des masques temps-fréquence estimés par un réseau de neurones. Afin de traiter le cas où plusieurs personnes parlent à un volume similaire, nous utilisons l'information de localisation pour faire un premier rehaussement à large bande dans la direction du locuteur cible. Nous montrons que donner cette information supplémentaire au réseau n'est pas suffisant dans le cas où deux locuteurs sont proches ; en revanche, donner en plus la version rehaussée du locuteur concurrent permet au réseau de renvoyer de meilleurs masques. Ces masques permettent d'en déduire un filtre multicanal qui améliore grandement la reconnaissance vocale. Nous évaluons cet algorithme dans différents environnements, y compris réels, grâce à un moteur de reconnaissance de la parole utilisé comme boîte noire.
Dans un dernier temps, nous combinons les systèmes de localisation et de rehaussement et nous évaluons la robustesse du second aux imprécisions du premier sur des exemples réels.

Abstract

This work was conducted in the fast-growing context of hands-free voice command. In domestic environments, smart devices are usually laid in a fixed position, while the human speaker gives orders from anywhere, not necessarily next to the device, or nor even facing it. This adds difficulties compared to the problem of near-field voice command (typically for mobile phones) : strong reverberation, early reflections on furniture around the device, and surrounding noises can degrade the signal. Moreover, other speakers may interfere, which make the understanding of the target speaker quite difficult.
In order to facilitate speech recognition in such adverse conditions, several preprocessing methods are introduced here. We use a spatialized audio format suitable for audio scene analysis : the Ambisonic format. We first propose a sound source localization method that relies on a convolutional and recurrent neural network. We define an input feature vector inspired by the acoustic intensity vector which improves the localization performance, in particular in real conditions involving several speakers and a microphone array laid on a table. We exploit the visualization technique called layerwise relevance propagation (LRP) to highlight the time-frequency zones that are correlate positively with the network output. This analysis is of paramount importance to establish the validity of a neural network. In addition, it shows that the neural network essentially relies on time-frequency zones where direct sound dominates reverberation and background noise.
We then present a method to enhance the voice of the main speaker and ease its recognition. We adopt a mask-based beamforming framework based on a time-frequency mask estimated by a neural network. To deal with the situation of multiple speakers with similar loudness, we first use a wideband beamformer to enhance the target speaker thanks to the associated localization information. We show that this additional information is not enough for the network when two speakers are close to each other. However, if we also give an enhanced version of the interfering speaker as input to the network, it returns much better masks. The filters generated from those masks greatly improve speech recognition performance. We evaluate this algorithm in various environments, including real ones, with a black-box automatic speech recognition system.
Finally, we combine the proposed localization and enhancement systems and evaluate the robustness of the latter to localization errors in real environments.

Remerciements

La liste de ceux qui ont illuminé mes trois dernières années est longue.

J'ai eu la chance, dans tous mes bureaux, de trouver des environnements conviviaux et des collègues remarquablement sympas. Merci en particulier à mes encadrants : Alexandre, pour ta présence chaleureuse et ta patience sans faille ; Romain, pour ton infatigable disponibilité ; Emmanuel, pour la pertinence de chacune de tes remarques, toutefois apportées avec délicatesse, ainsi que pour ton éthique.
J'ai apprécié tous les moments, cafés, concerts, footings, dîners partagés avec les membres de l'équipe TPS, de Multispeech, et de mon équipe d'adoption, Panama. Au passage, bravo à Stéphane et Amélie d'avoir osé se plonger dans mon code et d'en être ressorti·e·s indemnes. L'aide de l'équipe MAS m'a également été très précieuse.
Lors de mes journées de travail hors les murs, je suis fière d'avoir fait partie du groupe prometteur des Young Bold Researchers in Audio Signal Processing aux côtés d'Alex, Neil, Guillaume et Pierre.
Les conférences participent également à l'affection que je garderai pour le monde de la recherche. Je n'oublierai pas les échanges ouverts et curieux entre chercheur·se·s peu soucieux·ses de leur h-index. Patrick, you showed me that even great researchers are human beings, with their doubts, passions, and, in your case, an infinite kindness. Antoine, ton enthousiasme a failli me contaminer l'espace d'un instant. Les mots de soutien que tu m'as glissés, notamment lors de cet après-midi à Calgary, ont eu un impact que tu ne soupçonnes probablement pas. Delia, please keep on bringing this salutary energy to the world of research, the mix of consciousness and fantasy that it needs. J'ai aussi eu la chance d'y rencontrer une partie du jury qui évaluera ce travail et avec lequel j'ai hâte d'échanger lors de ma soutenance (et de son pot).

Merci à ma famille, de me soutenir et me supporter dans mes lubies, sans trop s'inquiéter, et parfois même en m'y accompagnant.

Mon quotidien aurait été bien terne sans les copains et les copines. Celleux qui ont ramé en même temps que moi en thèse, Marie, Jm, Vincent, Mika, Marti, Laura. Ça sera mieux après, vous verrez. Diego, Antoine et Clément, qui ont supporté ma mauvaise humeur à l'escalade ; qui, avec Quentin et Julian, m'ont suivie aux fêtes des galettes du mondes à la châtaigne et à la potion magique. Merci aux canapés qui m'ont accueillie lorsque mes pulsions de voyage étaient trop forte, et surtout à toi, Nico, de m'ouvrir ta porte même dans des conditions absurdes en dernière minute. À celles qui m'ont fait apercevoir d'autres mondes, telles des bouffées d'oxygène : Betty, Léa, Clara, Iona. À

Marc et Béa, de m'avoir fait me sentir chez moi à mon arrivée à Rennes, et celleux qui ont suivi, Eve, Pauline (les deux), Sandra, Orane, et les autres déjà cité·e·s.

Il me faudrait un manuscrit entier pour exprimer ma reconnaissance envers les burners, nobodies et autres spationautes. L'oasis que vous m'avez manutée dans ce désert fut vitale.

Merci aux lieux qui m'ont accueillie pour travailler lorsque mon allergie au bureau se faisait trop sévère : le café des champs libres, Block'out, la SNCF, la maison jaune, et, surtout, la Part des Anges où j'espère bien passer encore de nombreuses heures.

Enfin, il y a celleux pour lesquels je peine à poser les mots. Je compte sur vous pour percevoir la profondeur de ma gratitude. Firas, te voir en rentrant à l'appart fut un bonheur chaque soir renouvelé. Alex, tu as été un soutien sans faille dans toutes les situations que j'ai rencontrées ces trois dernière années, de l'écriture d'article à la gestion de la survie de 45 personnes dans le désert, en passant par nos premières expériences de DJs en public. Maelle, tu me rappelles sans cesse à l'essentiel, et tu as fait de ma rédaction une période merveilleuse, chose rare pour un·e· thésard·e·.

Note Compte tenu de la pénibilité d'utiliser l'écriture inclusive en LaTeX, ce manuscrit est seulement rédigé en écriture alternée. Il sera indifféremment fait référence aux locuteurs ou aux locutrices, sans que cela n'ait de rapport avec le contenu scientifique lui-même.

Table des matières

Table des figures

Liste des tableaux

Acronymes

ACI analyse en composantes indépendantes
BAN blind analytic normalization
CNN convolutional neural network
CRNN convolutional and recurrent neural network
CSIPD cosine-sine inter-phase difference
DNN deep neural network
DMR direct-to-mixture ratio
EM espérance-maximisation
ESPRIT estimation of signal parameters via rotational invariance techniques
FF feed-forward
FOA first-order Ambisonics
GAN generative adversarial networks
GCC-PHAT generalized cross correlation phase transphorm
GEV generalized eigenvalue
GEVD generalized eigenvalue decomposition
GLLiM Gaussian locally linear mapping
GMM Gaussian mixture model
HMM hidden Markov model
HOA high-order Ambisonics
ILD interaural level difference
ITD interaural time difference
LRP layerwise relevance propagation
LSTM long short-term memory
MFCC Mel frequency cepstral coefficient
MUSIC multiple signal classification
MVDR minimum variance distortionless response
MWF multichannel Wiener filter
MSE mean square error
NMF non-negative matrix factorization

PIT permutation invariant training
Q-CRNN quaternion convolutional and recurrent neural network
RAP reconnaissance automatique de la parole
ReLU rectified linear unit
RGPD réglement général sur la protection des données
RNN recurrent neural network
SDW-MWF speech distortion weighted multichannel Wiener filter
SIR signal-to-interference ratio
SNR signal-to-noise ratio
SRIR spatial room impulse response
SRP steered response power
SRP-PHAT steered response power with phase transform
TR60 temps de réverbération à 60 dB
TFCT transformée de Fourier à court terme
VVM velocity vector module
WER word error rate
WPE weighted prediction error

1. Introduction

Ce chapitre décrit le contexte scientifique et industriel dans lequel s'inscrit cette thèse, en particulier l'émergence des assistants personnels à commande vocale et les difficultés techniques soulevées par la reconnaissance vocale en environnement domestique. Nous définissons le cadre d'étude de la thèse et les objectifs fixés dans ce scénario. Enfin, nous détaillons l'organisation de ce document et présentons les principales contributions de la thèse.

1.1. Motivation et cadre

Séparation de sources audio et rehaussement de la parole Un individu essaye de comprendre ce que dit son interlocuteur au milieu d'un bar bondé. Une productrice de musique souhaite accéder à un instrument particulier dans un morceau pour utiliser ce sample dans sa prochaine création. Un malentendant appareillé se concentre sur les lèvres de son amie pour comprendre ses dires malgré la personne qui téléphone à côté d'eux. Un ingénieur du son veut récupérer tous les éléments d'une bande son de film afin d'en refaire un mixage spatialisé, qui permettra d'apprécier le film sur un système 5.1. Dans toutes ces situations, le problème est le même : séparer plusieurs sources sonores alors que l'on capte seulement leur mélange. Lorsque l'objectif est d'accéder à tous les constituants du mélange, comme par exemple dans le cas de la bande son de film dont on a besoin de chacun des éléments, on parle de séparation de sources. Lorsqu'une seule source doit être extraite d'un bruit de fond, on parle de débruitage. Si la ou les sources d'intérêt sont des sources de parole que l'on chercher à extraire, débruiter, ou déréverbérer, on parle de rehaussement de la parole.

Reconnaissance de la parole en champ lointain « Jarvis, active le bouclier. Jarvis ? JARVIS ! ». On imagine les conséquences pour Tony Stark, alias Iron Man, si l'intelligence artificielle qui commande son armure était incapable de comprendre ses ordres en présence de bruit ambiant (souvent présent au cœur de l'action). Science-fiction ? Certes, mais dont la réalité se rapproche avec l'apparition récente des assistant vocaux. Jarvis a cependant l'avantage d'avoir accès à un microphone qui enregistre la voix du donneur d'ordre au plus proche de sa bouche. C'est aussi le cas lorsque l'on commande vocalement un smartphone placé en face de soi. Le signal est alors relativement « propre » et les technologies actuelles de reconnaissance automatique de la parole (RAP) sont suffisamment performantes pour permettre la compréhension de l'ordre donné.
La situation se complique grandement pour les nouveaux assistants personnels, tels Google Home ou Amazon Echo, qui ont vocation à être placé dans la maison et aux-

quels on parle en en oubliant l'existence physique. De par la distance entre le locuteur et l'assistant, ce dernier capte tous les bruits ambiants en plus de la voix principale. Les autres sources de parole sont particulièrement gênantes pour la compréhension, ainsi qu'une réverbération importante, ou des réflexions du champ sonore si l'objet est placé sur une table ou dans un meuble, ou si le locuteur s'adresse à lui sans lui faire face (fonctionnalité indispensable pour obtenir des réponses aux questions qui l'assaillent lorsqu'il fait la vaisselle). Dans ces cas, il est fréquent que l'assistant ne comprenne pas un ordre. Si c'est généralement moins vital que pour Tony Stark, cela peut tout de même s'avérer problématique, comme par exemple lorsque les détenteurs d'Amazon Echo furent pris de sueurs froides en entendant un rire sardonique s'élever à l'improviste dans leur salon, à cause de la similarité phonétique de l'ordre « *Alexa laugh* » avec des phrases de la vie quotidienne [1], qui ont pu être mal interprétées en présence de bruit. Ce problème a été résolu par Amazon, dont l'intelligence artificielle ne rit maintenant qu'à la demande « *Alexa, can you laugh ?* ». Mais traiter une par une toutes les expressions paronymes dans l'implémentation de l'intelligence est sous-optimal.
Pour faciliter la compréhension des assistants vocaux en champ lointain, la stratégie favorisée aujourd'hui est de nettoyer la scène sonore captée en rehaussant la source d'intérêt, afin de transmettre au moteur de RAP un signal plus facile à interpréter sans ambiguïtés.

Scénario considéré et objectifs Pour permettre l'utilisation des travaux de cette thèse dans un contexte industriel, nous définissons un cadre applicatif précis. Il s'agit de mettre au point une méthode de rehaussement de la parole d'un locuteur cible, assimilé au donneur d'ordre, afin de faciliter la RAP qui s'ensuit. Nous nous intéressons en particulier à la présence simultanée de plusieurs sources de parole. Afin de s'assurer que la méthode est exploitable en pratique, des tests doivent être menés en conditions acoustiques domestiques réalistes, voire réelles. Les performances seront mesurées en terme de taux d'erreur sur les mots d'un système de RAP utilisé en aval comme une boîte noire.

1.2. Outils utilisés

Les réseaux de neurones La majorité des fonctionnalités d'« intelligence artificielle », de l'analyse d'images à la RAP [1], reposent aujourd'hui sur des réseaux de neurones artificiels. Il permettent de modéliser des fonctions complexes en optimisant un grand nombre de paramètres. Dans le cas de l'apprentissage supervisé, cela se fait grâce à des données dites d'apprentissage pour lesquelles la réponse désirée est connue. Ce type d'apprentissage a prouvé son efficacité dans la dernière décennie grâce à l'amélioration de la puissance de calcul, permettant de traiter de manière efficace un grand nombre de données et modéliser des fonctions complexes avec des réseaux de neurones « profonds ». De plus en plus souvent, ceux-ci sont capables de généraliser leurs estimations aux cas réels, c'est-à-dire s'adapter à une variété de cas de figure qui n'a pas été rencontrée pendant l'apprentissage. Ce progrès a également été rendu possible par des améliorations algo-

1. https://mashable.com/2018/03/06/amazon-echo-alexa-random-laugh/

rithmiques de l'apprentissage, en particulier pour les réseaux de neurones très profonds. Pour les problèmes de séparation de sources et de rehaussement de la parole, traditionnellement abordés par des méthodes de traitement du signal, les réseaux de neurones se sont également révélés être de valeureux alliés. Ils ont permis des progrès très substantiels concernant le débruitage de la parole monocanal en présence de bruits variés [2]. Dans le cas de plusieurs sources de paroles simultanées et/ou d'enregistrement multicanaux, le problème est plus complexe à résoudre, mais les progrès dans les autres champs du traitement du signal laissent à penser que la combinaison des réseaux de neurones à des outils de traitement du signal est l'approche la plus appropriée.
Un défaut des réseaux de neurones est que leurs résultats sont aujourd'hui difficiles à expliquer. Or, il est fondamental de comprendre l'origine de tel ou tel résultat, non seulement pour améliorer notre compréhension du phénomène physique étudié, mais surtout pour s'assurer de la robustesse du réseau de neurones et du fait que ce résultat est fondé sur une bonne modélisation du problème et non sur un biais au moment du test. Des techniques de visualisation ont été mises au point pour analyser le fonctionnement des réseaux de neurones. Nous en appliquerons une en particulier, appelée *layerwise relevance propagation* (LRP).

La localisation des sources La dénomination « réseau de neurones » traduit l'ambition originelle d'imiter les mécanismes d'apprentissage et de raisonnement du cerveau humain. Ce n'est plus exactement le cas avec les architectures récentes de réseaux. Cependant, les parallèles entre cognition et apprentissage automatique restent pertinents pour déterminer les informations nécessaires à la réalisation d'une tâche. Dans notre cas, pour isoler une voix du reste de la scène sonore, le cerveau utilise des indices de natures très variées. Les caractéristiques acoustiques du signal ou la sémantique sont par exemple exploitées par le cerveau et utilisées dans les algorithmes de traitement de la parole via la représentation temps-fréquence et les modèles de langage.
Un autre type d'information acoustique est primordial pour l'auditeur humain : la spatialité de l'environnement sonore. Par exemple, il est beaucoup plus facile de distinguer les différents instruments d'une chanson lorsqu'elle est mixée en stéréo plutôt qu'en mono. De même, lorsque l'enregistrement disponible contient une information spatiale, ce qui nécessite qu'il ait été fait à l'aide de plusieurs microphones (on parle de signal multicanal), le rehaussement de la parole est grandement facilité. Si, de surcroît, l'auditeur est face à un concert acoustique où il voit les différents instruments et leurs emplacements, il lui est encore plus facile de focaliser son attention sur un instrument en particulier. C'est également le cas en traitement du signal audio : connaître la localisation des sources dans la scène sonore facilite l'élimination des sources indésirables et le rehaussement des sources d'intérêt. Une partie de cette thèse visera donc à obtenir cette information de localisation.

Le format ambisonique Pour contenir l'information spatiale, l'enregistrement de la scène sonore doit être multicanal. Dans cette thèse, nous travaillons avec un format multicanal particulier, le format ambisonique. Il est obtenu à partir d'une antenne sphérique

de microphones dont les captations sont combinées afin d'obtenir des canaux facilement interprétables en termes de spatialité. Au niveau de l'analyse, ce format simplifie la localisation des sources dans la scène sonore. Il présente également l'avantage d'être isotropique, c'est-à-dire que toutes les directions de l'espace sont représentées avec une égale précision. C'est aussi un format approprié à la manipulation de scènes sonores : les rotations de scène ou la focalisation spatiale sont particulièrement simples et rapides. C'est pourquoi, malgré la légère perte d'information entre la captation originelle de l'antenne sphérique et la représentation ambisonique, ce format est de plus en plus plebiscité par les industriels comme Youtube, Facebook, ou Orange qui a fortement contribué à son développement. Il a également été intégré dans la norme audio 3D MPEG-H [3]. Cependant, les outils pour l'analyse de scène ambisonique, que ce soit pour la localisation ou le rehaussement, sont encore peu développés, et la puissance des réseaux de neurones commence seulement à être mise au service de ce format. Nous nous limiterons ici à ce cadre ambisonique.

1.3. Contributions et plan du document

Outre cette introduction, ce manuscrit est constitué d'une partie rappelant des notions préliminaires et l'état de l'art du domaine, suivie d'une partie présentant les contributions apportées par ce travail et leur validation expérimentale, avec notamment l'évaluation du traitement de bout-en-bout d'une scène sonore réelle depuis sa captation jusqu'au module de RAP.

Le chapitre 2 décrit en détail le formalisme ambisonique d'un point de vue mathématique, l'obtention de ce format à partir d'enregistrements réels et ses limites, ainsi que l'expression des grandeurs acoustiques de base dans ce formalisme et les avantages qui en découlent.

Le chapitre 3 expose l'état de l'art pour les différentes étapes du problème. Un rapide aperçu des techniques de reconnaissance vocale justifie la stratégie établie pour les prétraitements mis au point. S'ensuivent les présentations des techniques traditionnelles et actuelles de rehaussement de la parole et de localisation de sources audio. Enfin, nous donnons un aperçu des techniques de visualisation et d'analyse des réseaux de neurones, plus particulièrement la LRP. Nous précisons enfin les techniques sur lesquelles nous nous appuyons dans le cadre de cette thèse et les contributions proposées pour dépasser certaines de leurs limitations.

Le chapitre 4 présente la solution proposée pour localiser plusieurs locuteurs à partir d'un enregistrement ambisonique à l'aide d'un réseau de neurones convolutif et récurrent (CRNN, *convolutional recurrent neural network*). Elle s'appuie sur l'utilisation de la formulation ambisonique du vecteur d'intensité acoustique, ce qui assure un gain de robustesse en conditions réelles. Les performances sont mesurées dans des environnements

simulés et réels et les résultats analysés par LRP. Enfin, l'impact des différents paramètres est évalué.

Le chapitre 5 est dédié à la méthode mise au point pour rehausser un locuteur cible, connaissant sa position, notamment en présence d'un autre locuteur parlant à volume égal. Elle s'appuie sur l'utilisation d'un réseau de neurones récurrent (RNN, *recurrent neural network*) pour estimer un masque temps-fréquence correspondant à la parole cible, afin de pouvoir utiliser des techniques de filtrage de Wiener multicanales (MWF, *multichannel Wiener filter*). Les performances seront mesurées en terme de taux d'erreur sur les mots (WER, *word error rate*) du module de RAP choisi, et ce dans différents types d'environnements acoustiques. Ici encore, différents paramètres d'apprentissage et de filtrage seront discutés. D'autre part, nous combinons les modules de localisation et de rehaussement, en y ajoutant une étape intermédiaire de suivi des sources sonores. Cela permet d'évaluer la mise en œuvre du système global en situation réelle.

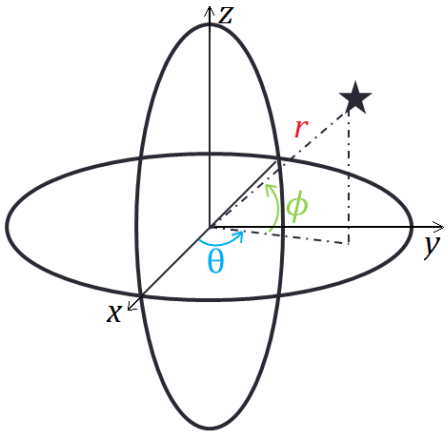
Le chapitre 6, enfin, rappelle les points principaux de ce manuscrit, y apporte une conclusion ainsi que des propositions de poursuite de ce travail de recherche.

2. Représentation ambisonique d'une scène sonore

Ce chapitre présente le format ambisonique ainsi que l'expression dans ce formalisme des grandeurs acoustiques dont nous nous servirons par la suite.

2.1. Décomposition du champ acoustique sur les harmoniques sphériques

2.1.1. Les fonctions harmoniques sphériques



$$\begin{cases} x = r\cos\theta\cos\phi \\ y = r\sin\theta\cos\phi \\ z = r\sin\phi \end{cases} \tag{2.1}$$

FIGURE 2.1. – Système de coordonnées sphériques : tout point de l'espace est repéré par sa distance à l'origine r, son azimut θ et son élévation ϕ. Le système d'équations (2.1) décrit le lien entre ces coordonnées et les coordonnées cartésiennes (x, y, z).

Les coordonnées sphériques (r, θ, ϕ) (distance à l'origine, azimut et élévation) utilisées pour décrire l'espace sont présentées sur la Figure 2.1. Décrire le champ acoustique dans ce système de coordonnées par le biais des fonctions harmoniques sphériques permet une représentation spatiale explicite et manipulable simplement. Les harmoniques sphériques peuvent être décrites dans leur version complexe ou réelle ; historiquement, les deux écoles coexistent [4, 5]. Nous relions ici les deux descriptions entre elles.
Les harmoniques sphériques complexes sont une famille de fonctions de (θ, ϕ) caractérisées

par leur degré $n \in \mathbb{N}$ et leur ordre $m \in \{-n, -n+1, ..., n-1, n\}$ [4, p. 190] :

$$Y_{nm}(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_{nm}(\sin\phi) e^{im\theta}, \tag{2.2}$$

où (.)! représente la fonction factorielle, les $P_{nm}(.)$ sont les polynômes de Legendre associés, et $i = \sqrt{-1}$.
On peut également définir les harmoniques sphériques réelles de degré $n \in \mathbb{N}$ et d'ordres $m \in \{0, 1, ..., n\}$ [6, p. 304] :

$$\tilde{Y}^{\sigma}_{nm}(\theta, \phi) = \begin{cases} (-1)^m \sqrt{2n+1} \tilde{P}_{nm}(\sin\phi)\cos(m\theta) \text{ si } \sigma = 1 \\ (-1)^m \sqrt{2n+1} \tilde{P}_{nm}(\sin\phi)\sin(m\theta) \text{ si } \sigma = -1, \end{cases} \tag{2.3}$$

où $\sigma \in \{-1, 1\}$ (sauf en $m = 0$ où $\sigma = 1$) et $\tilde{P}_{nm}(.)$ désigne la version normalisée des polynômes de Legendre associés :

$$\tilde{P}_{nm}(x) = \sqrt{\epsilon_m \frac{(n-m)!}{(n+m)!}} P_{nm}(x) \text{ avec } \begin{cases} \epsilon_0 = 1 \\ \epsilon_{m \geqslant 1} = 2. \end{cases} \tag{2.4}$$

Les harmoniques sphériques réelles peuvent être exprimées en fonction des parties réelles et imaginaires des harmoniques sphériques complexes :

$$\tilde{Y}^{\sigma}_{nm}(\theta, \phi) = (-1)^m \sqrt{\epsilon_m 4\pi} \begin{cases} \mathcal{R}[Y_{nm}(\theta, \phi)] \text{ si } \sigma = 1 \\ \mathcal{I}[Y_{nm}(\theta, \phi)] \text{ si } \sigma = -1, \end{cases} \tag{2.5}$$

où $\mathcal{R}(.)$ et $\mathcal{I}(.)$ désignent les parties réelle et imaginaire d'un nombre complexe.
La relation inverse s'écrit en utilisant la propriété aisément vérifiable $Y^*_{nm} = (-1)^m Y_{n\,-m}$, où $(.)^*$ désigne la conjugaison complexe :

$$Y_{nm}(\theta, \phi) = \frac{1}{\sqrt{\epsilon_m 4\pi}} \begin{cases} \tilde{Y}^{+1}_{-nm}(\theta, \phi) - i\tilde{Y}^{-1}_{-nm}(\theta, \phi) \text{ si } m < 0 \\ (-1)^m [\tilde{Y}^{+1}_{nm}(\theta, \phi) + i\tilde{Y}^{-1}_{nm}(\theta, \phi)] \text{ si } m \geqslant 0. \end{cases} \tag{2.6}$$

2.1.2. Transformée de Fourier sphérique

Les harmoniques sphériques complexes $Y_{nm}(\theta, \phi)$ (2.2) forment une base orthonormée [7, p. 16] de l'espace hermitien $\mathcal{L}^H_2$ des fonctions de (θ, ϕ) définies sur la sphère unité et dont le module au carré est intégrable sur la sphère, muni du produit hermitien défini par :

$$\langle f | g \rangle_{\mathcal{L}^H_2} = \frac{1}{4\pi} \int_{\theta=0}^{2\pi} \int_{\phi=-\pi/2}^{\pi/2} f(\theta, \phi) g(\theta, \phi)^* \cos\phi d\phi d\theta. \tag{2.7}$$

Toute fonction complexe $f(\theta, \phi)$ de $\mathcal{L}^H_2$ peut être décomposée sur cette base :

$$f(\theta, \phi) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} f_{nm} Y_{nm}(\theta, \phi). \tag{2.8}$$

De même, les harmoniques sphériques réelles $\tilde{Y}_{nm}^{\sigma}(\theta,\phi)$ (2.3) forment une base orthonormée [8, p. 18] de l'espace vectoriel $\mathcal{L}_2$ des fonctions réelles de carré intégrable sur la sphère unité, muni du produit scalaire suivant :

$$\langle f|g\rangle_{\mathcal{L}_2} = \frac{1}{4\pi}\int_{\theta=0}^{2\pi}\int_{\phi=-\pi/2}^{\pi/2} f(\theta,\phi)g(\theta,\phi)\cos\phi d\phi d\theta. \tag{2.9}$$

Toute fonction réelle $f(\theta,\phi)$ de $\mathcal{L}_2$ peut être décomposée sur cette base :

$$f(\theta,\phi) = \sum_{n=0}^{+\infty}\left[f_{n0}^{1}\tilde{Y}_{n0}^{1}(\theta,\phi) + \sum_{m=1}^{n}\sum_{\sigma=\pm1} f_{nm}^{\sigma}\tilde{Y}_{nm}^{\sigma}(\theta,\phi)\right]. \tag{2.10}$$

Cette projection peut aussi être considérée comme une décomposition en série de Fourier sphérique (appelée de manière abusive transformée de Fourier sphérique) de la fonction $f(\theta,\phi)$. Les coefficients vérifient alors :

$$\begin{aligned} f_{nm}^{\sigma} &= \langle f(\theta,\phi)|\tilde{Y}_{nm}^{\sigma}(\theta,\phi)\rangle \\ &= \frac{1}{4\pi}\int_{\theta=0}^{2\pi}\int_{\phi=-\pi/2}^{\pi/2} f(\theta,\phi)\tilde{Y}_{nm}^{\sigma}(\theta,\phi)\cos\phi d\phi d\theta. \end{aligned} \tag{2.11}$$

Le format ambisonique s'appuie sur cette décomposition pour représenter le signal en reflétant les propriétés spatiales du champ acoustique, indépendamment du mode d'enregistrement. Les différents canaux du format ambisonique représentent les coefficients de Fourier sphériques du champ de pression $p(\theta,\phi)$ à la surface d'une sphère [7, 6].

2.1.3. Équation de propagation des ondes

Pour trouver l'expression de ces coefficients, considérons le cas du champ de pression dû à la propagation d'une onde plane harmonique (contenant une seule fréquence), d'amplitude S_ω, de nombre d'onde $k = \frac{\omega}{c}$, avec ω la pulsation de l'onde et c sa vitesse de propagation, et provenant d'une direction (θ_i,ϕ_i). Le champ de pression complexe en tout point $\mathbf{r} = (r,\theta,\phi)$ de l'espace à l'instant τ s'écrit sous la forme

$$p_\omega(\mathbf{r},\tau) = S_\omega e^{ik\mathbf{r}.\mathbf{u}_i}e^{i\omega\tau}, \tag{2.12}$$

où $\mathbf{u}_i = (1,\theta_i,\phi_i)$ est le vecteur unité pointant dans la direction d'arrivée de l'onde. Les résultats sont facilement généralisables au cas d'une combinaison d'ondes harmoniques grâce à la transformée de Fourier temporelle, qui permet d'écrire

$$p(\mathbf{r},\tau) = \int_{\omega=0}^{+\infty} p_\omega(\mathbf{r},\tau)d\omega. \tag{2.13}$$

L'équation de propagation des ondes pour une telle onde plane en coordonnées sphériques s'écrit de la manière suivante :

$$\triangle_{\mathbf{r}}\, p_\omega(\mathbf{r},\tau) - \frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}p_\omega(\mathbf{r},\tau) = 0. \tag{2.14}$$

où $\triangle_{\mathbf{r}}$ désigne l'opérateur Laplacien en coordonnées sphériques. Cette équation peut être résolue en supposant la séparation des variables dans l'expression de p_ω :

$$p_\omega(\mathbf{r},\tau) = S_\omega R(r)\Theta(\theta)\Phi(\phi)T(\tau). \tag{2.15}$$

En injectant cette expression dans l'équation (2.14), on obtient les quatre équations partielles suivantes [7, p. 33] :

$$\begin{cases} \dfrac{d^2T}{d\tau^2} + \omega^2 T = 0 \\ \dfrac{d^2\Theta}{d\theta^2} + m^2\Theta = 0 \text{ pour tout } m \in \mathbb{Z} \\ \dfrac{d}{d\mu}\left[(1-\mu^2)\dfrac{d}{d\mu}\Phi\right] + \left[n(n+1) - \dfrac{m^2}{1-\mu^2}\right]\theta = 0 \\ \qquad\qquad \text{avec } \mu = \sin\phi, \text{ pour tout } n \in \mathbb{N}, |m| \leqslant n \text{ (équation de Legendre)} \\ \rho^2\dfrac{d^2}{d\rho^2}V + 2\rho\dfrac{d}{d\rho}V + \left[\rho^2 - n(n+1)\right]V = 0 \\ \qquad\qquad \text{avec } \rho = kr \text{ et } V(\rho) = R(r) \text{ (équation de Bessel sphérique),} \end{cases} \tag{2.16}$$

Ces équations admettent les solutions fondamentales suivantes :

$$\begin{cases} T(\tau) = \alpha_\tau e^{i\omega\tau} \\ \Theta(\theta) = \alpha_\theta e^{im\theta} \\ \Phi(\phi) = \alpha_\phi P_{nm}(\sin\phi) \\ R(r) = \alpha_r j_n(kr) + \alpha'_r h_n(kr) \end{cases}, \tag{2.17}$$

où les $j_n(.)$ sont les fonctions de Bessel sphériques de première espèce et $h_n(.)$ les fonctions de Hankel sphériques de première espèce. Ces dernières divergeant en 0, on peut les exclure puisque l'on considère le cas où la sphère d'observation ne contient aucune source acoustique qui pourrait être à l'origine d'une pression non bornée. Les α sont des coefficients définissant la réalisation d'un champ particulier.
En utilisant $Y_{nm} \propto P_{nm}(\sin\phi)e^{im\theta}$ (2.2), les solutions de l'équation des ondes s'écrivent comme une combinaison linéaire des solutions fondamentales, menant à une représentation de Fourier-Bessel :

$$p_\omega(\mathbf{r},\tau) = S_\omega \sum_{n=0}^{\infty}\sum_{m=-n}^{n} c_{nm} j_n(kr) Y_{nm}(\theta,\phi) e^{i\omega\tau}, \tag{2.18}$$

où $c_{nm} = 4\pi i^n Y^*_{nm}(\theta_i,\phi_i)$ est un coefficient déterminé dans le cas des ondes planes harmoniques [4, p. 226]. L'expression du champ de pression est alors la suivante :

$$p_\omega(\mathbf{r},\tau) = 4\pi S_\omega e^{i\omega\tau} \sum_{n=0}^{\infty} i^n j_n(kr) \sum_{m=-n}^{n} Y_{nm}(\theta,\phi) Y^*_{nm}(\theta_i,\phi_i), \tag{2.19}$$

où l'on rappelle que (θ_i, ϕ_i) est la direction d'arrivée de l'onde.
Afin d'identifier les coefficients ambisoniques dans le formalisme réel que nous utiliserons par la suite, il faut exprimer $p_\omega(\mathbf{r}, \tau)$ en fonction des harmoniques sphériques réelles $\tilde{Y}^\sigma_{nm}(\theta, \phi)$. En utilisant (2.6) et en regroupant les termes pour $m > 0$ et $m < 0$, on obtient :

$$\sum_{m=-n}^{n} Y_{nm}(\theta,\phi) Y^*_{nm}(\theta_i,\phi_i) = \frac{1}{4\pi}\left[\tilde{Y}^1_{n0}(\theta,\phi)\tilde{Y}^1_{n0}(\theta_i,\phi_i) + \sum_{m=1}^{n}\sum_{\sigma=\pm1}\tilde{Y}^\sigma_{nm}(\theta,\phi)\tilde{Y}^\sigma_{nm}(\theta_i,\phi_i)\right], \tag{2.20}$$

et donc, en injectant (2.20) dans (2.19) :

$$p_\omega(\mathbf{r},\tau) = S_\omega e^{i\omega\tau}\sum_{n=0}^{\infty} i^n j_n(kr)\left[\tilde{Y}^1_{n0}(\theta,\phi)\tilde{Y}^1_{n0}(\theta_i,\phi_i) + \sum_{m=1}^{n}\sum_{\sigma=\pm1}\tilde{Y}^\sigma_{nm}(\theta,\phi)\tilde{Y}^\sigma_{nm}(\theta_i,\phi_i)\right]. \tag{2.21}$$

Les coefficients ambisoniques B^σ_{nm} sont définis par rapport à la décomposition spatiale du champ de pression sur les harmoniques sphériques qui ne dépendent que de la direction d'observation (θ, ϕ). Ils ne prennent pas en compte les termes de propagation $i^n j_n(kr)$, et valent donc [6, Ann. A] :

$$B^\sigma_{nm} = S_\omega e^{i\omega\tau}\tilde{Y}^\sigma_{nm}(\theta_i,\phi_i). \tag{2.22}$$

Pour une onde quelconque (non nécessairement harmonique) portant le signal $s(\tau)$ provenant de (θ_i, ϕ_i), les coefficients ambisoniques s'expriment de même

$$B^\sigma_{nm}(\tau) = s(\tau)\tilde{Y}^\sigma_{nm}(\theta_i,\phi_i). \tag{2.23}$$

Les termes $i^n j_n(kr)$ étant connus pour tout r, un champ acoustique peut théoriquement être reconstitué en tout point de l'espace exempt de source sonore grâce à la connaissance de ses coefficients ambisoniques.

2.2. Ambisonie d'ordre 1

2.2.1. Troncature de la décomposition de Fourier sphérique

En appliquant la décomposition de Fourier sphérique (2.10) au champ de pression $p(k, r, \theta, \phi)$, il est donc possible de représenter parfaitement le champ acoustique à la surface d'une sphère de rayon r par ses coefficients de Fourier sphériques $p_{nm}(k, r)$. Cependant, il est impossible en pratique d'obtenir et de manipuler cette représentation infinie. On effectue donc une troncature à un degré donné N de la décomposition de Fourier sphérique (2.10). De façon abusive, on parle alors d'ambisonie d'« ordre » N :

$$p(k,r,\theta,\phi) \approx \sum_{n=0}^{N}\left[p^1_{n0}(k,r)\tilde{Y}^1_{n0}(\theta,\phi) + \sum_{m=1}^{n}\sum_{\sigma=\pm1} p^\sigma_{nm}(k,r)\tilde{Y}^\sigma_{nm}(\theta,\phi)\right]. \tag{2.24}$$

La série de Fourier sphérique d'une fonction de $\mathcal{L}_2$ converge vers ladite fonction. La troncature à l'ordre N en permet donc une approximation qui minimise l'erreur quadratique moyenne. Celle-ci est considérée négligeable lorsque $kr < N$ [7, p. 42]. Plus l'ordre est élevé, plus le champ acoustique est correctement représenté pour une zone de l'espace et jusqu'à une fréquence importante.

2.2.2. Interprétation des coefficients ambisoniques d'ordres 0 et 1

L'utilisation de la décomposition de Fourier sphérique pour représenter un champ acoustique a été introduite par Gerzon [9]. Contrairement aux développements postérieurs qui utilisent des composantes d'ordre élevé (HOA, *high-order Ambisonics*), celle-ci s'appuie sur la troncature à l'ordre 1 de la représentation. Appelée initialement « format B », cette représentation est aujourd'hui plus communément nommée FOA (*first-order Ambisonics*). On nomme classiquement W la composante d'ordre 0 et X, Y, et Z les composantes d'ordre 1. D'après leurs définitions (2.23) et les expressions des premières harmoniques sphériques réelles normalisées, pour une onde plane d'amplitude $s(\tau)$ venant de (θ_i, ϕ_i), elles valent :

$$\mathbf{x}(\tau) = \begin{bmatrix} W(\tau) \\ X(\tau) \\ Y(\tau) \\ Z(\tau) \end{bmatrix} = \begin{bmatrix} s(\tau)\tilde{Y}_{00}^{1}(\theta_i, \phi_i) \\ s(\tau)\tilde{Y}_{11}^{1}(\theta_i, \phi_i) \\ s(\tau)\tilde{Y}_{11}^{-1}(\theta_i, \phi_i) \\ s(\tau)\tilde{Y}_{10}^{1}(\theta_i, \phi_i) \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{3}\cos\theta_i \cos\phi_i \\ \sqrt{3}\sin\theta_i \cos\phi_i \\ \sqrt{3}\sin\phi_i \end{bmatrix} s(\tau). \tag{2.25}$$

En $r = 0$, les fonction de Bessel $j_n(kr)$ valent 0 pour $n \geqslant 1$. La fonction de Bessel de degré $n = 0$ vaut 1 en 0, tout comme les harmoniques sphériques réelles normalisées de degré 0 ($\tilde{Y}_{00}^{\sigma}$=1). D'après (2.21), on retrouve que $W = s(\tau)$ est la pression en $r = 0$. Ces quatre canaux FOA peuvent donc être interprétés comme la captation du champ acoustique en un point de l'espace par quatre microphones coïncidents : un microphone omnidirectionnel W et trois microphones bidirectionnels polarisés X, Y et Z. Les directivités des quatre canaux sont représentées sur la Figure 2.2.

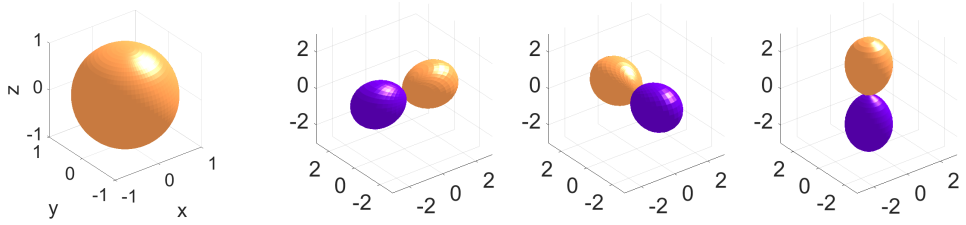


FIGURE 2.2. – Directivité $||\tilde{Y}_{nm}^{\sigma}(\theta, \phi)||^2$ des harmoniques sphériques correspondant aux canaux ambisoniques W, X, Y et Z. Les zones orange (claires) indiquent une grandeur positive et les violettes (plus foncées) une grandeur négative.

Cette formule est également valable en remplaçant les versions temporelles des signaux $s(\tau)$ et $\mathbf{x}(\tau)$ par leurs transformées de Fourier à court terme (TFCTs) $s(t, f)$ et $\mathbf{x}(t, f)$, où t représente la trame temporelle et f la bande de fréquence (voir Partie 3.2.1).

2.2.3. Encodage ambisonique d'une prise de son réelle

Il est impossible de mesurer les canaux HOA directement avec un microphone par canal, d'une part puisqu'il n'existe pas de microphones avec des directivités parfaitement égales aux harmoniques sphériques, et d'autre part car il serait de toute façon impossible de les placer de façon strictement coïncidente en $r = 0$.
Pour déterminer les coefficients ambisoniques, on utilise généralement une antenne de K microphones qui peut être considérée comme l'échantillonnage spatial d'une sphère. Un exemple typique en est l'Eigenmike [10] constitué de 32 capsules (voir Figure 2.3).

FIGURE 2.3. – Antenne de microphones Eigenmike, par Mh Acoustics [10]. Elle comporte 32 capsules réparties quasi-uniformément sur une sphère de 8,4 cm de diamètre, permettant un encodage HOA jusqu'à l'ordre 4.

Le vecteur $\mathbf{p}_J$ contenant la pression $p(r_{\text{mic}}, \theta_j, \phi_j)$ au niveau de chaque capsule j située en $(r_{\text{mic}}, \theta_j, \phi_j)$, appelé format A, s'écrit à l'ordre N [11, p. 27] :

$$\mathbf{p}_J = \mathbf{Y}\text{diag}(\mathbf{b}(kr_{\text{mic}}))\mathbf{x}_N, \tag{2.26}$$

où $\mathbf{x}_N$ est le vecteur colonne contenant les $(N+1)^2$ coefficients ambisoniques B_{nm}^{σ} jusqu'à l'ordre N. Dans le cas théorique d'une sphère acoustiquement transparente, $\mathbf{b}(kr_{\text{mic}})$ est le vecteur contenant les coefficients radiaux $b_n(kr_{\text{mic}}) = i^n j_n(kr_{\text{mic}})$ qui apparaissent dans l'expression (2.19) pour tous (n, m) correspondant à l'ordre N. Pour une sphère rigide comme l'Eigenmike, un terme correctif est en réalité introduit dans l'expression de $\mathbf{b}$, correspondant à la diffraction sur la sphère rigide sur laquelle sont placées les capsules et aux directivités de celles-ci [4, p. 228]. $\mathbf{Y}$ est la matrice de taille $J \times (N+1)^2$:

$$\mathbf{Y} = \begin{bmatrix} \tilde{Y}_{00}^{1}(\theta_1, \phi_1) & \cdots & \tilde{Y}_{NN}^{-1}(\theta_1, \phi_1) \\ \vdots & \ddots & \vdots \\ \tilde{Y}_{00}^{1}(\theta_J, \phi_J) & \cdots & \tilde{Y}_{NN}^{-1}(\theta_J, \phi_J) \end{bmatrix}. \tag{2.27}$$

L'encodage ambisonique à l'ordre N d'une prise de son avec une telle antenne sphérique se déduit de l'inversion de l'équation (2.26) :

$$\mathbf{x}_N = \text{diag}(\mathbf{b}(kr_{\text{mic}}))^{-1}\mathbf{Y}^{\dagger}\mathbf{p}_J. \tag{2.28}$$

où $.^{\dagger}$ est l'opération de pseudo-inversion.
Le calcul d'une représentation ambisonique à l'ordre N demande donc au minimum autant de microphones que de coefficients ambisoniques, à savoir $(N+1)^2$. Même dans ce cas, des erreurs d'encodage sont présentes en raison de plusieurs facteurs :

— Mathématiquement, la troncature de la décomposition sur les harmoniques sphériques introduit une imprécision en haute fréquence. À l'ordre 1 et pour une sphère

de rayon $r_{\text{mic}} = 4,2$ cm telle que l'Eigenmike, la fréquence maximale correspondant au critère $kr_{\text{mic}} < N$ est $f = 1,3$ kHz. Au-delà de ce seuil, l'encodage ambisonique induit certes une déformation du signal, mais elle est limitée comme le montrent les directivités réelles de l'Eigenmike mesurées par Baqué [12] (Figure 2.4).

— D'autre part, l'échantillonage de la sphère par des capsules microphoniques entraîne un recouvrement spatial dans les hautes fréquences. Mathématiquement, cela est dû à l'imperfection de l'inversion de la matrice $\mathbf{Y}$ [13]. Pour l'Eigenmike, qui a une distance inter-capsule $d_{\text{mics}} \approx 1,6$ cm, le critère communément adopté $\lambda > 2d_{\text{mics}}$, où $\lambda = \frac{c}{f}$ est la longueur d'onde, correspond à une fréquence $f < 10$ kHz.

— Enfin, à partir de l'ordre 1, les fonctions radiales $b_n(kr_{\text{mic}})$ sont très faibles en basse fréquence. Leur inversion pour le recouvrement des coefficients ambisoniques (2.28) est donc particulièrement sensible au bruit. Cela s'interprète physiquement par le fait qu'il est impossible de mesurer précisément les gradients de pression d'une onde en basse fréquence avec des microphones trop proches les uns des autres. Une étude de ce phénomène en fonction de différentes méthodes de régularisation est disponible dans l'ouvrage de Pulkki [11, p. 32-37]. Pour l'Eigenmike et l'encodage associé, on constate sur la partie de droite de la Figure 2.4 qu'à partir de 300 Hz, les directivités sont très proches des directivités optimales.

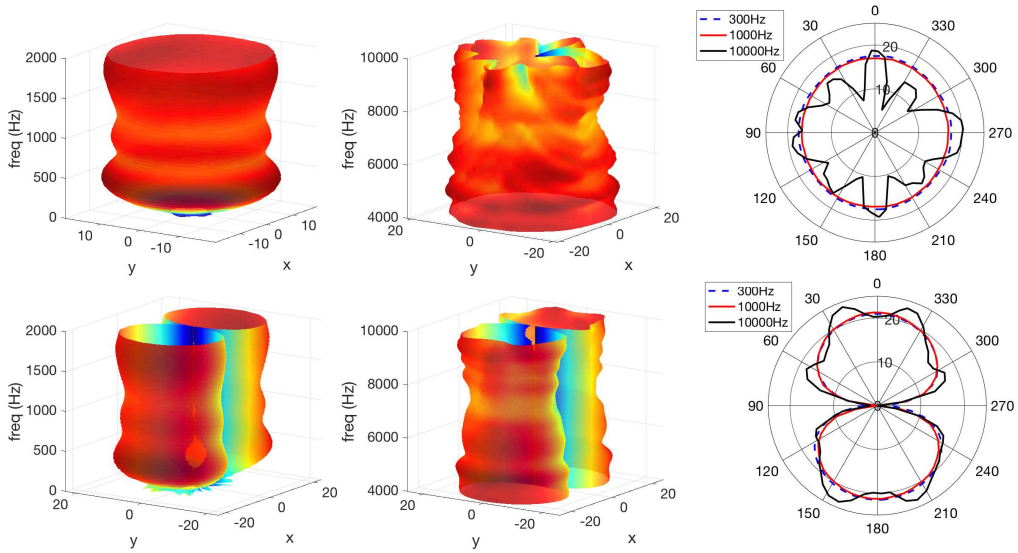


FIGURE 2.4. – Directivités réelle (dB) en fonction de la fréquence des composantes ambisoniques sur le plan horizontal (x, y) pour le microphone Eigenmike, d'après Baqué [12], pour les canaux W (haut) et X (bas). Gauche : basses fréquences. Centre : hautes fréquences. Droite : Coupes transversales.

Malgré ces limitations, nous considèrerons comme Baqué [12] que l'encodage FOA est valable jusque 8 kHz. Nous verrons par la suite que c'est suffisant pour nos applications de localisation et de rehaussement de la parole. Nous nous limitons donc à cet ordre par la suite.

2.3. Intérêt du format ambisonique

2.3.1. Représentation isotropique du champ acoustique

Le format ambisonique s'appuie sur une représentation mathématique intrinsèque au champ acoustique. En supposant l'encodage et le décodage parfaits, il est indépendant des systèmes de capture et de restitution. En cela, la représentation ambisonique peut être considérée comme un format pivot pour les contenus spatialisés. C'est un avantage majeur compte tenu du développement actuel des technologies audio 3D. Pour n'importe quelle antenne sphérique de microphones, il suffit de connaître la matrice d'encodage pour pouvoir conserver l'enregistrement au format ambisonique. Au moment de la restitution, il suffit cette fois de connaître la matrice de décodage correspondant au système de restitution, que ce soit le binaural, le 5.1, le Dolby ATMOS, ou tout autre format audio 3D. Cela permet une plus grande fluidité dans la production de contenus audio spatialisés. La première antenne dédiée à l'ambisonie fut le Soundfield conçu par Gerzon. Il en existe aujourd'hui bien d'autres, par exemple l'Ambeo de Sennheiser et les microphones Brahma pour l'ordre 1, ou l'Eigenmike de Mh Acoustics jusqu'à l'ordre 4.
Le format ambisonique est une représentation isotropique, c'est-à-dire qu'elle ne privilégie aucune direction de l'espace. Elle permet également des manipulations spatiales simples du champ acoustique, comme par exemple des rotations. La version transformée $\hat{\mathbf{x}}$ du champ $\mathbf{x} = [WXYZ]^T$ capté à l'ordre 1 par une rotation $\mathbf{R}$ s'effectue grâce à une simple multiplication matricielle [8, p. 53]

$$\hat{\mathbf{x}}(t,f) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & & & \\ 0 & & \mathbf{R} & \\ 0 & & & \end{bmatrix} \mathbf{x}(t,f), \tag{2.29}$$

où $\mathbf{R}$ peut toujours s'exprimer comme produit des matrices de rotation élémentaire autour des axes x, y, et z selon les angles de Cardan :

$$\begin{gathered} \mathbf{R}_x(\alpha_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_x & -\sin\alpha_x \\ 0 & \sin\alpha_x & \cos\alpha_x \end{bmatrix}, \mathbf{R}_y(\alpha_y) = \begin{bmatrix} \cos\alpha_y & 0 & -\sin\alpha_y \\ 0 & 1 & 0 \\ \sin\alpha_y & 0 & \cos\alpha_y \end{bmatrix}, \\ \mathbf{R}_z(\alpha_z) = \begin{bmatrix} \cos\alpha_z & -\sin\alpha_z & 0 \\ \sin\alpha_z & \cos\alpha_z & 0 \\ 0 & 0 & 1 \end{bmatrix}. \end{gathered} \tag{2.30}$$

C'est particulièrement utile dans le cas de la captation d'une scène sonore effectuée par un robot capable de tourner la tête, auquel cas le champ ambisonique peut facilement être déduit en temps réel indépendamment des mouvements du robot. C'est également utile lors de la restitution binaurale d'une scène sonore, où le suivi des mouvements de la tête de l'auditeur permet plus de réalisme.

2.3.2. Formation de voie

L'ambisonie permet également une formulation simple des filtres pleine bande de formation de voie, qui visent à rehausser le son venant d'une direction donnée (θ_i, ϕ_i). Un filtre $\mathbf{w}$ est appliqué à l'enregistrement ambisonique $\mathbf{x}$ de la façon suivante :

$$y(t, f) = \mathbf{w}^H \mathbf{x}(t, f), \tag{2.31}$$

où $y(t, f)$ est le signal monocanal rehaussé et $(.)^H$ est l'opérateur de transconjugaison. La formation de voie dite « non contrainte » rehaussant la direction (θ_i, ϕ_i) s'effectue simplement en utilisant le vecteur d'encodage d'une onde plane $\mathbf{d}_{\theta_i,\phi_i}$ entre la source et le point d'observation, appelé vecteur directionnel :

$$\mathbf{w}_{\text{dir}} = \mathbf{d}_{\theta_i,\phi_i} = \begin{bmatrix} 1 \\ \sqrt{3}\cos\theta_i \cos\phi_i \\ \sqrt{3}\sin\theta_i \cos\phi_i \\ \sqrt{3}\sin\phi_i \end{bmatrix}. \tag{2.32}$$

De même, si plusieurs sources sont présentes (au maximum autant que le nombre de canaux dans le mélange ambisonique, ici 4), le filtre rehaussant la direction (θ_i, ϕ_i) et annulant les autres directions s'obtient en inversant la matrice de mélange constituée des vecteurs d'encodage des directions d'arrivée [12]

$$\mathbf{w}_{\text{dir}}^H = \mathbf{u_i}^T \left[\mathbf{d}_{\theta_0,\phi_0}\, \mathbf{d}_{\theta_1,\phi_1} \ldots \mathbf{d}_{\theta_I,\phi_I}\right]^\dagger, \tag{2.33}$$

où $\mathbf{u_i}$ est le vecteur colonne unitaire de dimension égale au nombre de sources I, avec un 1 en position i. Ce filtre pleine bande dépend uniquement des directions des sources et non de la fréquence. S'il présente l'avantage de supprimer les directions d'arrivées souhaitées, la directivité de ce filtre dit « contraint » possède d'importants lobes dans des directions annexes (Figure 2.5).

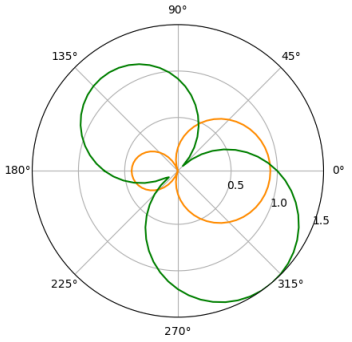


FIGURE 2.5. – Directivités dans le plan horizontal des filtres de formation de voie. En orange : rehaussant la direction (0°, 0°). En vert : rehaussant la direction (0°, 0°) et annulant la direction (0°, 45°)

2.3.3. Vecteur intensité acoustique

L'un des outils permettant de caractériser un champ sonore est le vecteur d'intensité acoustique active [14], qui représente le flux d'énergie en un point de l'espace :

$$\mathbf{I}_{\text{a}}(t, f) = \mathcal{R}\{p(t, f)\mathbf{v}^*(t, f)\}, \tag{2.34}$$

avec $\mathbf{v}(t,f)$ la vélocité particulaire et $p(t,f)$ le champ de pression généré par l'onde. Ici, le champ de pression est quelconque, et non pas généré par une onde plane harmonique. La vélocité particulaire s'exprime simplement dans le formalisme FOA [11, p. 90] :

$$\mathbf{v}(t,f) = -\frac{1}{\rho_0 c\sqrt{3}} \begin{bmatrix} X(t,f) \\ Y(t,f) \\ Z(t,f) \end{bmatrix}, \tag{2.35}$$

où ρ_0 est la densité de l'air. D'après (2.23), le canal ambisonique W d'ordre 0, aussi appelé canal monophonique, contient le champ de pression au point de mesure : $W(t,f) = p(t,f)$. À un facteur constant près que nous négligeons par la suite, le vecteur d'intensité acoustique active s'écrit donc à partir des canaux ambisoniques :

$$\mathbf{I}_\mathrm{a}(t,f) = -\begin{bmatrix} \mathcal{R}\{W(t,f)X^*(t,f)\} \\ \mathcal{R}\{W(t,f)Y^*(t,f)\} \\ \mathcal{R}\{W(t,f)Z^*(t,f)\} \end{bmatrix}. \tag{2.36}$$

On peut définir la partie réactive du vecteur d'intensité en utilisant cette fois la partie imaginaire $\mathbf{I}_\mathrm{r}(t,f) = \mathcal{I}\{p(t,f)\mathbf{v}^*(t,f)\}$ [14]. Elle représente les transferts locaux d'énergie dissipative. Son expression dans le formalisme FOA est :

$$\mathbf{I}_\mathrm{r}(t,f) = -\begin{bmatrix} \mathcal{I}\{W(t,f)X^*(t,f)\} \\ \mathcal{I}\{W(t,f)Y^*(t,f)\} \\ \mathcal{I}\{W(t,f)Z^*(t,f)\} \end{bmatrix}. \tag{2.37}$$

Ces deux pendants du vecteur intensité, la partie active et la partie réactive, donnent des indications sur les ondes qui génèrent le champ de pression observé. Lorsqu'une seule onde plane est présente, la partie active est opposée à la direction de propagation de l'onde. La partie réactive, elle, est nulle. Cela se vérifie dans le formalisme FOA : d'après (2.25) et (2.35), pour une onde plane, $p(t,f)\mathbf{v}^*(t,f)$ est nécessairement réel. Lorsque le champ est créé par une onde non plane, ou par superposition d'ondes, la partie réactive est plus fortement présente. C'est en particulier le cas lorsqu'une onde se déplace en milieu réverbérant.

2.4. Résumé

Cette partie présente le format ambisonique, qui est une description spatiale du champ de pression en un point. Elle est isotropique et se veut indépendante du système de capture dans la limite où celui-ci est adapté à l'encodage ambisonique (en particulier, il est nécessaire d'utiliser une antenne sphérique de microphones). Mathématiquement, c'est la projection du champ de pression sur la base des fonctions harmoniques sphériques. Nous avons en particulier présenté la troncature à l'ordre 1 de cette projection, nommée FOA. C'est une approximation de la représentation ambisonique exacte qui a l'avantage d'être légère à manipuler, avec seulement 4 canaux facilement interprétables en termes spatiaux. Elle permet également une expression simple des grandeurs dont nous nous

servirons par la suite : le vecteur d'intensité acoustique et le filtre de formation de voies pleine bande. Nous considérons dans cette thèse qu'une représentation de la scène sonore au format FOA est disponible, et c'est sur celle-ci que nous travaillerons, sans prendre en compte les potentielles erreurs dues à l'encodage.

3. État de l'art

Dans ce chapitre, nous présentons un état de l'art de reconnaissance vocale, de rehaussement de la parole et de localisation de sources, vus à travers le prisme de notre cas d'application : la facilitation de la commande vocale à distance en environnement domestique. Nous posons un cadre mathématique afin de définir le problème de rehaussement de la parole dans ces conditions. Nous mettons également en évidence les défis qui sont posés par ce cas d'application et qui ne sont pas encore résolus par les méthodes actuelles, en particulier en présence de plusieurs locuteurs. Cela nous permet de définir le cadre dans lequel se placent les travaux de cette thèse.

3.1. Reconnaissance vocale

3.1.1. De l'audio au texte

Les systèmes de RAP [15] sont conçus pour renvoyer le texte correspondant à un signal de parole. Plusieurs étapes sont fondamentales. Avant toute chose, il faut adopter une paramétrisation appropriée du signal audio par une séquence de vecteurs de descripteurs, par exemple les coefficients cepstraux en échelle Mel (MFCCs, *Mel frequency cepstral coefficients*). Ensuite, le modèle acoustique calcule la probabilité des senones correspondant au signal, c'est-à-dire des états acoustiques liés dépendant du contexte. Un algorithme de décodage détermine la séquence de mots la plus probable en prenant en compte les spécificités de la langue considérée représentée par un modèle de langage. Mathématiquement, ceci est décrit par la règle de Bayes [16] :

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{w}|\mathbf{x}) = \arg\max_{\mathbf{w}} P(\mathbf{x}|\mathbf{w})P(\mathbf{w}) \tag{3.1}$$

où $\mathbf{w}$ représente une séquence de mots, $\hat{\mathbf{w}}$ est la séquence de mots la plus probable retournée par le système, $P(.)$ est l'opérateur de densité de probabilité, et $\mathbf{x}$ l'observation acoustique. $P(\mathbf{w})$ est déterminé par le modèle de langage, tandis que le modèle acoustique représente $P(\mathbf{x}|\mathbf{w})$. Les différentes étapes de RAP sont représentées dans la Figure 3.1 et décrites ci-dessous.
Les modèles acoustiques utilisent traditionnellement des chaînes de Markov cachées (HMMs, *hidden Markov models*) où les états cachés représentent les senones. La densité de probabilité conditionnelle de vecteurs de descripteurs sachant l'état caché peut être modélisée par un mélange de gaussiennes (GMM, *Gaussian mixture model*). Aujourd'hui, les GMMs sont souvent remplacés par un réseau de neurones profond (DNN, *deep neural networks*) [17] qui estime la probabilité a posteriori des états cachés correspondant à la séquence de descripteurs acoustiques observés, et la paramétrisation par des MFCCs est remplacée par un banc de filtres de type TFCT.

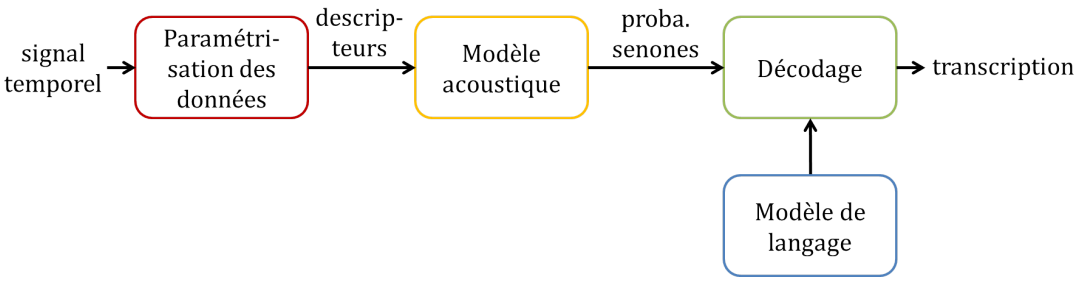


FIGURE 3.1. – Architecture classique d'un système de RAP.

Récemment, l'utilisation de réseaux de neurones récurrents (RNNs, *recurrent neural networks*) [18] a permis la mise au point de systèmes incluant à la fois le modèle acoustique et le modèle de langage dans un même réseau de neurones [19], permettant un apprentissage conjoint directement lié à l'objectif final, à savoir la transcription textuelle du signal de parole. La piste d'un apprentissage bout-en-bout incluant l'apprentissage des descripteurs directement à partir de signaux temporels commence également à être explorée [20]. Les premiers résultats atteignent des performances proches des meilleures paramétrisations fréquentielles, tout en permettant une plus grande flexibilité vis à vis de la tâche et du scénario considérés.

3.1.2. Robustesse en champ lointain

Défis du monde réel Les systèmes décrits ci-dessus fonctionnent remarquablement bien lorsque les signaux de parole à transcrire sont propres, c'est-à-dire avec une faible réverbération et sans bruit ou locuteurs concurrents. Ce n'est malheureusement pas le cas dans la situation domestique qui nous intéresse [21, 22], où un appareil de type enceinte connectée doit traiter les commandes données par une locutrice à une distance d'1 m ou plus. Le signal mesuré par les microphones inclut alors des réflexions précoces, de la réverbération et potentiellement du bruit ambiant et d'autres locuteurs [21].

Apprentissage robuste La première réponse à ces situations difficiles, appelée apprentissage multi-conditions, consiste à procéder à l'apprentissage du réseau de neurones de RAP dans toutes les situations susceptibles d'être rencontrées, que ce soit pour parer au bruit [23] ou à la réverbération [24]. Le système de RAP doit alors apprendre à reconnaître les mots malgré le bruit corrompant les signaux d'entrée. Il peut être rendu encore plus robuste par l'utilisation d'une paramétrisation moins sensible au bruit [25]. Ces techniques présentent l'avantage de ne pas nécessiter de données d'apprentissage propres. Elles ont prouvé leur efficacité et sont désormais systématiquement utilisées lors de l'apprentissage des systèmes de RAP, mais ne sont pas suffisantes. Comme l'ont montré les résultats des challenges CHiME [26], ainsi que les études suivantes [27], les données utilisées pour l'apprentissage dans de tels challenges sont généralement proches des données de test en termes de conditions acoustiques et de mode de capture, de sorte que les systèmes appris sur ces données peinent à se généraliser à d'autres cas réels. Para-

doxalement, une piste d'amélioration est l'utilisation de données simulées [28] : elles sont certes plus éloignées encore des conditions réelles, mais la possibilité de les synthétiser en grand nombre et avec une diversité importante facilite la généralisation. Un problème persiste cependant pour les situations multi-locuteurs, où il est nécessaire que le système possède une information supplémentaire lui permettant de se focaliser sur le locuteur qui donne l'ordre [29, 30].

Solutions bout-en-bout La tendance actuelle en RAP est de rassembler des chaînes de traitement entières dans un seul réseau de neurones très profond. Dans la continuité, plusieurs travaux récents intègrent une étape de rehaussement directement au système de RAP. Certains utilisent un réseau de neurones pour estimer un masque indiquant la présence de parole en chaque point temps-fréquence, afin d'en déduire un filtre multicanal à partir des formules traditionnelles de filtrage [31, 32], tandis que d'autres estiment directement les coefficients du filtre à appliquer [33]. Dans les deux cas, les étapes de traitement déterministes sont différentiables et l'apprentissage est fait conjointement pour tous les composants. Cela permet d'optimiser toutes les étapes en fonction de l'objectif final de transcription. De surcroît, il est ici aussi possible de procéder à l'apprentissage avec des données bruitées.
L'apprentissage reste néanmoins délicat, nécessitant une quantité de données et une puissance de calcul considérables. D'autre part, les systèmes bout-en-bout sont dépendants du système de captation. L'apprentissage doit être fait avec la même antenne de microphones que celle qui sera disponible pour le test (bien que [32] propose une technique pour introduire une certaine flexibilité).

Solutions modulaires incluant un rehaussement explicite en amont L'utilisation de techniques de filtrage multicanal (voir partie 3.2.2) en amont de la RAP a permis une amélioration significative des performances en environnement défavorable. En particulier, les techniques les plus performantes pour les challenges CHiME et REVERB utilisent une étape préalable de filtrage spatial qui s'appuie sur l'estimation de masques temps-fréquence. Ceux-ci sont estimés par des réseaux de neurones [34, 24, 35].
Pour les challenges CHiME et REVERB, le modèle de RAP est généralement réappris sur les données d'apprentisage traitées par le système de rehaussement afin de gagner autant que possible en performance. Cependant, la dissociation entre module de rehaussement et module de RAP permet d'avoir un système de RAP qui fonctionne sur tous types de signaux, tandis que le système de rehaussement est spécifique au mode de capture. Étant donné le coût d'apprentissage d'un système de RAP comparé à celui d'un système de rehaussement, cela représente un avantage non-négligeable des solutions modulaires sur les solutions bout-en-bout.
D'autre part, les auteurs du système bout-en-bout pourtant prometteur Beamnet [31] ont constaté une irrégularité de performance : leur système est efficace sur le corpus CHiME-4, où les enregistrements sont effectués en champ relativement proche et parole interférente. En revanche, sur une base de données rassemblée par Google correspondant à des cas réels de commande vocale en champ lointain, ce qui est notre cas d'application,

les performances chutent en-dessous de celles de systèmes ou le rehaussement et la RAP sont disjoints [2].

3.1.3. Résumé et positionnement

Cette partie présente un état de l'art de la RAP et l'intégration relativement récente des réseaux de neurones dans ce domaine. Si le problème est aujourd'hui résolu en champ proche avec un seul locuteur, les situations où le locuteur est en champ lointain induisent des difficultés encore non résolues. Plusieurs stratégies ont été proposées pour rendre les systèmes robustes aux interférences et à la réverbération qui en découlent : l'apprentissage robuste, l'apprentissage bout-en-bout ou encore le traitement modulaire du signal. Cette dernière solution est aujourd'hui la plus efficace ; nous nous plaçons donc dans ce contexte, en nous intéressant plus précisément aux techniques de rehaussement de la parole pour le pré-traitement des signaux avant la RAP. Le système de RAP que nous utilisons pour mesurer l'efficacité du rehaussement s'appuie sur des réseaux de neurones. Il est utilisé ici comme une boîte noire. Dans la partie suivante, nous présentons un état de l'art du rehaussement de la parole.

3.2. Rehaussement de la parole

3.2.1. Définition du problème

Séparation de sources et rehaussement de la parole Dans une situation de *cocktail party*, les voix de plusieurs locuteurs se mêlent dans un environnement bruyant et réverbérant (voir Figure 3.2). Face à ce problème, plusieurs objectifs peuvent être poursuivis : si l'on cherche à obtenir les signaux audio séparés de tous les locuteurs, on parle de séparation de sources. Si le but est de « nettoyer » un ou plusieurs signaux de parole, ce qui peut inclure le débruitage et la déréverbération en plus de la séparation de sources, on parle de rehaussement de la parole [36]. Les signaux séparés ou rehaussés peuvent être destinés à l'écoute ou à un post-traitement tel que la reconnaissance vocale.
Dans le cas d'un signal destiné à l'écoute, une simple erreur quadratique entre le signal visé et le signal estimé est rarement pertinente. Plusieurs indicateurs estimant le degré de séparation et de distorsion du signal ont été proposés [37]. Le rapport signal sur interférences mesure le rapport de puissance entre le signal cible et les interférences ou le bruit provenant d'autres sources sonores. Le rapport signal sur artefacts évalue la quantité de « bruits musicaux » présents dans le signal estimé.
Dans notre cas, on considère un problème de rehaussement avec un seul locuteur cible. L'objectif est ici la reconnaissance vocale. La qualité perceptive du signal de parole estimé importe finalement peu ; la métrique finale s'appuie sur la validité de la transcription de celui-ci, mesurée par le taux d'erreur sur les mots, ou WER, défini dans la partie 5.2.5.

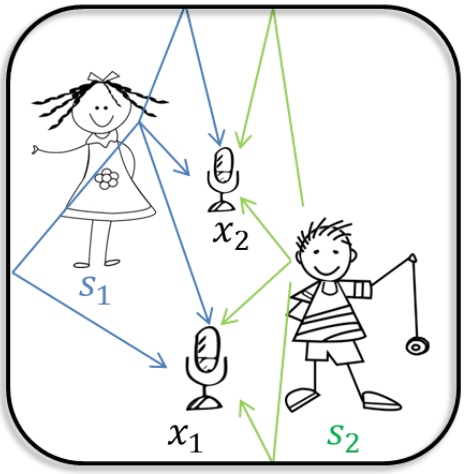


FIGURE 3.2. –
Exemple de *cocktail party* : un goûter d'anniversaire avec deux enfants enregistrés par deux microphones.

Formulation temporelle D'un point de vue mathématique, la situation de *cocktail party* peut être décrite par l'équation

$$\mathbf{x}(n) = \sum_{i=0}^{I-1} \mathbf{c}_i(n) + \mathbf{n}_{\text{diff}}(n) \tag{3.2}$$

où n est un échantillon temporel donné, $\mathbf{x}(n)$ est le signal multicanal capté comprenant J canaux (dans le cas du FOA, $J = 4$), I est le nombre de sources présentes dans le mélange, les $\mathbf{c}_i(n)$ sont les contributions multicanales de chaque source au mélange appelées « images spatiales » des sources et $\mathbf{n}_{\text{diff}}(n)$ est un bruit additionnel provenant de sources spatialement diffuses ou de la mesure. Les $\mathbf{c}_i(n)$ résultent de la propagation acoustique des signaux émis par les sources jusqu'au point de mesure modélisée par une convolution :

$$\mathbf{c}_i(n) = \sum_{m=0}^{\infty} \mathbf{h}_i(m) * t_i(n-m). \tag{3.3}$$

$\mathbf{h}_i(m)$ est le vecteur des réponses impulsionnelles caractérisant le chemin acoustique entre la source i et l'antenne de microphones et $t_i(n)$ est le signal émis par la source i. On considère les sources immobiles à l'échelle du temps d'analyse, $\mathbf{h}_i(m)$ est donc constant au cours du temps.
La séparation de sources vise à reconstituer soit les images spatiales $\mathbf{c}_i(n)$, soit les signaux émis $t_i(n)$ [38].
Pour notre application de rehaussement de la parole, si la métrique est claire (la validité de la transcription), le choix du signal intermédiaire que l'on vise à isoler n'est pas évident. On peut chercher à reconstituer $\mathbf{c}_i(n)$, ou bien sa version déréverbérée $t_i(n)$, ou encore un canal de $\mathbf{c}_i(n)$, puisqu'il suffit d'un signal monocanal pour procéder à la RAP. Cet objectif intermédiaire dépend notamment du système de RAP utilisé, plus ou moins robuste à la réverbération, aux distorsions, au bruit ambiant ou aux interférences

directionnelles. On peut poser le problème de la façon suivante :

$$\mathbf{x}(n) = \mathbf{s}(n) + \mathbf{n}(n) \tag{3.4}$$

où $\mathbf{s}(n)$ est un signal multicanal contenant le locuteur cible, tandis que $\mathbf{n}(n)$ contient tous les locuteurs interférents et le bruit diffus. Selon l'application, on cherche à estimer $\mathbf{s}(n)$ ou un signal monocanal sous-jacent, par exemple l'un de ses canaux.

Formulation temps-fréquence Dans le paragraphe précédent, le problème est formulé dans le domaine temporel. En pratique, il est plus facile de travailler dans le domaine temps-fréquence qui permet de faire apparaître la parcimonie des signaux et facilite leur séparation [39].
La représentation choisie est généralement la TFCT. L'étape d'analyse consiste à déterminer la représentation fréquentielle d'un signal à partir de sa version temporelle. Ceci est fait grâce à l'application d'une fenêtre d'analyse glissante $w_a(n)$ de longueur N, avec un pas H entre chaque application de la fenêtre. Une version fenêtrée $s_t(n)$ du signal $s(n)$ est extraite à chaque trame t :

$$s_t(n) = w_a(n)s(tH + n) \text{ avec } w_a(n) = 0 \text{ pour } n \notin \{0, ..., N-1\}. \tag{3.5}$$

On procède ensuite à la transformation de Fourier du signal fenêtré, ce qui fournit une représentation du signal dans chaque trame t et bande de fréquence f. On choisit en général une représentation fréquentielle sur un nombre de points égal à la longueur N de la fenêtre d'analyse. Les coefficients de la TFCT du signal $s(n)$ sont donnés par :

$$s(t, f) = \sum_{n=0}^{N-1} s_t(n)e^{-2i\pi nf/N}. \tag{3.6}$$

Pour un signal $s(n)$ réel, la transformée de Fourier étant symétrique hermitienne, on ne considère que la moitié des bandes de fréquence $f \in \{0, ..., N/2\}$.
L'opération inverse, qui permet de reconstituer un signal temporel à partir de ses coefficients de Fourier, s'appelle la synthèse. Dans un premier temps, il s'agit d'appliquer une transformation de Fourier inverse à $s(t, f)$:

$$s_t(n) = \frac{1}{N} \sum_{f=0}^{N-1} s(t, f)e^{2i\pi nf/N} \text{ avec } n \in \{0, ..., N-1\} \tag{3.7}$$

puis le signal temporel $s(n)$ peut être reconstitué par la méthode dite d'*overlap and add* :

$$s(n) = \sum_t s_t(n - tH)w_s(n - tH). \tag{3.8}$$

Afin que la reconstruction soit exacte, le pas H et les fenêtres d'analyse $w_a(n)$ et de synthèse $w_s(n)$ doivent vérifier le critère suivant :

$$\sum_t w_a(n - tH)w_s(n - tH) = 1 \text{ pour tout } n \in \{0, ..., N-1\}. \tag{3.9}$$

Bien qu'il soit courant de choisir une fenêtre d'analyse qui permette de satisfaire ce critère sans recours à une fenêtre de synthèse, par exemple une fenêtre de Hamming, cela pose problème lorsque des manipulations sont effectuées dans le domaine temps-fréquence : des discontinuités peuvent apparaître dans la reconstruction temporelle du signal, notamment à la frontière entre les trames. Pour limiter ces discontinuités, il est préférable d'utiliser une fenêtre de synthèse [40]. Dans notre cas, nous utiliserons la fonction sinus pour les deux fenêtres, avec un pas $H = N/2$.
Les équations (3.2) et (3.3) peuvent être réécrites dans le domaine temps-fréquence :

$$\mathbf{x}(t,f) = \sum_{i=0}^{I-1} \mathbf{c}_i(t,f) + \mathbf{n}_{\text{diff}}(t,f). \tag{3.10}$$

Sous l'hypothèse de bande étroite, c'est-à-dire si la fenêtre d'analyse est suffisamment grande par rapport au temps caractéristique de réverbération, la convolution dans (3.3) peut être approchée par une multiplication dans le domaine fréquentiel :

$$\mathbf{c}_i(t,f) = \mathbf{h}_i(f) t_i(t,f), \tag{3.11}$$

où $\mathbf{h}_i(f)$ est la transformée de Fourier discrète de taille N de $\mathbf{h}_i(n)$. En pratique, cette hypothèse est rarement vérifiée [41, 42]. Nous baserons notre travail sur l'équation issue de (3.4)

$$\mathbf{x}(t,f) = \mathbf{s}(t,f) + \mathbf{n}(t,f). \tag{3.12}$$

3.2.2. Méthodes historiques de rehaussement pour une seule source

En rehaussement, lorsqu'une seule source doit être mise en valeur, on utilise souvent des techniques de filtrage. Puisque nous traitons des signaux ambisoniques, nous ne présenterons que les techniques de filtrage multicanal. En s'appuyant sur la dimension spatiale du mélange capté, elles induisent beaucoup moins de distorsions qu'un filtre monocanal. Il a été prouvé que ceci est largement avantageux pour la RAP [34]. On cherche donc un filtre $\mathbf{w}$, c'est-à-dire un vecteur de taille J tel que

$$y(t,f) = \mathbf{w}(t,f)^H \mathbf{x}(t,f), \tag{3.13}$$

où $y(t,f)$ est une estimation monocanale du locuteur cible. Estimer directement le signal source $t_i(t,f)$ n'est pas faisable en pratique [41]. On cherche donc souvent à retrouver un des canaux de son image spatiale $\mathbf{s}(t,f) = \mathbf{c}_i(t,f)$. Un module de déréverbération peut être appliqué a posteriori pour approcher $t_i(t,f)$. $\mathbf{w}(t,f)$ est calculé selon un critère à définir, par exemple d'erreur quadratique entre le signal estimé $y(t,f)$ et le signal cible.

Formation de voies directionnelle Cette partie présente un filtre dérivé exclusivement selon des critères spatiaux. On considère que la direction d'arrivée (θ_0, ϕ_0) de la source à rehausser est connue. Le filtre ne dépend pas du signal émis lui-même ; il se contente de

rehausser le chemin direct de l'onde, considérée plane et sous l'hypothèse de champ lointain. La connaissance de la géométrie de l'antenne de microphones permet la connaissance du vecteur directionnel $\mathbf{d}_{\theta_0,\phi_0}(f)$. Le filtre de formation de voies associé est alors

$$\mathbf{w}(f) = \mathbf{d}_{\theta_0,\phi_0}(f). \tag{3.14}$$

Par exemple, pour une antenne de microphones linéaire uniforme avec une distance δ entre les microphones,

$$\mathbf{d}_{\theta_0,\phi_0}(f) = [e^{-2i\pi\frac{\delta f f_s}{cN}\cos(\theta_0)}...e^{-2i\pi\frac{J\delta f f_s}{cN}\cos(\theta_0)}]^T \tag{3.15}$$

avec f_s la fréquence d'échantillonage en Hz. Le filtre $\mathbf{w}(f)$ correspondant est alors appelé *delay-and-sum* [43].
Pour une antenne ambisonique, le vecteur directionnel ne dépend pas de la fréquence. Le filtre résultant est alors pleine bande : il possède la même directivité quelque soit la fréquence. Son expression est donnée dans l'équation (2.32) de la partie 2.3.2, ainsi qu'une extension permettant d'annuler le signal provenant d'autres directions.

Filtre de Wiener multicanal La formation de voies s'appuyant uniquement sur la direction d'arrivée ne permet pas de prendre en compte la réverbération, contrairement à certains filtres dépendant de la fréquence et des statistiques des signaux [44]. Ceux-ci sont plus complexes et induisent plus de distorsion que les filtres présentés précédemment, mais ils ont un pouvoir séparateur bien plus important.
Parmi eux, la famille des filtres de Wiener multicanaux (MWF, *multichannel Wiener filter*) cherche une estimation optimale du signal cible au sens de l'erreur quadratique moyenne. En reprenant l'expression du mélange (3.12), le MWF vise à estimer le premier canal $s_1(t,f)$ de $\mathbf{s}(t,f)$ en minimisant l'erreur quadratique moyenne entre $s_1(t,f)$ et le signal filtré :

$$\mathbf{w}_{\text{MWF}}(t,f) = \min_{\mathbf{w}(t,f)} \mathbb{E}\{|\mathbf{w}^H(t,f)\mathbf{x}(t,f) - s_1(t,f)|^2\} \tag{3.16}$$

où $\mathbb{E}\{.\}$ est l'opérateur d'espérance. En supposant tous les signaux centrés, l'optimisation de ce critère mène à l'expression du MWF :

$$\mathbf{w}_{\text{MWF}}(t,f) = [\mathbf{R}_{\mathbf{ss}}(t,f) + \mathbf{R}_{\mathbf{nn}}(t,f)]^{-1}\mathbf{R}_{\mathbf{ss}}(t,f)\mathbf{u}_0 \tag{3.17}$$

où $\mathbf{u}_0$ est le vecteur $[1\,0\,...\,0]^T$ et $\mathbf{R}_{\mathbf{ss}}(t,f)$ et $\mathbf{R}_{\mathbf{nn}}(t,f)$ sont les matrices de covariance de $\mathbf{s}(t,f)$ et $\mathbf{n}(t,f)$ (qui sont hermitiennes positives). La matrice de covariance d'un signal $\mathbf{u}(t,f)$ est définie par

$$\mathbf{R}_{\mathbf{uu}}(t,f) = \mathbb{E}\{\mathbf{u}(t,f)\mathbf{u}(t,f)^H\}. \tag{3.18}$$

Il est possible d'introduire un coefficient de pondération μ pour régler le compromis entre l'atténuation des sources interférentes et du bruit et la distorsion du signal de parole estimé. Le filtre pondéré est appelé SDW-MWF (*speech distortion weighted MWF*) [45] :

$$\mathbf{w}_{\text{SDW-MWF}}(t,f) = [\mathbf{R}_{\mathbf{ss}}(t,f) + \mu\mathbf{R}_{\mathbf{nn}}(t,f)]^{-1}\mathbf{R}_{\mathbf{ss}}(t,f)\mathbf{u}_0 \tag{3.19}$$

Pour $\mu \to \infty$, l'atténuation des interférences et du bruit est maximale, mais les distorsions sont importantes. Pour $\mu \to 0$, en faisant l'hypothèse que $\mathbf{R_{ss}}(t,f)$ est de rang 1, on peut reformuler le filtre et montrer qu'il tend vers le filtre minimisant les distorsions appelé MVDR (*minimum variance distortionless response*) [46, 47].

Maximum SNR Le filtre maximisant le rapport signal à bruit (SNR, *signal-to-noise ratio*) [48, 49] est défini par

$$\mathbf{w}_{\text{max-SNR}}(t,f) = \underset{\mathbf{w}(t,f)}{\arg\max} \frac{\mathbf{w}^H(t,f)\mathbf{R_{ss}}(t,f)\mathbf{w}(t,f)}{\mathbf{w}^H(t,f)\mathbf{R_{nn}}(t,f)\mathbf{w}(t,f)}. \tag{3.20}$$

dont la solution est donnée par le vecteur propre généralisé principal de $\mathbf{R_{ss}}(t,f)$ et $\mathbf{R_{nn}}(t,f)$, d'où le nom de GEV (*generalized eigenvalue*) qui lui est également donné [49] :

$$\mathbf{w}_{\text{max-SNR}}(t,f) = \mathcal{P}[\mathbf{R_{nn}}^{-1}(t,f)\mathbf{R_{ss}}(t,f)]. \tag{3.21}$$

MWF avec une approximation de rang 1 de la matrice de covariance Afin de rendre le MWF plus robuste aux erreurs d'estimation de $\mathbf{R_{ss}}(t,f)$, on peut la remplacer dans (3.17) par une approximation de rang 1 :

$$\mathbf{R_{ss-r1}}(t,f) = \sigma(t,f)\mathbf{a}(t,f)\mathbf{a}(t,f)^H. \tag{3.22}$$

En particulier pour le filtre s'appuyant sur la décomposition généralisée en valeurs propres (GEVD, *generalized eigenvalue decomposition*) [50], $\mathbf{a}(t,f)$ est obtenu comme le vecteur propre généralisé principal de $\mathbf{R_{xx}}$ et $\mathbf{R_{nn}}$ (définies par (3.18)) :

$$\mathbf{a}(t,f) = \mathcal{P}[\mathbf{R_{nn}}(t,f)^{-1}\mathbf{R_{xx}}(t,f)]. \tag{3.23}$$

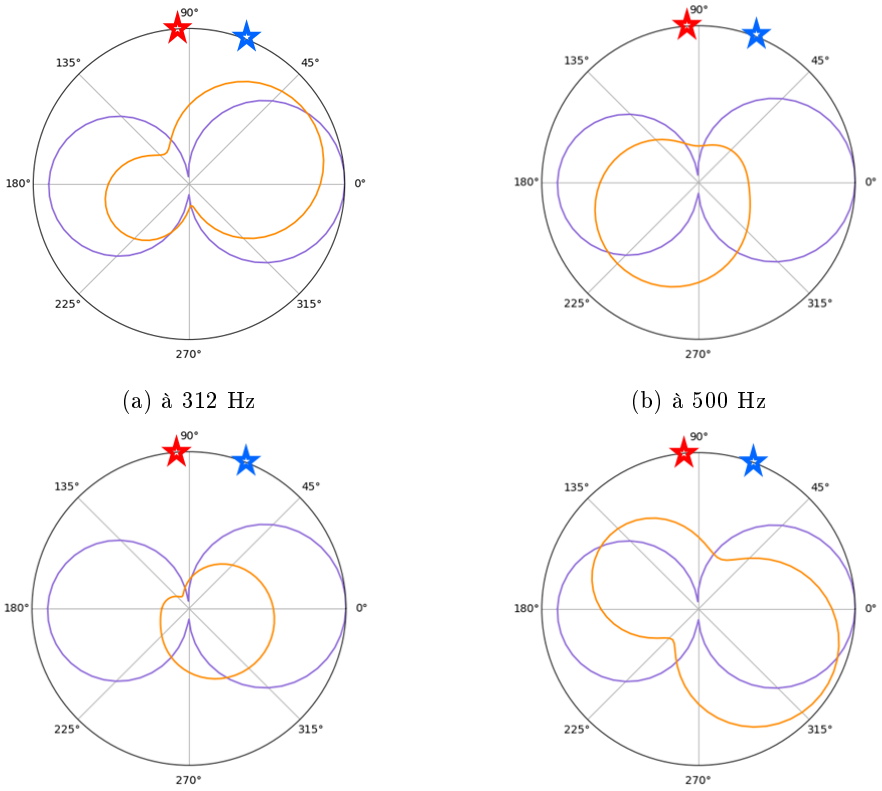
$\sigma(t,f)$ peut être estimée par [51] :

$$\sigma(t,f) = \frac{\text{tr}[\mathbf{R_{ss}}(t,f)]}{\text{tr}[\mathbf{a}(t,f)\mathbf{a}(t,f)^H]}. \tag{3.24}$$

Le MWF de rang 1 s'écrit alors

$$\mathbf{w}_{\text{MWF-r1}}(t,f) = [\mathbf{R_{ss-r1}}(t,f) + \mathbf{R_{nn}}(t,f)]^{-1}\mathbf{R_{ss-r1}}(t,f)\mathbf{u}_0 \tag{3.25}$$

Une implémentation efficace du filtre MWF-r1 est proposée dans [50, eq. (61)]. Une comparaison des directivités à différentes fréquences du MWF-r1 et du filtre de formation de voies FOA pleine bande (2.32) est montrée sur la Figure 3.3. On observe que le MWF-r1 varie de façon importante en fonction de la fréquence. En pratique, cela permet de prendre en compte la réverbération.



(a) à 312 Hz (b) à 500 Hz

(c) à 1000 Hz (d) à 3125 Hz

FIGURE 3.3. – Directivités à 312 Hz, 500 Hz, 1000 Hz et 3125 Hz du filtre FOA pleine bande contraint (en violet) et du MWF-r1 (en orange). La source cible est située à 70° (étoile bleue) et la source interférente à 90° (étoile rouge).

Post-traitement monocanal Les filtres ci-dessus peuvent induire certaines distorsions. Pour y remédier, une normalisation BAN (*blind analytic normalization*) a été proposée pour le max-SNR [48] afin d'assurer un ratio unité entre la réponse spatiale du filtre et la fonction de transfert entre la source cible et le microphone. Le coefficient pondérateur à appliquer en chaque point temps-fréquence est :

$$w_{\mathrm{BAN}}(t,f) = \frac{\sqrt{\mathbf{w}(t,f)^H \mathbf{R_{nn}}(t,f)\mathbf{R_{nn}}(t,f)\mathbf{w}(t,f)/J}}{\mathbf{w}(t,f)^H \mathbf{R_{nn}}(t,f)\mathbf{w}(t,f)} \tag{3.26}$$

On peut également utiliser cette normalisation pour le MWF-r1.

Déréverbération La déréverbération pour la parole est un domaine de recherche large où la littérature abonde [52]. Ce n'est pas l'objet de cette thèse, mais nous en présentons les principes généraux car il est prouvé qu'elle améliore largement les performances de RAP [53]. Il s'agit de retrouver le signal source t_i à partir de la source image $\mathbf{s} = \mathbf{c}_i$

captée par un ensemble de microphones. Nous omettons ici les indices temporels ou spectro-temporels, puisque ce qui suit est valable dans les deux cas.

— Le premier type d'algorithmes, dits d'annulation de réverbération, considèrent le modèle convolutif (3.3). Certains estiment les réponses impulsionnelles $\mathbf{h}_i$ entre la source et les microphones puis un filtre inverse [54]. Il est préférable d'éviter l'inversion directe, rarement stable. D'autres algorithmes estiment conjointement les réponses impulsionnelles $\mathbf{h}_i$ et les signaux sources t_i grâce à un modèle de Markov couplé à un algorithme d'espérance-maximisation [55]. Enfin, les algorithmes d'estimation linéaire multicanale, dont le WPE (*weighted prediction error*) [56] déterminent un filtre linéaire qui minimise l'erreur d'estimation du signal source en considérant la parole comme un processus auto-régressif dans le domaine fréquentiel.

— Les algorithmes de suppression de la réverbération considèrent quant à eux le modèle acoustique additif

$$\mathbf{s} = \mathbf{s}^{(e)} + \mathbf{s}^{(r)} + \mathbf{v} \tag{3.27}$$

où $\mathbf{s}^{(e)}$ est la contribution à la source image due au champ direct et éventuellement aux premières réflexions, tandis que $\mathbf{s}^{(r)}$ est due à la réverbération tardive. $\mathbf{v}$ modélise un bruit diffus ou bruit de mesure. La réverbération tardive est alors supprimée en considérant un modèle de décroissance exponentielle de la réverbération nécessitant uniquement l'estimation du TR60 (temps de réverbération à 60 dB) [57], avec une extension prenant également en compte le rapport direct-à-réverbéré [58]. Dans le cas multicanal, des filtres tels qu'un MVDR ou un MWF peuvent être utilisés pour bloquer le champ direct (à condition de connaître sa direction d'arrivée) [59]. En soustrayant le signal résultant au mélange, on obtient une estimation du signal direct.

— Enfin, il est également possible d'estimer directement le signal source, soit par le biais d'une modélisation source-filtre du processus articulatoire [60], soit en utilisant des réseaux de neurones [61].

Dans le cadre de ce travail, nous utilisons le filtre WPE comme pré-traitement juste avant d'utiliser le module de RAP pour la mesure de performance.

Limites de ces méthodes En pratique, la formation de voies directionnelle n'est pas robuste à la réverbération, tandis que les méthodes de filtrage dépendant de la fréquence reposent sur l'estimation des matrices de covariance $\mathbf{R}_{\mathbf{ss}}(t,f)$ et $\mathbf{R}_{\mathbf{nn}}(t,f)$. Cela nécessitait historiquement d'avoir accès à la détection de l'activité vocale et supposait la stationnarité du bruit, hypothèse qui n'est pas réaliste dans le cas multi-locuteurs. Ces limites sont aujourd'hui traitées par l'utilisation de réseaux de neurones pour l'estimation des matrices de covariance (voir partie 3.2.4).

3.2.3. Méthodes historiques de séparation de sources

Même si nous avons établi que notre cas d'application nécessite plutôt le rehaussement d'une source, il est important de passer rapidement en revue les techniques de séparation

de sources. Ces deux domaines ne se sont pas développés de façon disjointe mais se sont au contraire influencés l'un l'autre.

Séparation aveugle de sources Les méthodes dites de séparation aveugle sont des méthodes non supervisées qui s'appuient uniquement sur l'hypothèse que les signaux cibles, issus de sources différentes, sont indépendants. Les premières méthodes ont été développées dans les années 1990 pour les mélanges instantanés de sources, où la convolution de l'équation (3.3) est une simple multiplication :

$$\mathbf{c}_i(n) = \mathbf{h}_i t_i(n). \tag{3.28}$$

L'algorithme JADE [62] estime les sources grâce à la diagonalisation conjointe de tranches de cumulants d'ordre 4. L'analyse en composantes indépendantes (ACI) généralise ce principe à d'autres mesures de l'indépendance et différents algorithmes d'optimisation [63].
L'algorithme SOBI [64] relâche l'hypothèse d'indépendance des échantillons pour une même source en les considérant temporellement correlées. Les algorithmes utilisant la non-stationnarité considèrent quant à eux que les échantillons ne sont pas identiquement distribués et possèdent une variance qui évolue au cours du temps [65]. Les informations apportées par cette variation des statistiques d'ordre 2, avec l'hypothèse d'indépendance des sources, permettent de maximiser une fonction de vraisemblance et d'estimer les sources.
Les méthodes conçues pour des mélanges instantanés ont été adaptées à des mélanges convolutifs de sources. On remarque en effet que l'expression fréquentielle du mélange (3.11), qui repose sur l'hypothèse de bande étroite, correspond à un mélange instantané dans chaque bande de fréquence [66]. Outre l'approximation due à cette hypothèse, l'application des méthodes de séparation instantanées dans chaque bande de fréquence crée des ambiguïtés de permutation entre les bandes, car les sources sont estimées dans un ordre arbitraire. Pour les résoudre, les solutions proposées s'appuient sur la similarité de l'enveloppe temporelle des signaux entre les bandes de fréquence adjacentes [67], sur l'estimation des directions d'arrivée, ou bien sur une combinaison de ces méthodes [68], avec des résultats cependant mitigés.
Une autre limitation critique des algorithmes d'ACI est qu'ils ne peuvent être utilisés que dans le cas déterminé, où le mélange $\mathbf{x}$ contient autant de canaux que le nombre de sources I à trouver.

Méthodes utilisant la parcimonie Les méthodes de masquage temps-fréquence permettent de traiter des cas sous-déterminés. Le masque $M_{\mathbf{c}_i}(t, f)$ estime la présence du signal $\mathbf{c}_i$ au point (t, f). Il prend des valeurs dans $\{0, 1\}$ s'il s'agit d'un masque binaire ou dans le segment $[0, 1]$ pour un masque continu. Le signal $\mathbf{c}_i(t, f)$ est ensuite estimé à partir du mélange $\mathbf{x}(t, f)$:

$$\hat{\mathbf{c}}_i(t, f) = M_{\mathbf{c}_i}(t, f)\mathbf{x}(t, f). \tag{3.29}$$

Ces masques peuvent être estimés par des techniques de regroupement des points temps-fréquence s'appuyant sur des caractéristiques communes. L'algorithme DUET [39] utilise des informations de localisation, à savoir les différences interaurales de temps et de niveau (ITD, *interaural time difference* et ILD, *interaural level difference*) dans leur forme pleine-bande. MESSL [69] utilise de surcroît des GMMs afin d'exploiter ces informations de façon plus robuste. Cependant, l'approche pleine-bande est problématique en présence de réverbération, puisque l'information spatiale est affectée différemment dans chaque bande de fréquence. Le regroupement en bandes étroites apporte une solution à ce problème mais nécessite une étape supplémentaire pour résoudre les ambiguïtés de permutation entre les différentes bandes de fréquence.

Modèle probabiliste gaussien Des hypothèses statistiques supplémentaires permettent également de traiter le cas sous-déterminé, que ce soit pour des mélanges instantanés [70, 71] ou réverbérants [72]. Dans le cadre probabiliste gaussien, on suppose qu'une image spatiale $\mathbf{c}_i(t,f)$ prend des valeurs indépendantes en chaque point (t,f), suivant une distribution gaussienne complexe isotrope centrée de matrice de covariance $\mathbf{R}_{\mathbf{c}_i\mathbf{c}_i}(t,f)$ qui prend la forme suivante :

$$\mathbf{c}_i(t,f) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{\mathbf{c}_i\mathbf{c}_i}(t,f)) \text{ avec } \mathbf{R}_{\mathbf{c}_i\mathbf{c}_i}(t,f) = v_i(t,f)\mathbf{R}_i(f). \tag{3.30}$$

$v_i(t,f) \in \mathbb{R}^+$ est la densité spectrale de puissance et $\mathbf{R}_i(f)$ est la matrice de covariance spatiale (indépendante du temps pour une source immobile) associées à la source i.
Les $v_i(t,f)$ et $\mathbf{R}_i(f)$ peuvent être estimées itérativement grâce à un algorithme d'espérance-maximisation (EM) [73]. Les estimations des images spatiales $\hat{\mathbf{c}}_i(t,f)$ sont alors obtenues à partir du mélange grâce à un filtre de Wiener variant dans le temps (voir aussi la partie 3.2.2) :

$$\hat{\mathbf{c}}_i(t,f) = \mathbf{R}_{\mathbf{c}_i\mathbf{c}_i}(t,f)\mathbf{R}_{\mathbf{xx}}(t,f)^{-1}\mathbf{x}(t,f) \tag{3.31}$$

avec $\mathbf{R}_{\mathbf{xx}}(t,f) = \sum_i \mathbf{R}_{\mathbf{c}_i\mathbf{c}_i}(t,f)$ la matrice de covariance du mélange $\mathbf{x}(t,f)$.

Factorisation matricielle positive L'utilisation de la factorisation en matrices à valeurs positives (NMF, *non-negative matrix factorization*) fait quant à elle des hypothèses fortes sur la structure des signaux à reconstruire. Les spectres d'amplitude ou de puissance des signaux sont supposés être constitués d'une somme d'éléments de rang 1, chacun décomposable en produit d'un spectre de base et d'une séquence d'amplitudes [74]. Cela correspond par exemple à la décomposition de la parole en phonèmes, ou de la musique en notes, et à la signature spectro-temporelle spécifique de chacun de ces éléments. Une telle grandeur $\mathbf{V}$ est alors modélisée par

$$\hat{\mathbf{V}}(t,f) = \sum_k \mathbf{b}_k(f)\mathbf{h}_k(t) \tag{3.32}$$

où $\mathbf{b}_k = [b_k(0), ..., b_k(F)]^T$ et $\mathbf{h}_k = [h_k(0), ..., h_k(T)]$ sont des vecteurs à valeurs positives représentant respectivement les motifs spectraux et les activations temporelles de l'élément k.

Ces éléments peuvent être estimés de façon non-supervisée par un algorithme de partitionnement de type k-moyennes [75] combiné à une estimation itérative des paramètres par EM. Les résultats ne sont satisfaisants que si les sources à séparer ont des supports suffisamment disjoints. Des versions supervisées de la NMF ont donc été développées, impliquant un apprentissage préalable de dictionnaires spectraux [76]. Cela nécessite d'avoir accès aux sources de manière indépendante pendant un laps de temps avant de traiter leur mélange. Enfin, des techniques intermédiaires dites « semi-supervisées » initient l'apprentissage des dictionnaires spectraux hors-ligne, puis affinent cet apprentissage à partir du mélange [77]. Ces différentes variantes sont initialement conçues pour être appliquées à un mélange monocanal, mais des versions multicanales ont également été mises au point [78]. La NMF a classiquement été combinée avec le modèle probabiliste Gaussien présenté dans le paragraphe précédent [78]. Dans ce cas, la NMF modélise la variance d'une source ou d'un mélange de source. Si chaque composante NMF représente une source, la séparation est directement obtenue. En revanche, si une source est représentée par plusieurs composantes, il est de surcroît nécessaire de regrouper ces composantes, ce qui peut par exemple être fait selon un critère spatial [78].

3.2.4. Méthodes de rehaussement utilisant des réseaux de neurones

Les méthodes de rehaussement de la parole utilisant des réseaux de neurones profonds ont d'abord eu un succès notable dans le contexte monocanal. Elles s'appuient sur différentes architectures de réseau, basées sur des couches totalement connectées (FF, *feed-forward*) [79] ou RNN [80], dont LSTM (*long short-term memory*) [81, 82] et LSTM bidirectionnelles (biLSTM) [83]. De façon similaire aux méthodes de la partie 3.2.3, le réseau peut être entraîné à reconstruire le signal source lui-même [84], parfois de façon générative [85], ou bien un masque qui permet d'estimer la proportion de signal cible présente dans le mélange en chaque point temps-fréquence [86]. Nous nous intéressons ci-dessous plus en détail à l'utilisation de réseaux de neurones dans le cadre multicanal.

3.2.4.1. Rehaussement d'un locuteur seul

L'information multicanale a été utilisée sous diverses formes pour faciliter l'apprentissage du réseau : en combinant l'ITD et l'ILD à une information monocanale [87, 84] ; en comparant les différences de phase effectives et celles, théoriques, correspondant à la direction supposée connue de la source d'intérêt [88] ; ou en mesurant le caractère diffus du signal [89]. Bien que certains travaux cherchent à estimer directement le signal cible $\mathbf{s}(t, f)$ en sortie du réseau [84], la plupart reposent sur l'estimation de masques temps-fréquence comme étape intermédiaire [90]. Ces masques peuvent être appliqués tels quels à chaque canal du signal [87], mais ils induisent alors des distorsions défavorables à la RAP [34].

Filtrage multicanal utilisant un masque temps-fréquence Pour éviter les distortions résultant d'un traitement monocanal, il a été proposé d'utiliser les masques estimés par un réseau de neurones pour l'estimation de filtres multicanaux [34, 49, 51, 91]. Le principe

de cette approche est présenté dans la Figure 3.4 : le masque $M_{\mathbf{s}}(t,f)$ permet d'estimer le signal cible $\mathbf{s}(t,f)$ et le signal interférent $\mathbf{n}(t,f)$:

$$\begin{aligned}\hat{\mathbf{s}}(t,f) &= M_{\mathbf{s}}(t,f)\mathbf{x}(t,f)\\ \hat{\mathbf{n}}(t,f) &= (1-M_{\mathbf{s}}(t,f))\mathbf{x}(t,f)\end{aligned} \tag{3.33}$$

Les matrices de covariance peuvent ensuite être estimées en moyenne sur un intervalle de temps

$$\begin{aligned}\mathbf{R}_{\mathbf{ss}}(f) &= \sum_t \hat{\mathbf{s}}(t,f)\hat{\mathbf{s}}(t,f)^H\\ \mathbf{R}_{\mathbf{nn}}(f) &= \sum_t \hat{\mathbf{n}}(t,f)\hat{\mathbf{n}}(t,f)^H.\end{aligned} \tag{3.34}$$

et être utilisées pour calculer un des filtres présentés dans la partie 3.2.2, qui est alors invariant sur cet intervalle de temps. Certains travaux calculent un seul masque pour l'ensemble des canaux [92]. D'autres estiment un masque par canal, puis un unique masque médian est appliqué à tous les canaux pour plus de robustesse [49, 93].

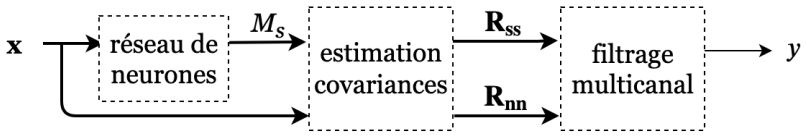


FIGURE 3.4. – Filtrage multicanal utilisant des masques temps-fréquence estimés par un réseau de neurones.

Cette formulation permet de relâcher les hypothèses de stationnarité du bruit et d'indépendance des signaux, ainsi que la nécessité d'avoir accès à une estimation préalable de l'activité vocale, conditions auparavant indispensables à l'utilisation des filtres de la partie 3.2.2. Elle a permis d'améliorer significativement les performances de rehaussement d'une source de parole dans un bruit diffus, y compris en environnements réels [51].

Estimation directe des paramètres de filtrage Enfin, certaines méthodes proposent d'utiliser les réseaux de neurones pour estimer directement les paramètres des filtres, sans passer par le calcul des matrices de covariance. Par exemple, Xiao et al. [94] optimisent directement les coefficients des filtres selon des critères de RAP.
Nugraha et al. [95] utilisent des DNNs pour estimer les spectres des signaux sur lesquels repose le MWF dans le cadre probabiliste gaussien (3.30), permettant le calcul de filtres multicanaux variant dans le temps.

3.2.4.2. Le cas multi-locuteurs

Le problème se complexifie lorsque plusieurs locutrices sont présentes simultanément. En effet, deux signaux de parole ont des caractéristiques spectro-temporelles plus proches qu'un signal de parole et un signal de bruit ambiant, ce qui rend leur séparation intrinsèquement plus difficile. Pour autant, les auditeurs humains sont généralement capables

d'extraire la voix d'une locutrice en présence de parole interférente, y compris lorsque le niveau de celle-ci est plus élevé que celui de la cible [96]. Cela est certes plus facile lorsque les voix sont déjà connues, que ce soit pour un auditeur humain ou pour les réseaux de neurones [97]. Comprendre le sens des discours facilite également leur séparation. Utiliser des informations de reconnaissance vocale a été proposé pour faciliter le rehaussement [83] en fournissant au réseau de rehaussement le senone le plus probable selon un système de RAP. À notre connaissance, l'information de RAP n'a jamais été utilisée dans le cas multi-locuteurs.
Dans les paragraphes suivant, nous exposons les difficultés propres à la situation multi-locuteurs et les pistes existantes pour y remédier.

Problème de permutation des étiquettes d'apprentissage Pour un système de séparation automatique de la parole, en plus de la difficulté propre à la situation multi-locuteurs, des problèmes techniques se posent. Imaginons deux personnes parlant aussi fort l'une que l'autre, Anwen et Bodmaël. Lors de l'apprentissage, si le réseau estime parfaitement bien les sorties correspondant à Bodmaël et Anwen, mais que la fonction de coût est calculée avec les étiquettes correspondant, dans l'ordre, à Anwen et Bodmaël, le réseau sera pénalisé pour une réponse pourtant correcte. Il lui est impossible d'apprendre dans ces conditions. Ce problème est appelé « l'ambiguïté des étiquettes » (voir Figure 3.5).

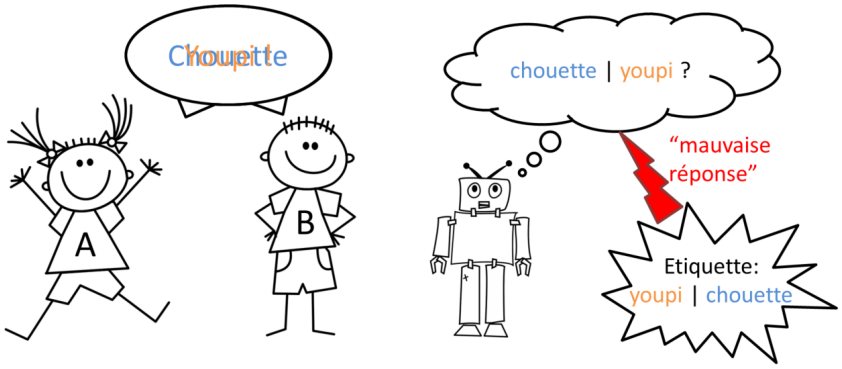


FIGURE 3.5. – Illustration du problème de permutation des étiquettes. Lors de l'apprentissage, le réseau peut être pénalisé pour une estimation correcte présentée dans un ordre différent de celui des étiquettes d'apprentissage.

Désambiguïsation par RAP Weng et al. [98] se servent de la RAP pour résoudre ce problème de permutation (mais pas pour faciliter la séparation elle-même). Ils utilisent l'hypothèse, presque toujours vérifiée, que sur chaque trame la puissance de l'un des deux signaux de parole est plus forte que l'autre. Deux réseaux de neurones sont appris pour la RAP, l'un apprenant à renvoyer le senone correspondant au signal le plus fort, l'autre au signal le moins fort. Les senones correspondant à chaque locuteur sont regroupés au fil des trames par un critère de vraisemblance portant sur les transcriptions finales

des deux signaux, avec une pénalité empêchant un changement trop fréquent entre les senones de puissance plus élevée ou plus faible. Ce système ne permet pas de reconstituer les signaux audio séparés, mais est adapté à notre application finale de reconnaissance de la parole. Cependant, il ne fonctionne que dans un scénario avec exactement deux locuteurs. D'autre part, selon les auteurs eux-mêmes, la désambiguïsation des senones trame-à-trame telle qu'elle est proposée n'est robuste que dans le cas d'un vocabulaire et d'une grammaire réduits.

L'apprentissage invariant aux permutations (PIT, *permutation invariant training*) [99], propose une solution simple qui consiste à calculer les fonctions de coût pour toutes les permutations et choisir celle dont la valeur est minimale. Le calcul de la fonction de coût n'étant pas très complexe, le surcoût engendré est minime, et ne concerne que la phase d'apprentissage. Il peut donc être étendu dans le cas d'un nombre de locuteurs quelconque, mais connu et fixe, car il détermine l'architecture du réseau. Initialement présentée pour le cas monocanal, une extension de cette méthode a également été proposée pour exploiter l'information multicanale [100].

Modèles génératifs Une autre façon de se libérer du problème de permutation des étiquettes est de supprimer lesdites étiquettes. C'est possible en adoptant une approche non-supervisée. Dans le cadre de l'apprentissage profond, on peut utiliser des réseaux génératifs adversariaux, ou GANs (*generative adversarial networks*) [101]. Un réseau est conçu pour synthétiser les signaux sources à reconstituer étant donné le mélange, tandis qu'un second doit distinguer les signaux synthétisés des vrais signaux. En apprenant ces réseaux simultanément et sans utiliser d'étiquettes, le problème de l'ambiguïté est évité.

Partitionnement profond Le *deep clustering* [102] propose une méthode différente. Au lieu d'utiliser un réseau [98] ou une sortie de réseau [99] pour chaque source, elle utilise la capacité des techniques de partitionnement à gérer des groupes non étiquetés. Les techniques classiques type k-moyennes ne sont pas adaptées à des situations aussi complexes que le partitionnement de points temps-fréquence pour plusieurs sources de parole, faute d'une métrique adaptée dans ce domaine. Le *deep clustering* utilise des réseaux de neurones pour associer à chaque point temps-fréquence une représentation dans un espace de dimension supérieure, où la distance canonique est effectivement représentative de l'identité du locuteur prédominant en chaque point. Dans cet espace, il est alors possible d'utiliser les k-moyennes pour regrouper les points temps-fréquence par locuteur. Une variante, nommée *deep attractor networks* [103], intègre directement l'étape de partitionnement au réseau.
Un réseau appris avec cette méthode peut être appliqué à un nombre quelconque de sources, dont la connaissance n'est nécessaire qu'à l'étape finale de partitionnement. Le *deep clustering* est très flexible en théorie, mais l'étape supplémentaire de partitionnement ajoute une certaine complexité lors du décodage. De plus, pour être robuste, le décodage dans l'espace de haute dimension doit être effectué sur plusieurs trames. Ce coût calculatoire empêche son utilisation en temps quasi-réel. Comme dans le cas du PIT,

une extension multicanale a été proposée [104].

Une seule locutrice cible Les méthodes présentées ci-dessus cherchent à extraire toutes les sources de parole. Comme nous l'avons vu, dans notre cas d'application, le problème est plus restrictif : il s'agit d'extraire le discours d'une seule locutrice. Cela peut sembler plus simple, mais nécessite une information supplémentaire qui permette au système d'identifier la locutrice cible. Si cette information est présente, cela résout un problème non traité par les méthodes ci-dessus : suivre des locutrices d'une phrase à l'autre, afin que les transcriptions finales soient cohérentes.
Au niveau de la RAP, il est possible d'utiliser un mot-clé qui focalise directement l'attention du moteur de RAP sur la locutrice cible [30]. En amont, au niveau signal, il a par exemple été proposé d'utiliser une phrase où la locutrice cible est seule afin de générer une représentation de la locutrice via un réseau auxiliaire. Cette représentation permet ensuite d'adapter les poids du réseau principal à la locutrice et de l'extraire du mélange par la suite [93]. Cependant, cette identification n'est pas suffisante dans les cas difficiles, notamment lorsque les locutrices sont de même genre.

3.2.5. Résumé et positionnement

Nous venons d'exposer les techniques existant pour le rehaussement de sources de paroles. Aujourd'hui, qu'il y ait une ou plusieurs sources, les plus efficaces s'appuient sur des réseaux de neurones. En particulier, en amont de la RAP, le cadre du filtrage multicanal s'appuyant sur des masques calculés par un réseau de neurones a fait ses preuves pour le débruitage d'une source de parole. En effet, la composante spatiale de ces filtres permet de limiter les distorsions, ce qui en fait une méthode adaptée au pré-traitement en amont de la RAP. D'autre part, aucune méthode utilisant les réseaux de neurones n'a jusqu'ici été présentée pour le rehaussement de signaux ambisoniques.
Nous présentons donc dans le chapitre 5 une technique qui s'inscrit dans le filtrage multicanal utilisant un réseau de neurones, mise au point afin de pouvoir être utilisée sur des contenus ambisoniques.
Par ailleurs, en raison du problème de la permutation d'étiquette, les techniques existant aujourd'hui dans le cadre du filtrage utilisant des réseaux de neurones ne peuvent être utilisées dans le cas multi-locuteurs. Nous proposons donc une méthode adaptée à cette situation qui utilise la connaissance de la direction d'arrivée des sources pour lever l'ambiguïté. La partie suivante présente un état de l'art de la localisation de sources sonores.

3.3. Localisation de sources sonores

De nombreuses techniques de rehaussement de sources multicanales nécessitent la localisation préalable des sources sonores [105, 106, 107, 91]. Nous dressons ici un état de l'art de ce domaine, avant de proposer une solution adaptée à l'ambisonie dans le chapitre 4. Les techniques classiques de localisation utilisent des principes théoriques variés présen-

tés ci-dessous. Ces principes sont encore d'actualité, comme le prouvent notamment les différentes soumissions au challenge LOCATA en 2018 [108]. Nous n'incluons dans la tâche de localisation que l'estimation de la direction d'arrivée des sons, c'est-à-dire du couple azimut et élévation (θ, ϕ), sans chercher à déterminer la distance entre la source et le microphone.

3.3.1. Techniques traditionnelles

Une revue des méthodes classiques de localisation de sources peut être trouvée dans [109]. Nous présentons ici les grands principes de ces différentes méthodes, ainsi que les façons dont elles sont exploitées dans des travaux récents.

Différence de temps d'arrivée Étant donné l'enregistrement d'une scène sonore par plusieurs microphones, la différence de temps d'arrivée entre ces microphones est une quantité souvent utilisée pour la localisation. Combinée à la connaissance de l'antenne de microphones, elle permet d'obtenir la direction d'arrivée. Dans le cas d'une seule paire de microphones, la direction ne peut pas être déterminée complètement. S'ils sont placés dans le plan médian par exemple, seul l'azimut peut être estimé, et une ambiguïté subsiste entre le demi-plan avant et le demi-plan arrière.
La différence de temps d'arrivée sur une trame t peut être estimée par GCC-PHAT (*generalized cross correlation phase transform*) [110], c'est-à-dire la transformée de Fourier inverse de la corrélation spectrale croisée entre les signaux captés par deux microphones j et j' :

$$\psi_{jj'}(t,\tau) = \frac{1}{F}\sum_{f=0}^{F-1} \frac{x_j(t,f)x_{j'}^*(t,f)}{|x_j(t,f)||x_{j'}^*(t,f)|} e^{2i\pi\frac{f\tau}{F}}. \tag{3.35}$$

Le suffixe PHAT désigne le fait de normaliser par l'amplitude en ne conservant que l'information de phase. $x_j(t,f)$ et $x_{j'}(t,f)$ sont les coefficients de TFCT (voir partie 3.2.1) des signaux captés par les micros j et j' ; * est l'opérateur de conjugaison complexe ; F est le nombre de bandes de fréquence de la TFCT. La différence de temps d'arrivée à la trame t est alors le délai maximisant GCC-PHAT :

$$\Delta_{jj'}(t) = \arg\max_{\tau} \psi_{jj'}(t,\tau). \tag{3.36}$$

Pour plus de deux microphones, la différence de temps d'arrivée est calculée pour chaque paire. La direction d'arrivée est déduite par triangulation, ce qui revient à faire une fusion tardive de la localisation pour toutes les paires de microphones (j, j'). Cette technique n'est pas adaptée à des antennes compactes de microphones, car la triangulation n'est pas robuste, ou au cas de plusieurs sources.

Cartes acoustiques L'algorithme SRP-PHAT (*steered response power with phase transform*) [111, 109] étend cette technique grâce à la fusion précoce des GCC-PHAT pour toutes les paires de microphones. Une carte acoustique $\mathcal{M}$ de la scène sonore est alors

créée :

$$\mathcal{M}(t,\theta,\phi) = \sum_{jj'} \psi_{jj'}(t, \tau_{jj'}(\theta,\phi)), \tag{3.37}$$

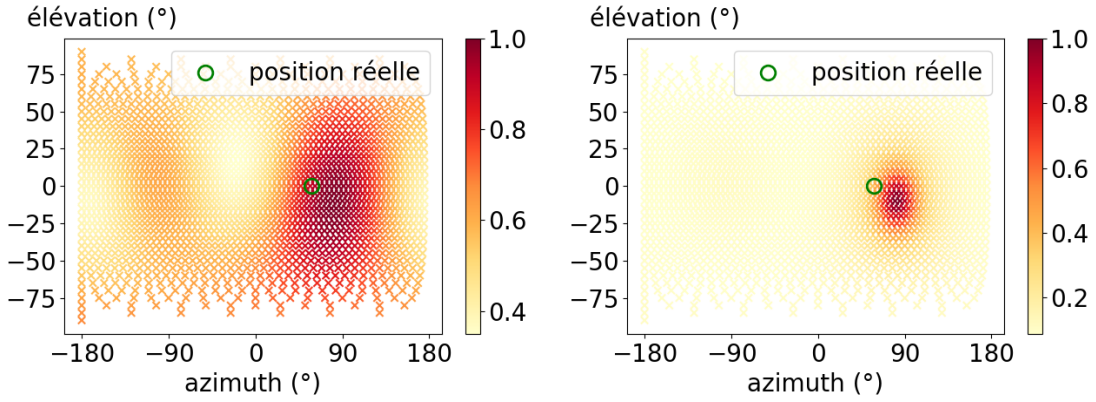
où $\tau_{jj'}(\theta,\phi)$ représente la différence de temps d'arrivée entre les micros j et j' pour une onde sonore provenant de la direction (θ,ϕ).
Cela est mathématiquement équivalent à appliquer un filtre de formation de voie de type *delay-and-sum* pour chaque direction possible, ce qui évite le calcul explicite des différences de temps d'arrivée :

$$\mathcal{M}(t,\theta,\phi) = \sum_{f=0}^{F-1} \mathbf{d}_{\theta,\phi}(f)^H \frac{\mathbf{x}(t,f)}{|\mathbf{x}(t,f)|} \frac{\mathbf{x}(t,f)^H}{|\mathbf{x}(t,f)|} \mathbf{d}_{\theta,\phi}(f), \tag{3.38}$$

où $\mathbf{d}_{\theta,\phi}(f)$ est le vecteur directionnel correspondant à l'antenne de microphones et à la direction (θ,ϕ), défini dans la partie 3.2.2. Une soumission au challenge LOCATA [112] propose une implémentation efficace de SRP-PHAT et l'évalue en conditions réelles pour une source, en obtenant de meilleurs résultats que MUSIC (voir paragraphe suivant) dans ces conditions. Un exemple de carte acoustique établie par SRP-PHAT est présenté dans la Figure 3.6a.
Les méthodes précédentes utilisent une normalisation en amplitude pour plus de robustesse à la réverbération. Cependant, en supprimant l'information d'amplitude, cela empêche de traiter correctement le cas de plusieurs sources concurrentes [106, 113]. Pour traiter ce cas, d'autres méthodes proposent d'estimer la direction de la source prédominante en chaque point temps-fréquence. On peut alors regrouper les points temps-fréquence pour faire apparaître les directions des sources [113], par exemple avec un algorithme de partitionnement de type k-moyennes, ou bien en sélectionnant les maxima d'un histogramme. De manière générale cependant, l'estimation de la direction est sensible à la réverbération. En particulier, les premières réflexions peuvent apparaître comme des sources supplémentaires.

Sous-espaces vectoriels Une autre famille d'algorithmes de localisation utilise la décomposition en éléments propres de la matrice de covariance $\mathbf{R}_{\mathbf{xx}}(t,f)$ (3.18) entre les différents canaux du signal capté $\mathbf{x}(t,f)$.
La méthode ESPRIT (*estimation of signal parameters via rotational invariance techniques*) [114] utilise des propriétés d'invariance rotationnelle de l'antenne de microphones pour faire apparaître le sous-espace significatif (par opposition à l'espace du bruit).
MUSIC (*multiple signal classification*) [115] sélectionne les vecteurs propres associés aux valeurs propres jugées suffisamment grandes pour correspondre à une source ponctuelle et non à du bruit. Les vecteurs propres fournissent alors une estimation des vecteurs directionnels $\mathbf{d}_{\theta,\phi}(f)$ correspondant aux directions des sources. La carte acoustique correspondante est établie en calculant la grandeur suivante (voir Figure 3.6b) :

$$\mathcal{M}(\theta,\phi,t,f) = \frac{1}{\mathbf{d}_{\theta,\phi}^H(f)\mathbf{U}_\mathbf{n}(t,f)\mathbf{U}_\mathbf{n}(t,f)^H\mathbf{d}_{\theta,\phi}(f)} \tag{3.39}$$



(a) SRP-PHAT (b) MUSIC

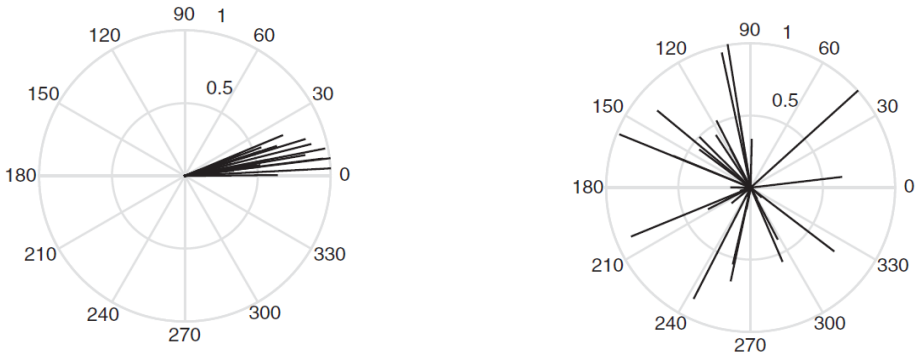
FIGURE 3.6. – Cartes acoustiques d'un signal ambisonique générées par (a) SRP-PHAT en utilisant le filtre de formation de voie (2.32) et (b) MUSIC.

où $\mathbf{U_n}(t, f)$ est le sous-espace vectoriel constitué des vecteurs propres de $\mathbf{R_{xx}}(t, f)$ correspondant au bruit.
MUSIC peut être combiné avec des techniques de partitionnement des points temps-fréquence. Le critère de dominance du chemin direct [116] permet de déterminer les points où une seule source prédomine, auquel cas il suffit de ne conserver que la composante la plus énergétique de la matrice de covariance. On regroupe ensuite les directions apparentes pour tous les points temps-fréquence. Cette technique a été évaluée dans le cadre de LOCATA [117] ; les performances ne sont cependant pas rapportées.

Analyse en composantes indépendantes Les méthodes d'ACI visent à retrouver la matrice de mélange qui relie des phénomènes indépendants à l'observation de leur mélange instantané. Dans le cas de l'audio, elles ont avant tout été mises au point pour la séparation de sources (voir la partie 3.2.3) ; il a cependant été constaté que ces matrices de mélange permettent de retrouver les vecteurs directionnels pointant vers les sources [118, 119, 120]. L'ACI s'est montrée plus robuste que MUSIC [118] et SRP-PHAT [119] dans le cas de plusieurs sources, notamment si celles-ci sont proches. Cependant, ces méthodes ne sont pas adaptées à la séparation de sources en présence de réverbération importante qui met à mal l'hypothèse de mélange instantané (voir partie 3.2.3). C'est également le cas pour la localisation.

Vecteur d'intensité acoustique Le vecteur d'intensité acoustique active, présenté pour le cas ambisonique dans la partie 2.3.3, pointe par définition dans le sens de propagation de l'énergie acoustique, c'est-à-dire l'opposé de la direction d'arrivée. Il est donc naturel de s'en servir pour la localisation [121, 122]. Il nécessite de connaître la vélocité, estimée à partir du gradient de pression entre plusieurs microphones ou directement mesurée grâce à des capteurs acoustiques coûteux [123] (plus courants dans le domaine des radars qu'en

audio). On peut encore une fois utiliser le regroupement des points temps-fréquence pour gérer la présence simultanée de plusieurs sources. Cependant, malgré les efforts pour modéliser la réverbération [124], ces méthodes y restent très sensibles [125] (voir Figure 3.7).



(a) Environnement anéchoïque (b) Environnement réverbérant

FIGURE 3.7. – Vecteur d'intensité acoustique actif pour une fréquence fixée et plusieurs trames consécutives, (a) sans réverbération et (b) avec réverbération selon Pulkki [11].

3.3.2. Techniques adaptées à l'ambisonie

Différences de niveau La plupart des concepts de localisation présentés ci-dessus sont transposables aux signaux ambisoniques. Le cas le moins évident est celui de la différence de temps d'arrivée, puisque pour un encodage parfait les canaux ambisoniques correspondent à des capteurs coïncidents : il n'y a donc pas de différences de temps d'arrivée entre eux. Cependant, les différences de temps d'arrivée au format A (capté par l'antenne de microphones réels) sont utilisées dans l'encodage FOA. Cette information est donc présente de façon implicite dans les rapports d'amplitude des canaux FOA. La direction d'arrivée y apparaît de façon simple d'après (2.25) [126] :

$$\theta(t,f) = \arctan\left(\frac{|Y(t,f)|}{|X(t,f)|}\right) \tag{3.40}$$

$$\phi(t,f) = \arctan\left(\frac{|Z(t,f)|}{\sqrt{|X(t,f)|^2 + |Y(t,f)|^2}}\right) \tag{3.41}$$

L'encodage (2.25) n'est valable que dans le cas d'une onde plane. Cette limitation apparente est mise à profit dans une participation au challenge LOCATA [127] afin d'identifier les points temps-fréquence dominées par le champ sonore direct. Le « critère onde plane » y est défini de la façon suivante :

$$C(t,f) = \frac{|X(t,f)|^2 + |Y(t,f)|^2 + |Z(t,f)|^2}{|W(t,f)|^2}. \tag{3.42}$$

Il permet de pondérer l'importance des directions d'arrivée calculées en chaque point temps-fréquence grâce à (3.40) et (3.41) : plus $C(t, f)$ est proche de 1, plus le point (t, f) est significatif.

SRP-PHAT et sous-espaces On a vu dans la partie 2.3.2 l'expression du filtre de formation de voies ambisonique. En l'utilisant pour scanner l'espace des directions, on produit une carte acoustique où les sources principales sont identifiables [128] comme dans un algorithme SRP-PHAT classique.
Les méthodes de décomposition en sous-espaces, MUSIC et ESPRIT, ont aussi été appliquées à des contenus FOA [116, 129].

L'analyse en composantes indépendantes peut également être appliquée à des contenus ambisoniques [130, 12]. Dans ce cas, la matrice de mélange est

$$\mathbf{A} = [\mathbf{d}_{\theta_0,\phi_0}\, \mathbf{d}_{\theta_1,\phi_1} \dots \mathbf{d}_{\theta_J,\phi_J}], \tag{3.43}$$

où les $\mathbf{d}_{\theta_j,\phi_j}$ sont les vecteurs directionnels pointant vers les sources (2.32). Les directions sont déduites de ces colonnes par (3.40) et (3.41).

Le vecteur d'intensité est particulièrement adapté au format FOA [131, 121, 128] où il est exprimé simplement en fonction des canaux W, X, Y et Z (voir partie 2.3.3). Mathématiquement, utiliser la direction pointée par le vecteur d'intensité acoustique active correspond exactement à faire le rapport entre les canaux comme dans (3.40) et (3.41). Des améliorations augmentant la robustesse des estimations ont été présentées à LOCATA [127, 132]. L'intensité réactive, quant à elle, n'a jamais été utilisée pour la localisation à notre connaissance.

3.3.3. L'apprentissage supervisé non neuronal

Les méthodes précédentes reposent sur une modélisation mathématique des phénomènes physiques mis en jeu, mais celle-ci ne peut prendre en compte toute la complexité et la diversité des phénomènes réels. Il est difficile de prendre en compte le bruit, les interférences et la réverbération, qui changent dans chaque cas.
Les techniques d'apprentissage supervisé cherchent à modéliser ces phénomènes de façon implicite. Elles utilisent des fonctions comprenant un grand nombre de paramètres, qu'elles optimisent automatiquement grâce à des exemples dont on connaît la vérité terrain. Avec des données bien choisies, cela revient dans certains cas à effectuer une classification automatique des points temps-fréquence qui prend en compte des phénomènes acoustiques complexes.

Estimateurs par noyau Les travaux précurseurs de Roman et al. [107, 133] proposent d'apprendre des fonctions non-linéaires et dépendant de la fréquence afin de procéder conjointement à la localisation et à la séparation de deux sources. Une première fonction est apprise afin d'associer les différences de temps d'arrivée mesurées dans un contexte

binaural aux azimuts des sources. Dans un second temps, cela permet d'apprendre des estimateurs par noyau pour tous les azimuts possibles. Ceux-ci associent les indices binauraux (ITD et ILD) à la différence d'énergie entre les sources. Grâce à ces différences d'énergie, il est possible d'estimer un masque temps-fréquence et de séparer les sources présentes.

Régression d'arête Wilson et Darrell [134] utilisent également l'apprentissage supervisé pour parer aux limitations de GCC-PHAT, en particulier en présence de réverbération. L'apprentissage ne vise pas à retrouver directement les directions d'arrivée, mais à estimer la fiabilité de GCC-PHAT pour chaque point temps-fréquence afin d'en pondérer les estimations. Plus précisément, un ensemble d'apprentissage est synthétisé avec la méthode des sources images [135]. On connaît donc les directions d'arrivée théoriques des sources, qui sont également estimées par GCC-PHAT. Un apprentissage par régression d'arête (régression linéaire par morceaux avec régularisation L2) permet d'associer un spectrogramme donné à une précision de localisation par GCC-PHAT, définie en chaque point temps-fréquence comme étant l'inverse de l'erreur de GCC-PHAT au carré. Fait intéressant, les points les plus fiables pour la localisation avec GCC-PHAT correspondent aux attaques des sons, ce qui a également été démontré pour l'être humain en psychoacoustique [136].

Mélanges de gaussiennes Les fonctions utilisées dans les travaux cités ci-dessus, même s'ils sont plus complexes que les modèles de la partie 3.3.1, possèdent un pouvoir de modélisation encore limité. En revanche, les GMMs peuvent théoriquement modéliser n'importe quelle fonction si le nombre de composantes gaussiennes est suffisamment élevé. Pour chaque direction considérée, un GMM est entraîné à modéliser la distribution des indices choisis (en général ILD et ITD) pour tous les points temps-fréquence [69]. Pour l'analyse d'un mélange inconnu, il suffit ensuite de sélectionner la direction correspondant à la probabilité maximale a posteriori [137, 138]. Une approche similaire peut être adoptée en utilisant des machines à vecteurs support à la place des GMMs [139].
L'algorithme GLLiM (*Gaussian locally linear mapping*) [140] utilise des techniques de modélisation par GMMs en y incorporant une information précieuse : l'espace d'arrivée est de dimension 2 (azimut et élévation). Puisqu'il existe une fonction projetant l'espace de départ de très haute dimension (typiquement celui des indices binauraux, ITD et ILD, calculés en chaque point du spectrogramme) sur l'espace d'arrivée, Deleforge et al. prouvent que les données binaurales reposent sur une surface de dimension 2 dans l'espace de départ. L'idée que les données de départ reposent sur une variété de dimension inférieure à celle de l'espace qui les contient a été utilisée pour plusieurs problèmes de traitement du signal audio [141].

Pour l'ambisonie ? À notre connaissance, pour des contenus ambisoniques, il n'existe aucune technique de localisation reposant sur un apprentissage supervisé mais ne faisant pas appel à des réseaux de neurones profonds.

Référence	Données d'entrée	Contexte	Architecture	Estimations	Sources	Apprentissage	Test
Xiao [142]	GCCs	0.2 s	FF	θ 360 classes	1	3 salles simu.	salles simu. et réelles sur la grille
Takeda [143]	vecteurs propres des covariances	0.2 s 812-4812Hz	FF	θ 72 classes	multi (nb connu)	1 salle réelle anéch.	anéch. et réverb. sur la grille
Chakrabarty [144]	phases TFCT	1 trame	CNN (micros et freq.)	θ 37 classes	1	35 salles simu. signaux de bruit	salles simu. et réelles sur la grille
Chakrabarty [145]	phases TFCT	1 trame	CNN (micros)	θ 37 classes	multi (nb connu)	35 salles simu. signaux de bruit	salle simu. sur la grille
Ma [146]	GCCs et ILDs	1 trame	FF	θ 72 classes	multi (nb connu)	salle réelle anéch.	salle réelle reverb. sur la grille
Adavanne [147]	TFCT FOA	2 s	CRNN	θ, ϕ 614 classes	multi (nb inconnu)	salle simu.	salles simu. sur la grille
Adavanne [148]	TFCT FOA	1.5 s	CRNN	x, y, z sur la sphère régression	multi (max. 1 par type)	simu. et réelles	salles réelles
Chakrabarty [149]	phases TFCT	1 trame	CNN à dilatation (micros)	θ 37 classes	multi (nb connu)	35 salles simu. signaux de bruit	salle simu. sur la grille
Comminiello [150]	FOA temporel	3 s	Q-CRNN à dilatation	x, y, z sur la sphère régression	multi (max. 1 par type)	salle simu. anéch.	salle simu. anéch. et reverb.
He [151]	TFCT	0.6 s	CRNN à résidu	θ 360 classes	multi (nb inconnu)	mesures réelles	mesures réelles salle identique
Nguyen [152]	ITDs et ILDs	1 trame	CNN	θ, ϕ régression	1	mesures réelles signaux de bruit	mesures réelles salle identique
Perotin [153]	intensité acoustique FOA	0.8 s	CRNN	θ, ϕ 429 classes	1	42,900 salles simu.	salles simu. et réelles
Salvati [154]	SRPs normalisés	trame	CNN	coeffs. SRP	1	salle simu.	salles simu. et réelles
Sivasankaran [155]	CSIPD[1] et estimation des points temps- fréquence cibles	0.1 s	CNN	θ 181 classes	1 cible locuteur interférent	50 salles simu.	salles simu.
Suvorov [156]	signal temporel	1 trame	CNN à résidu (micros)	θ 20 classes	1	salle réelle	salle réelle différente
Perotin [157]	intensité acoustique FOA	0.8 s	CRNN	θ, ϕ 429 classes	multi (nb connu)	42,900 salles simu.	salles simu. et réelles

TABLEAU 3.1. – Comparaison des principaux systèmes de localisation utilisant des DNNs. Sauf indication contraire, les CNNs procèdent aux convolutions selon les axes de temps, de fréquence et les canaux. La mention de « salle » indique que le signal a été convolué à une RIR, tandis que « mesure » indique un enregistrement direct. [1]CSIPD : cosinus et sinus de la différence de phase (*cosine-sine inter-phase difference*)

3.3.4. L'apprentissage neuronal

Les réseaux de neurones profonds ont été appliqués avec succès à de nombreux problèmes en audio. Pour l'application qui nous intéresse, les assistants vocaux domestiques, ils ont été appliqués dans l'ordre inverse de la chaîne de traitement. D'abord utilisés pour la reconnaissance vocale (partie 3.1), puis le rehaussement de la parole (partie 3.2.4), ce n'est que récemment qu'ils ont été utilisés pour la localisation, en premier lieu grâce la robotique [142, 143, 151, 152]. L'utilisation de réseaux de neurones profonds permet de modéliser une réalité acoustique diverse, rendant les modèles robustes aux conditions acoustiques difficiles incluant du bruit et de la réverbération [142].

Architectures Diverses architectures ont été testées et sont résumées dans le Tableau 3.1. Les premiers réseaux utilisent simplement des couches FF [142, 146]. Ils ont par la suite été remplacés par des réseaux convolutifs. Parmi les variantes de ces approches, les réseaux résiduels permettent l'apprentissage de réseaux très profonds [151, 156] ; les convolutions dilatées augmentent la taille du champ récepteur du réseau sans augmenter le nombre de paramètres (de façon intéressante, Chakrabarty et Habets [149] les utilisent pour augmenter le nombre de microphones de l'antenne, et non pour prendre en compte un contexte temporel important) ; les réseaux de neurones convolutifs et récurrents (CRNN) [147, 148, 150, 153, 157] prennent en compte le caractère séquentiel des signaux audios. Aucun système ne s'appuie cependant sur des réseaux purement récurrents.
Les travaux de Comminiello [150] considèrent les échantillons temporels des signaux FOA comme des quaternions et utilisent une adaptation des couches convolutives à cette représentation dans un Q-CRNN (*quaternion convolutional and recurrent neural network*).

Paramétrisation des données De multiples paramétrisations des signaux sont utilisées dans la littérature. Elles dépendent du mode de capture, auquel succède en général un post-traitement explicite pour transformer les données brutes. Certains réseaux s'appuient simplement sur les méthodes classiques : ils utilisent en entrée les corrélations croisées généralisées entre les microphones [142, 146], les vecteurs propres des matrices de corrélation (comme MUSIC) [143], ou encore les résultats d'un SRP afin d'en optimiser la normalisation [154]. Grâce aux couches convolutives, d'autres sont capables d'extraire l'information de données de plus bas niveau, comme les spectrogrammes [147, 151] ou leur phase uniquement [144]. Certains travaux explorent même la possibilité d'utiliser des données temporelles brutes [150, 156]. Cependant, plus les données d'entrée sont de bas niveau, plus le modèle doit être complexe et plus le corpus d'apprentissage et la puissance de calcul doivent être importants. Il est donc (pour l'instant) intéressant d'utiliser des données d'entrée adaptées à la localisation, par exemple les indices binauraux ITD et ILD [152], les cosinus et sinus des différences de phase [155], ou les vecteurs d'intensité acoustique [153].

Espace des directions d'arrivée L'utilisation des réseaux de neurones ne résout pas pour autant le problème de la localisation. Pour assurer la généralisation, un maximum de situations doivent être rencontrées lors de l'apprentissage. En plus des conditions acoustiques variées précédemment citées, cela implique de rencontrer tout l'éventail des directions à estimer. Plus il est grand, plus la capacité de modélisation doit être importante. Pour cette raison, la plupart des techniques d'apprentissage supervisées (neuronales ou non) se limitent à estimer l'azimut, sans considérer la sphère unité complète — ce qui est de toute façon impossible géométriquement pour les antennes de microphones linéaires. Or, ajouter un degré de liberté complique significativement le problème. Les travaux d'Adavanne [147, 148] s'intéressent malgré tout à ce cas (voir Tableau 3.1). La scène y est captée au format FOA, dont l'isotropie est un avantage majeur.

Mode d'estimation Une question se pose lorsque l'on traite un problème par apprentissage automatique : doit-il être considéré comme un problème de classification ou de régression ? Étant donné le caractère ordonné et continu de l'azimut et de l'élévation, l'estimation des directions d'arrivée semble naturellement être un problème de régression. Le fait que l'azimut soit cyclique doit être pris en compte. Pour cela, il a été proposé d'estimer les coordonnées cartésiennes (x, y, z) sur la sphère unité correspondant à la direction (θ, ϕ) [148].
Cependant, comme le montre le Tableau 3.1, une large majorité des systèmes considèrent la localisation comme un problème de classification. Il a été raporté dans une note de bas de page dans l'article de Xiao et al. [142] que les résultats sont en effet meilleurs de cette façon. Dans ce cas, l'espace des directions d'arrivée est discrétisé en n_{DOA} classes. Puisqu'aucune relation d'ordre n'apporte d'information complémentaire, il est nécessaire d'avoir un grand nombre d'exemples d'apprentissage pour chaque classe [153]. Pour vérifier la généralisation, le système doit être testé y compris pour des directions qui sont pas exactement égales à celles représentant les classes. L'apprentissage doit donc également inclure des exemples de telles directions d'arrivées.
Afin de réintroduire une structure dans l'espace discret d'arrivée, au lieu d'un résultat binaire pour chaque classe, il a été proposé d'estimer la probabilité de chacune d'entres elles [151]. Les classes voisines de la classe réelle (au sens de la distance angulaire) se voient alors attribuer une probabilité non nulle, ce qui est censé aider l'apprentissage. Les résultats ne sont pas comparés à ceux obtenus en considérant une cible binaire. Nous avons effectué une expérience similaire, présentée dans la partie 4.5.4.
Enfin, traiter la localisation comme un problème de classification permet de s'attaquer à un cas important : la localisation simultanée de plusieurs sources.

Le problème des sources multiples La localisation de plusieurs sources simultanées soulève deux types de difficultés. D'une part, le problème est intrinsèquement plus difficile à résoudre : il est donc nécessaire d'utiliser des réseaux avec une capacité d'apprentissage suffisante et de les confronter à tous les cas de figure possibles lors de l'apprentissage. D'autre part, cela présente également un défi pour concevoir l'architecture du réseau : comment estimer plusieurs directions d'arrivée ?

Avec la régression, il serait nécessaire d'intégrer la connaissance du nombre de sources n_{src} dans l'architecture même du réseau, qui aurait alors $n_{src} \times 2$ sorties (pour l'azimut et l'élévation). L'ensemble des paires d'étiquettes possibles doit être ordonné afin d'éviter toute ambiguïté d'étiquette qui perturberait l'apprentissage (Figure 3.8a). Dans la version proposée par Adavanne et al. [148], la localisation est jointe à une tâche de classification du type d'évènement sonore. Une direction d'arrivée est alors estimée pour chaque type d'évènement, en utilisant $(x, y, z) = (0, 0, 0)$ lorsqu'un type n'est pas représenté. Cependant, il ne peut y avoir de cette façon plus d'une source localisée par type d'évènement. Cette méthode n'est donc pas adaptée à la situation multi-locuteurs.

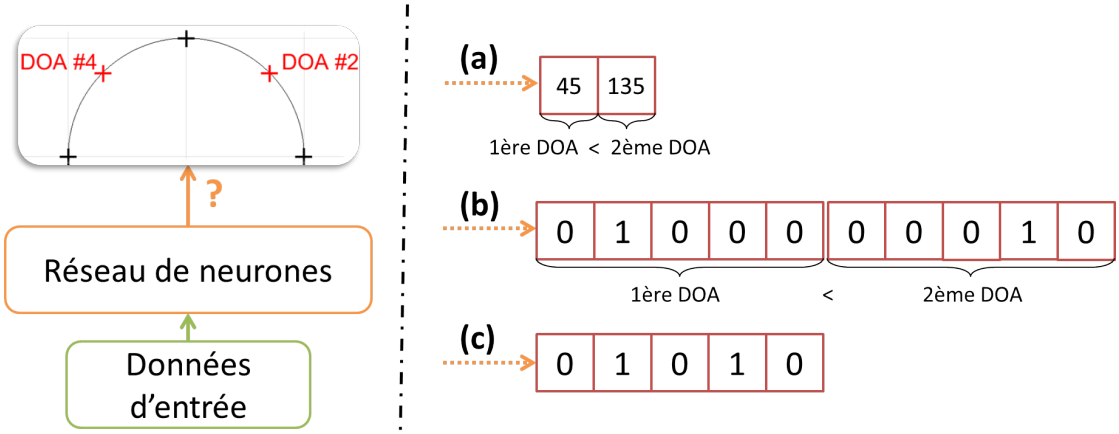


FIGURE 3.8. – Implémentations permettant l'estimation de plusieurs directions. Dans cet exemple, pour plus de lisibilité, seul l'azimut est estimé. Pour la classification, $n_{\text{DOA}} = 5$. (a) Estimation par régression, (b) par classification $n_{\text{DOA}} \times n_{src}$ neurones en sortie [143], (c) avec un nombre de classes fixe égal à n_{DOA}.

La plupart des systèmes de localisation adoptent une approche de classification. Connaissant le nombre de sources, Takeda et al. [143] proposent une architecture avec $n_{\text{DOA}} \times n_{src}$ sorties (Figure 3.8b), afin d'avoir exactement une distribution de probabilité par source. Il est alors nécessaire d'imposer un ordre dans les étiquettes pour éviter le problème d'ambiguïté d'étiquettes et d'apprendre un réseau pour chaque configuration $(n_{\text{DOA}}, n_{src})$.
Les autres systèmes utilisent un réseau avec n_{DOA} sorties (Figure 3.8c), où les sorties correspondant à la direction d'une source doivent être proche 1 et les autres proches de 0. Pour plusieurs sources, cela ne correspond plus à une distribution de probabilité ; il est donc impossible d'utiliser la non-linéarité *softmax* avant la dernière couche et la fonction de coût de corrélation croisée associée, qui permettent pourtant un apprentissage efficace. On utilise à la place la non-linéarité *sigmoïde*. Elle peut être associée à la fonction de coût aux moindres carrés MSE (*mean square error*) ou à l'entropie croisée binaire ou multiclasse. Lorsque le nombre de sources est connu, on peut sélectionner de manière adéquate les directions les plus vraisemblables parmi les estimations (à supposer que toutes les sources aient des directions différentes). Cette approche a un avantage majeur par rapport à celle de Takeda et al. [143] : un seul réseau est nécessaire quel que soit le

nombre de sources. Ce nombre n'étant utilisé qu'a posteriori, il peut par exemple être estimé par un système annexe et changer au cours du temps. Enfin, si le nombre de sources est inconnu, il est nécessaire d'estimer un hyper-paramètre de seuillage. Cela permet de sélectionner les pics prédominants dans les sorties du réseau, censés correspondre aux sources réelles, par opposition aux pics dus aux interférences et à la réverbération.

3.3.5. Résumé et positionnement

L'estimation de la direction d'arrivée de sources sonores est un problème complexe, étant donnée la variété des environnements acoustiques que l'on peut rencontrer dans le monde réel. Les techniques les plus efficaces pour traiter ce problème, encore une fois, s'appuient sur les réseaux de neurones. En particulier, des techniques ont été proposées pour traiter le cas bidimensionnel (recherche de l'azimut et de l'élévation). Certaines s'appliquent notamment aux signaux ambisoniques, qui comprennent une information spatiale presque explicite, et dont l'isotropie est adaptée à la localisation en deux dimensions.
Cependant, ces techniques ne sont pas complètement robustes aux conditions réelles, notamment en présence de sources directives ou de premières réflexions énergétiques due aux murs ou au mobilier. En combinant un réseau de neurones et l'information portée par le vecteur d'intensité acoustique, particulièrement facile à calculer pour des signaux ambisoniques, nous proposons dans le chapitre 4 un système de localisation robuste aux conditions citées ci-dessus.

3.4. Visualisation

3.4.1. De l'importance de l'éthologie neuronale

L'amélioration des performances des réseaux de neurones permet aujourd'hui leur utilisation en contextes réels. Il devient indispensable de comprendre leur comportement. Le règlement général sur la protection des données (RGPD) de l'Union européenne inclut même un « droit à l'explication » des décisions prises automatiquement par des algorithmes. Cependant, la nature de cette « explication » n'est pas clairement définie et la loi est difficilement applicable avec les connaissances actuelles [158]. « Comprendre » un réseau de neurones peut avoir de nombreuses significations [159] :

— l'explication algorithmique de la convergence de l'apprentissage vers une solution, son optimalité ou son unicité ;

— l'analyse du fonctionnement interne d'un réseau, la signification de chaque couche ou chaque paramètre [160] ;

— l'interprétation post-hoc des estimations d'un modèle entraîné [161].

Nous nous concentrerons sur cette dernière question. En particulier, une façon parlante d'interpréter les estimations d'un réseau est de faire une cartographie des données d'entrées pertinentes. Par analogie avec l'analyse d'image, nous parlerons de « pixels » pour désigner les différentes variables d'entrées du réseau. Un exemple de visualisation est présenté sur la Figure 3.9.

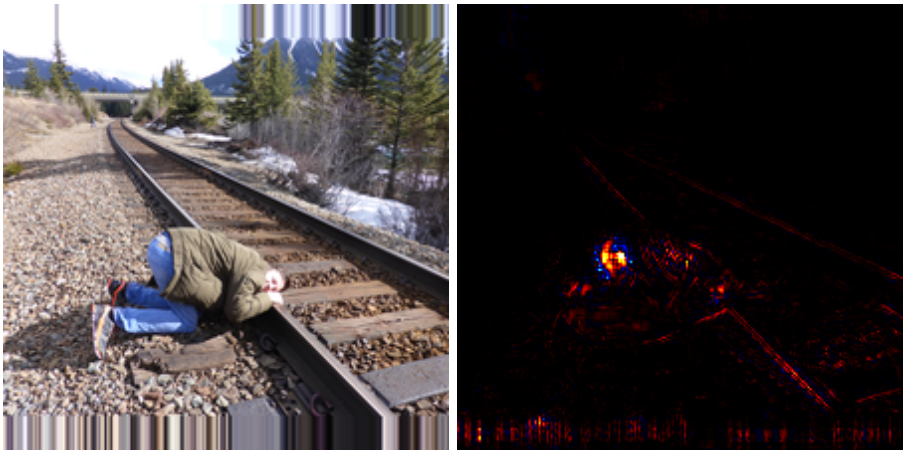


FIGURE 3.9. – Exemple de visualisation expliquant une estimation du réseau de reconnaissance d'images CaffeNet. Ici, la cartographie correspond à l'estimation « prayer rug ». Réalisé avec le site heatmapping.org.

De telles cartographies sont d'une importance majeure pour plusieurs raisons :

1. Elles facilitent la vérification de la capacité de généralisation du modèle, en faisant apparaître les biais potentiels dans les données d'apprentissage. Par exemple, dans [162], un algorithme de reconnaissance d'images est entraîné sur une base de données où toutes les occurrences de la classe « cheval » contiennent le copyright du photographe. La visualisation de la pertinence permet de mettre en évidence que le réseau utilise ce copyright, et non le cheval lui-même, pour estimer la classe. L'estimation est correcte, mais l'algorithme sera vraisemblablement incapable de généraliser à d'autres images de chevaux. Une autre pratique, plus couramment répandue, permettant de détecter ce type de biais (mais pas de l'expliquer) consiste à utiliser un ensemble de test réel et mesuré dans des conditions très différentes de l'ensemble d'apprentissage.

2. Éthiquement, ces analyses permettent de s'assurer que le modèle ne reproduira pas des comportements indésirables, fussent-ils représentatifs de l'ensemble d'apprentissage. Ce type de problème a été rencontré par Amazon avec leur système d'aide à l'embauche qui sélectionnait automatiquement les « meilleurs » CVs. En s'appuyant sur les exemples d'embauche effective des dix dernières années, celui-ci apprit que le genre était un critère plus discriminant que les diplômes ou l'expérience [1].

3. Enfin, les techniques de visualisation aident à mieux comprendre le problème, à découvrir des relations de causalité entre observations et estimations, et donc à trouver de nouvelles pistes de résolution dudit problème [163, 164]. Elles peuvent même être utilisées comme données d'apprentissage pour un réseau de neurones auxiliaire [165].

1. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

3.4.2. Principales techniques de visualisation

Un tour d'horizon des techniques de visualisation applicables aux réseaux de neurones est présenté dans les travaux de Montavon et al. [161, 166].

Maximisation de l'estimation Il est possible de générer l'entrée type la plus caractéristique, selon un réseau de classification, d'une classe donnée [167]. En fixant les poids du système, il suffit de procéder à une ascension de gradient sur les données d'entrée pour maximiser la sortie. La convergence vers un maximum absolu n'est pas garantie ; selon l'initialisation, différents maxima locaux peuvent être atteints. Le gradient peut être calculé de façon simple par rétro-propagation à travers le réseau, comme pendant l'apprentissage, mais cette fois les poids sont fixes et les activations et entrées sont variables. Les images synthétisées de cette façon ne semblent généralement pas naturelles, formées d'un amas chaotique d'arêtes et de motifs élémentaires. Il a donc été proposé d'ajouter un régulariseur, lui-même entraîné à favoriser les images réalistes [168], dans un principe similaire à celui des GANs [169]. Le réglage de ce régulariseur est délicat : s'il est trop sévère, les images générées seront représentatives de l'ensemble d'apprentissage, et non plus de la classe à optimiser.

Cette technique génère une image pour une classe donnée. Les suivantes, elles, sont conçues pour expliquer l'estimation du réseau pour un exemple d'entrée particulier : elles génèrent donc une carte différente pour chaque exemple d'entrée.

Analyse de sensibilité Les méthodes utilisant les dérivées partielles furent parmi les premières à être utilisées pour analyser le comportement des réseaux de neurones [170, 167]. L'algorithme Grad-CAM [171] s'appuie également sur des dérivées partielles locales mais s'avère plus discriminant vis-à-vis des différentes classes. Ces techniques peuvent être appliquées dès lors que la fonction modélisée par le réseau est localement dérivable. Les calculs peuvent également être faits par rétro-propagation. Cependant, l'analyse de sensibilité effectuée avec les dérivées partielles n'explique pas les estimations du système elles-mêmes ; elle révèle plutôt les pixels dont le changement ferait appartenir plus ou moins l'exemple donné à la classe estimée. Cela n'est pas toujours significatif. Samek et al. [166] donnent l'exemple d'une photo de rue avec des scooters : changer les pixels des scooters diminuerait l'appartenance de la photo à la classe « scooter », et changer les pixels de la route vide augmenterait l'appartenance à cette classe. Ces deux types de pixels sont mis en valeur de la même manière dans la visualisation produite, ce qui mène à des résultats peu interprétables.

Occlusion partielle de l'entrée Afin d'identifier les pixels d'un exemple qui sont les plus significatifs pour une estimation du réseau, il est possible d'appliquer des patchs annulant la valeur de l'entrée sur certaines zones. On mesure alors la différence d'estimation entre l'exemple original et l'exemple patché [172, 173]. Pour obtenir une cartographie complète,

il est nécessaire de reproduire l'expérience pour de nombreuses configurations, ce qui est rapidement coûteux.

Déconvolution Zeiler et al. [172] proposent une technique de déconvolution, qui vise à reconstruire l'entrée d'une couche convolutive à partir de sa sortie. Ils mettent au point un réseau complémentaire au CNN analysé, appelé DeconvNet, qui applique des opérations inverses à celles du CNN, dans l'ordre opposé. Ces opérations inverses n'ont été présentées que pour les réseaux convolutifs avec des couches de *max pooling* (mise en commun par maximum) et les fonctions de rectification linéaires (ReLU, *rectified linear unit*).

Propagation de la *relevance* L'analyse par propagation de la grandeur appelée *relevance* dans chaque couche (LRP, *layerwise relevance propagation*) n'est pas fondée sur des calculs de gradient, contrairement à toutes les méthodes précédemment citées (à l'exception de la déconvolution). Elle entre dans le cadre mathématique de la décomposition de Taylor [161]. Originellement, les règles de calcul de la LRP furent déterminées dans l'optique de respecter trois principes [173] :

- la visualisation produite doit être représentative de l'estimation du réseau, et donc dépendre des poids et activations liés à un exemple donné ;
- la propagation doit être conservative, c'est-à-dire que la somme de la *relevance* constante d'une couche à l'autre ;
- la *relevance* de la couche de sortie est fixée comme étant égale à l'activation du neurone correspondant à la classe à analyser.

Plusieurs règles de propagation vérifient ces principes et sont présentées plus en détail dans la partie 4.4. La LRP peut être appliquée à tout réseau dont les fonctions d'activation sont monotones.

3.4.3. Précautions d'utilisation

L'utilisation massive des réseaux de neurones étant relativement récente, le domaine de la visualisation explicative est encore ouvert. De nombreuses techniques existent, dont nous avons vu les principales, mais leur validité est généralement justifiée par des exemples ponctuels et des descriptions subjectives. La stabilité de ces méthodes n'est pas toujours vérifiée. Une visualisation peu représentative de l'intuition humaine du problème peut être due à une mauvaise technique d'analyse, mais aussi à un réseau non fiable, ou encore au fait que le problème inclut des phénomènes non appréhendés naturellement par un observateur humain. Il est donc fondamental de savoir si ces techniques sont valides, afin de pouvoir éliminer la première explication.

Qu'est-ce qu'une bonne visualisation ? Nous listons ci-dessous les critères invoqués dans la littérature [166, 174, 175] pour justifier la validité des techniques d'analyse. S'ils semblent nécessaires à une bonne visualisation, ils ne sont pas forcément suffisants, et d'autres apparaîtront probablement dans des travaux futurs.

— Invariance par rapport à l'implémentation [174] : si deux réseaux sont fonctionnellement équivalents, c'est-à-dire qu'ils retournent la même estimation pour n'importe quel exemple, alors les visualisations correspondantes doivent toujours être identiques. La LRP ne respecte pas ce principe.

— Sensibilité [174] : si la modification de certains pixels modifie l'estimation, leur importance est non nulle. Inversement, si la fonction implémentée par le réseau est indépendante d'une variable d'entrée, sa visualisation doit être neutre. Ce n'est pas le cas pour la déconvolution ni pour les méthodes de gradient.

— Explication continue [166] : des modifications mineures de l'entrée qui ne changent pas l'estimation doivent résulter en des changements mineurs de la cartographie. Cela n'est pas satisfait par les méthodes fondées sur des calculs de gradient.

— Invariance par rapport à l'entrée [175] : l'addition aux données d'entrée d'une constante qui ne change pas l'estimation ne doit pas changer la cartographie. Ce n'est pas toujours respecté par la LRP.

— Explication représentative de l'image [166] : si l'on applique à l'entrée une transformation qui ne change pas l'estimation, par exemple une rotation, la cartographie devrait mettre en valeur les mêmes caractéristiques (à la transformation près). Cela n'est pas respecté par la déconvolution. Cette propriété se rapproche de l'invariance par rapport à l'entrée ; pourtant, les travaux de Samek et al. [166] affirment que la LRP verifie cette formulation du critère d'invariance.

— Contributions positives et négatives [166] : la cartographie doit faire apparaître de façon distincte les zones qui appuient l'estimation, et celles qui contredisent l'estimation. C'est le cas de la LRP en utilisant des règles bien choisies, mais pas de la déconvolution et des méthodes utilisant le gradient.

— Explication globale [166] : la cartographie obtenue doit mesurer la pertinence absolue des pixels par rapport à une estimation. Ce n'est pas le cas de l'analyse de sensibilité qui fournit des informations locales, à savoir quels changements de pixels changeraient l'estimation.

De toutes les techniques présentées dans la partie 3.4.2, aucune ne respecte toutes ces conditions. Cela met en évidence la nécessité de poursuivre les efforts pour créer des techniques de visualisation robustes.

Évaluation quantitative Une étude propose une évaluation quantitative des différentes techniques de visualisation [166]. En s'inspirant des techniques d'occlusion partielle présentées dans la partie 3.4.2, il s'agit d'évaluer l'importance effective des zones mises en valeur par la visualisation. Pour cela, on perturbe les zones concernées, puis on mesure la modification de l'estimation du réseau qui en résulte. D'après cette étude, la LRP est plus pertinente que les autres méthodes.

3.4.4. Résumé et positionnement

Bien que l'application des réseaux de neurones soit aujourd'hui largement répandue, ce succès est assez récent et aucune méthode unique n'est plébiscitée pour vérifier le

comportement et la fiabilité de ceux-ci.
Mettre au point ces nouvelles techniques n'étant pas l'objet de cette thèse, nous avons choisi une technique existante, la LRP, malgré ses limitations. Elle s'est en effet montrée plus significative que la plupart des autres méthodes de visualisation sur un certain nombre de critères exposés dans cette partie.

4. Localisation de sources par réseau de neurones convolutif et récurrent

Ce chapitre est consacré à notre travail sur la localisation de sources sonores à partir d'un enregistrement ambisonique. Nous exposons la technique mise au point, qui s'appuie sur un CRNN. Nous présentons les performances dans des conditions variées : sur des exemples créés à partir de réponses impulsionnelles spatiales (SRIR, *spatial room impulse response)* simulées, de SRIRs réelles, ou encore sur des enregistrements réels en conditions domestiques. Nous analysons par LRP les comportements des CRNNs face à différents exemples. Enfin, nous étudions l'impact de la formulation du problème par classification ou régression, des paramètres choisis et des conditions d'apprentissage sur les performances du réseau.

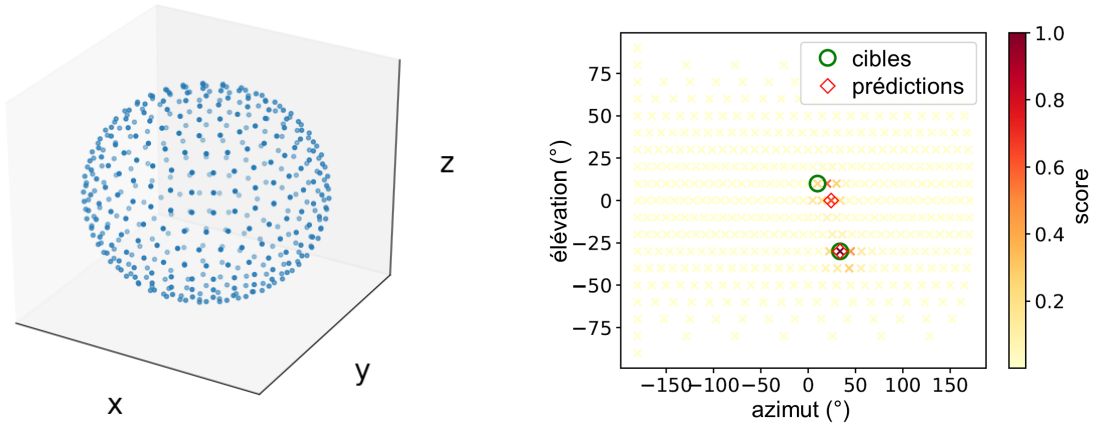
4.1. Solution proposée

4.1.1. Formulation du problème de classification

Nous avons évoqué dans la partie 3.3.4 les différentes façons de poser le problème de localisation de sources sonores rencontrées dans la littérature. Comme la plupart des autres travaux, nous considérons un problème de classification, c'est-à-dire qu'il s'agit d'estimer un score pour chaque classe dans un ensemble discret prédéfini. Cela permet d'avoir une architecture de réseau indépendante du nombre de sources à localiser. Afin d'obtenir les élévations $\phi \in [-90, 90]$ et les azimuts $\theta \in [-180, 180]$ des directions d'arrivée qui pourront être estimées par le réseau, nous utilisons la discrétisation quasi-uniforme de la sphère unité suivante :

$$\begin{cases} \phi_i & = -90 + \frac{i}{I} \times 180 \qquad \text{avec } i \in \{0, \dots, I\} \\ \theta_j^i & = -180 + \frac{j}{J^i+1} \times 360 \text{ avec } j \in \{0, \dots, J^i\}, \end{cases} \tag{4.1}$$

où $I = \lfloor \frac{180}{\alpha} \rfloor$, $J^i = \lfloor \frac{360}{\alpha} \cos \phi_i \rfloor$ et α est la résolution souhaitée, en degrés. Cette discrétisation est représentée sur la Figure 4.1a. Le réseau renvoie ensuite un score pour chaque direction possible, correspondant aux valeurs continues dans $[0, 1]$ attribuées à tous les points sur la Figure 4.1b.



(a) Sphère discrétisée

(b) Exemple d'estimation du réseau

FIGURE 4.1. – (a) Échantillonnage de la sphère unité afin de définir les classes possibles pour l'estimation des directions d'arrivée. (a) Exemple d'estimation du réseau sur cet échantillonnage.

4.1.2. Structure du réseau

Le réseau de neurones utilisé pour calculer les scores de chaque direction (θ_j^i, ϕ_i) est présenté sur la Figure 4.2. Il prend en entrée une séquence d'un nombre fixe de trames d'un vecteur FOA qui sera présenté par la suite. La première partie du réseau est constituée de trois blocs convolutifs qui permettent d'extraire des informations haut niveau à partir de paramètres d'entrée bas niveau. Chaque couche convolutive est suivie d'une normalisation par lot (*batch normalization*) qui permet d'accélérer l'apprentissage et de le rendre plus robuste à l'initialisation [176], puis d'un sous-échantillonnage par maximum (*max pooling*) selon l'axe fréquentiel. Dans un deuxième temps, deux couches récursives biLSTM prennent en compte l'aspect séquentiel du signal sonore. Enfin, l'application de deux couches FF permet d'estimer un score pour chaque classe.

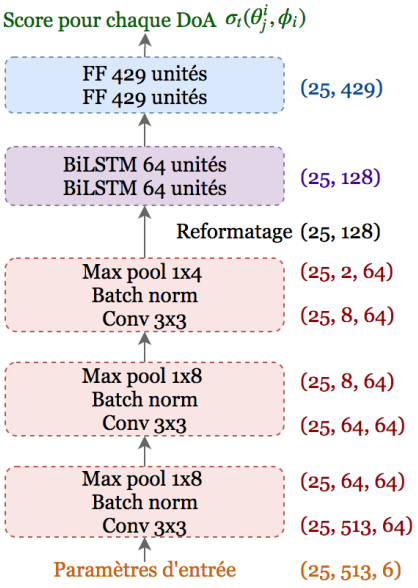


FIGURE 4.2. – Structure du réseau de neurones d'estimation de directions d'arrivée.

Les dimensions de sortie de chaque couche sont indiquées par les triplets (trames, bandes de fréquence, canaux). Les tailles des noyaux de convolution sont indiquées en trames x

bande de fréquence ; les convolutions sont également appliquées à tous les canaux. Cette architecture à la fois convolutive et récurrente a montré son efficacité pour extraire une information haut niveau d'une représentation temps-fréquence bas niveau, que ce soit la direction d'arrivée et le type d'évènement audio [148] ou le nombre de sources [164]. On verra dans la partie 4.5.3 que se limiter à une architecture purement convolutive ou purement récurrente fonctionne moins bien.
On suppose que les sources sont immobiles : leur direction d'arrivée est donc identique pour toutes les trames d'une séquence. Cependant, le réseau est conçu pour estimer les scores $\sigma_t(\theta^i_j, \phi_i)$ à chaque trame. Un post-traitement est nécessaire afin d'obtenir une estimation de la direction d'arrivée par séquence de trames. On effectue une moyenne temporelle des sorties du réseau sur la séquence afin d'obtenir un score global $\sigma(\theta^i_j, \phi_i)$ pour chaque classe. Ce score global est ensuite lissé en faisant la moyenne des scores dans un voisinage défini par une distance angulaire Δ :

$$\bar{\sigma}(\theta^i_j, \phi_i) = \frac{\sum_{i'j'} w_{iji'j'} \sigma(\theta^{i'}_{j'}, \phi_{i'})}{\sum_{i'j'} w_{iji'j'}}. \tag{4.2}$$

Les poids

$$w_{iji'j'} = \max\left\{0, 1 - \frac{\delta[(\theta^i_j, \phi_i), (\theta^{i'}_{j'}, \phi_{i'})]}{\Delta}\right\}. \tag{4.3}$$

décroissent linéairement avec la distance angulaire δ calculée de la façon suivante :

$$\begin{aligned}\delta[(\hat{\theta}, \hat{\phi}), (\theta, \phi)] = \arccos\{ &\sin(\hat{\phi})\sin(\phi) \\ &+ \cos(\hat{\phi})\cos(\phi)\cos(\hat{\theta} - \theta)\}.\end{aligned} \tag{4.4}$$

En supposant connus le nombre de sources I, les estimations finales sont obtenues en sélectionnant les I pics avec les scores lissés les plus importants. On considère qu'une classe est un pic si c'est un maximum sur son voisinage. L'étape de lissage permet de s'assurer que deux pics différents correspondent effectivement à deux sources différentes. Les losanges rouges sur la Figure 4.1b correspondent aux estimations finales sur une séquence. Si le nombre de sources n'était pas considéré connu, il serait possible de définir un seuil pour sélectionner les pics correspondant aux sources. Cependant, cela nécessite l'estimation d'un paramètre de seuillage, qui est généralement variable d'une salle à l'autre ou si le niveau sonore des sources varie.

4.1.3. Paramètres d'entrée du réseau

Contrairement à d'autres travaux qui fournissent simplement au réseau les phases et magnitudes des spectrogrammes TFTC des signaux FOA [147], nous proposons d'utiliser des paramètres d'entrée issus des parties active (2.36) et réactive (2.37) du vecteur d'intensité acoustique (voir partie 2.3.3), sous la forme suivante :

$$\frac{-1}{|W(t,f)|^2 + \frac{1}{3}(|X(t,f)|^2 + |Y(t,f)|^2 + |Z(t,f)|^2)} \begin{bmatrix}\mathbf{I}_\mathrm{a}(t,f) \\ \mathbf{I}_\mathrm{r}(t,f)\end{bmatrix}. \tag{4.5}$$

Un exemple est présenté en Annexe A. Le terme de normalisation permet de s'assurer que cette grandeur demeure bornée en chaque point temps-fréquence. Le vecteur d'intensité active donne au réseau une information sur la direction du flux d'énergie du champ acoustique. L'intensité réactive indique si les points temps-fréquence sont dominés par un signal direct venant d'une source prédominante, ou si plusieurs sources ou du bruit diffus sont présents [177]. Cependant, l'information portée par le vecteur d'intensité réactive est très bruitée, ce qui explique qu'elle n'ait jamais été exploitée en localisation jusqu'à présent. Nous étudierons son impact dans la partie 4.5.2. Le réseau utilisant cette paramétrisation sera appelé par la suite CRNN-Intensité.

4.2. Protocole expérimental

4.2.1. Paramètres audio

Tous les signaux sont échantillonnés à 16 kHz. La TFCT est calculée avec un fenêtrage de 1024 points, soit 64 ms, et un recouvrement de 50%. Une fenêtre sinusoïdale est appliquée lors de l'analyse et lors de la synthèse.

4.2.2. Paramètres d'apprentissage

La sphère est échantillonnée avec une résolution $\alpha = 10°$ dans (4.1). Cela correspond à $n_{\text{DOA}} = 429$ classes.
Les données sont présentées au réseau par séquences de 25 trames, avec 513 bandes de fréquence par trame et 6 canaux pour chaque point temps-fréquence (pour les différentes composantes du vecteur d'intensité). Un extrait de 1 s est donc scindé en 2 séquences de 25 trames (832 ms) avec 12 trames de recouvrement. Les dernières trames de la deuxième séquence sont fixées à 0. Les couches convolutives utilisent 64 filtres de taille 3×3. D'autres tailles sont testées dans la partie 4.5.3. L'opération de *max pooling* est effectuée uniquement selon l'axe fréquentiel, sur 8 bandes pour les deux premières couches, puis sur 4 bandes pour la dernière. Les deux couches biLSTM contiennent chacune 64 unités cachées, et les deux couches FF en contiennent 429 (soit le nombre de classes à estimer). La fonction d'activation utilisée pour les couches convolutives et la première couche FF est une ReLU. Pour les couches biLSTM, la fonction d'activation récurrente est une approximation linéaire par morceaux de la fonction sigmoïde (*hard sigmoid*), et la fonction d'activation globale est tanh. La dernière couche FF est suivie d'une sigmoïde, afin d'avoir une sortie comprise entre 0 et 1 pour chaque classe.
Pour l'apprentissage, on utilise l'optimiseur Nadam [178] avec un pas d'apprentissage initial de 10^{-3}. On applique une régularisation par abandon (*dropout*) de 30%. Le surapprentissage est évité en interrompant l'apprentissage lorsque la performance de classification cesse de s'améliorer sur l'ensemble de validation (voir ci-dessous), avec une patience de 20 époques. L'apprentissage nécessite environ 150 époques.

4.2.3. Ensembles d'apprentissage et de validation

SRIRs simulées pour l'apprentissage Pour l'apprentissage, il est nécessaire de constituer une base de données d'exemples audio qui permette au réseau de généraliser en conditions de test réelles. Deux approches sont possibles pour cela :

— Utiliser une base d'enregistrements réels. Le réseau est ainsi confronté pendant l'apprentissage à toute la complexité du problème, y compris des phénomènes qui ne sont pas toujours modélisés par des SRIRs simulées comme par exemple la diffraction ou la directivité des sources. Cependant, acquérir une base de données de taille conséquente est très coûteux en temps et en matériel, en particulier si l'on veut varier les conditions de mesure (lieu d'enregistrement, personnes effectuant les manipulations, sources audio, matériel de capture...). De plus, ces enregistrements sont soumis à une incertitude lors de l'acquisition de la vérité terrain, ce qui peut compromettre l'apprentissage.

— Utiliser une base constituée à partir de SRIRs simulées. La simulation utilise un modèle simplifié par rapport à la réalité, mais il est possible de générer un grand nombre de SRIRs avec des conditions variées pour toutes les directions d'arrivée possibles.

Une comparaison qualitative d'une SRIR réelle et d'une SRIR simulée est présentée sur la Figure 4.3. On observe que la SRIR simulée représente le champ direct et les premières réflexions avec des pics très nets, tandis que la SRIR réelle est beaucoup plus bruitée.
Nous avons choisi la dernière méthode en synthétisant une base de SRIRs selon la méthode présentée dans l'Algorithme 1. Nous utilisons la méthode image [135] implémentée dans le générateur de Habets [179], que nous avons modifié afin de générer des SRIRs selon l'encodage FOA parfait (2.25).
Puisque le but final est d'estimer les directions d'arrivée des sources présentes dans un mélange, il est nécessaire que tout l'espace des directions sur la sphère unité soit représenté dans la base de données. Pour cela, nous commençons par sélectionner aléatoirement une direction sur la sphère (appelée DoA_0 dans l'Algorithme 1) en utilisant un tirage quasi-uniforme[1]. Il est important de sélectionner DoA_0 en premier afin qu'aucune contrainte, comme par exemple la géométrie de la salle, ne puisse biaiser l'uniformité du tirage. Les caractéristiques de la salle sont choisies dans un deuxième temps : les dimensions horizontales sont tirées aléatoirement entre 2,5 et 10 mètres, la hauteur entre 2 et 3 mètres et le TR60 entre 200 et 800 millisecondes. L'antenne de microphones est positionnée à plus de 0,5 mètre des parois et la distance entre les sources et l'antenne est fixée entre 1 et 3 mètres. Enfin, dans un dernier temps, deux autres sources sont tirées aléatoirement dans la salle, ce qui permet de constituer des mélanges comprenant jusqu'à trois sources. Les directions ne sont donc pas uniformément représentées dans l'ensemble d'apprentissage, mais elles le sont sur au moins un tiers de l'ensemble, ce qui permet de s'assurer que toutes les directions possibles seront rencontrées un nombre significatif de fois. Les histogrammes récapitulant les tirages aléatoires pour l'ensemble des SRIRs d'apprentissage sont présentés en Annexe B.

1. http://mathworld.wolfram.com/SpherePointPicking.html

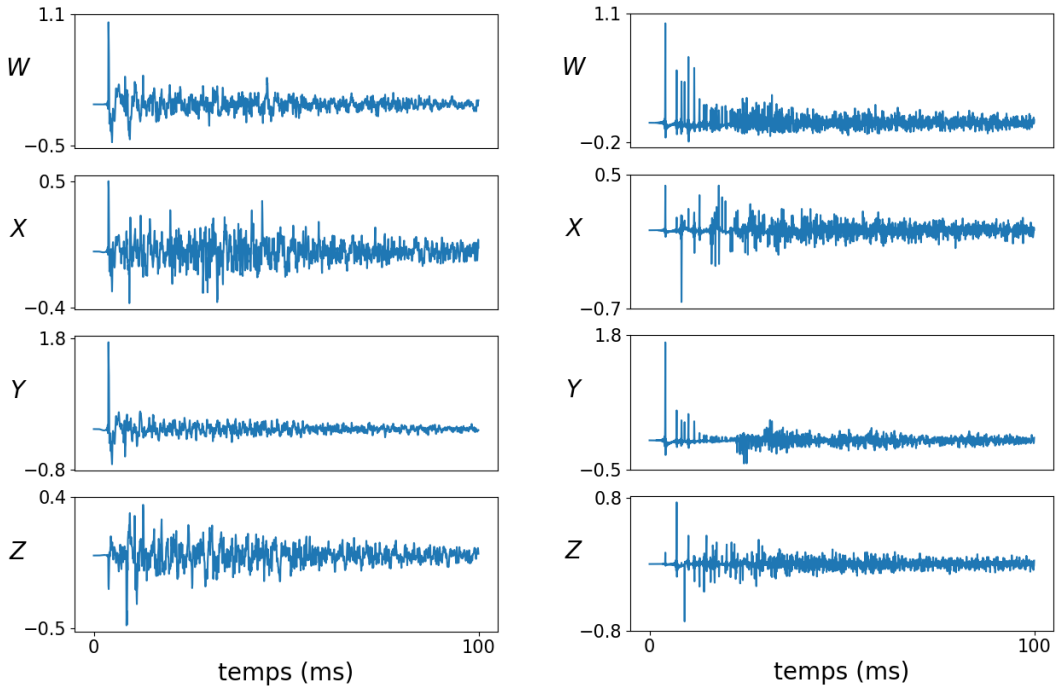


FIGURE 4.3. – Comparaison entre une SRIR FOA réelle (gauche) et simulée (droite). La SRIR réelle correspond à une source située à (71°, -5°) et la SRIR simulée à une source située à (77°, 5°). Les salles sont, de fait, différentes.

Au total, 42 900 salles sont générées, avec 3 SRIRs dans chaque salle, pour un total de 128 700 SRIRs. Les graines aléatoires sont imposées à chaque itération afin de s'assurer que toutes les salles et SRIRs sont différentes.

Signaux audio pour l'apprentissage L'ensemble d'apprentissage est constitué de deux sous-ensembles :

— Une source avec du bruit diffus : ce sous-ensemble contient des extraits audio longs de 1 s contenant une seule source ponctuelle et du bruit diffus. Les scènes sonores sont générées en convoluant chacune des 128 700 SRIRs de la base simulée avec un signal de parole. Le bruit diffus est synthétisé par convolution entre un bruit de foule et une SRIR diffuse. Pour chaque bruit diffus, on crée une SRIR diffuse en faisant la moyenne des parties diffuses de deux SRIRs piochées aléatoirement dans une base de SRIRs enregistrées dans une salle réverbérante. Le SNR entre la source ponctuelle et le bruit diffus est tiré aléatoirement entre 0 et 20 dB.

— Deux sources avec du bruit diffus : ici, tous les extraits audio contiennent deux sources et du bruit diffus. Chacune des SRIRs de la base est utilisée tour à tour pour générer l'image spatiale de la première source, tandis que la deuxième est générée avec une des deux autres SRIRs correspondant à la même configuration de salle, choisie au hasard. Les signaux de parole piochés pour chacune des deux sources

Algorithme 1 Protocole pour déterminer les paramètres des SRIRs simulées. δ est la distance angulaire (4.4).

pour toute direction DoA_0 tiré aléatoirement sur la sphère **:**
  répéter
    procédure SALLE
      $l = rand(2,5;10)$
      $L = rand(2,5;10)$ ▷ en mètres
      $h = rand(2;3)$
      $TR_{60} = rand(0,2;0,8)$ ▷ en secondes
    fin procédure

    procédure POSITION DU MICRO ET DE LA PREMIÈRE SOURCE
      $d_{mic-srcs} = rand(1;3)$ ▷ en mètres, pour toutes les sources
      $x_0, y_0, z_0 \in$ salle
      $x_{mic}, y_{mic}, z_{mic} \in$ salle ▷ à plus de 0,5 m des murs
      tels que : $\arg(x_{mic} - x_0, y_{mic} - y_0, z_{mic} - z_0) = DoA_0$
    fin procédure

    procédure POSITIONS DES AUTRES SOURCES
      Tirer aléatoirement $DoA_{1,2}$
        avec $\delta(DoA_i, DoA_j) > 10°$ pour tout $(i,j) \in \{0,1,2\}^2$
    fin procédure
  jusqu'à ce qu' une configuration plausible soit trouvée.
fin pour

durent 1 s et sont différents l'un de l'autre. Le rapport d'énergie entre la première et la deuxième source (SIR, *signal-to-interference ratio*) est tiré aléatoirement entre 0 et 10 dB. La première source est donc presque toujours prépondérante dans le mélange. Un bruit de foule diffus est ajouté avec un SNR de 20 dB par rapport à la première source.

Pour chacun des deux ensembles, on crée donc 128 700 mélanges de 1 s. Les signaux de parole sont extraits d'un sous-ensemble du corpus Bref [180]. Ce sous-ensemble contient 5 h de parole prononcée par 44 locuteurs différents. Les bruits de foule sont piochés aléatoirement parmi un ensemble de 33 minutes sélectionné manuellement sur Freesound [2].
Deux sous-ensembles de validation (à une et deux sources) permettant d'ajuster les hyperparamètres sont générés de la même façon, avec pour chaque ensemble 1 287 SRIRs différentes de celles de l'ensemble d'apprentissage mais synthétisées dans les mêmes conditions. Les locuteurs et bruits sont issus de Bref et Freesound mais n'ont pas été rencontrés à l'apprentissage.
Le réseau est d'abord appris avec l'ensemble à une source jusqu'à ce que les performances ne s'améliorent plus sur l'ensemble de validation à une source. L'apprentissage est ensuite

2. http://freesound.org

affiné sur l'ensemble à deux sources, avec une interruption de l'apprentissage lorsque les performances ne s'améliorent plus sur l'ensemble de validation à deux sources. Au total, 514 800 séquences sont utilisées pour l'apprentissage, pour un total de près de 115 h de signal.

4.2.4. Ensembles de test

Base de test à partir de SRIRs simulées Le réseau est tout d'abord testé sur deux ensembles créés de façon similaire aux ensembles d'apprentissage et de validation. Il utilise 1 287 nouvelles SRIRs simulées dans des conditions comparables. Les mélanges sont constitués respectivement d'une ou de deux sources. Pour les mélanges à une seule source, un bruit de foule diffus inconnu est ajouté de manière à ce que le SNR soit compris entre 0 et 20 dB. Pour les mélanges à deux sources, le SIR est tiré aléatoirement entre 0 et 10 dB et le SNR est fixé à 20 dB. La distance angulaire entre les deux sources est de 25° minimum. Les signaux de parole sont issus du corpus en anglais fourni pour le challenge SiSEC 2008 [181].

Base de test à partir de SRIRs réelles Le deuxième ensemble de test s'éloigne des conditions d'apprentissage. Les signaux audio sont générés avec le même protocole que l'ensemble de test précédent, à l'exception des SRIRs qui sont cette fois réelles. Elles ont été mesurées avec un Eigenmike positionné dans 36 positions différentes dans une salle moyennement réverbérante, avec un TR60 de 500 ms environ, mesurant 4 m sur 7 m pour une hauteur de 2,5 m. Pour chaque position de l'antenne, 16 haut-parleurs ont émis un sweep, permettant la collecte de 576 SRIRs. Les configurations sont présentées dans la Figure 4.4.

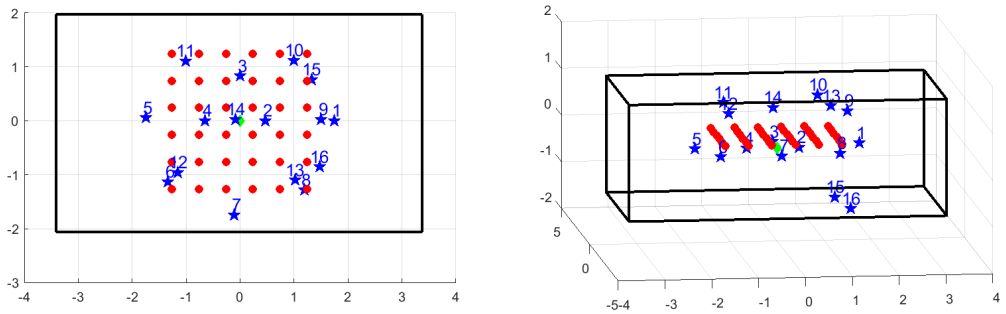


FIGURE 4.4. – Configuration d'enregistrement des SRIRs réelles. Gauche : vue du dessus. Droite : vue de côté. Les haut-parleurs sont représentés par des pentagrammes bleus, les positions de l'Eigenmike par des disques rouges. Tous les haut-parleurs sont dirigés vers le centre de la salle repéré par un losange vert, sauf le numéro 4 qui pointe vers le mur du bas (d'après la vue du dessus). Les dimensions sont données en mètres.

Base de test à partir d'enregistrements réels Afin de valider le système sur des données réelles, nous avons enregistré des signaux de parole dans un salon. L'Eigenmike est positionné au-dessus d'une table basse, comme le montre la Figure 4.5, donnant lieu à des premières réflexions très énergétiques. Deux locuteurs et une locutrice lisent des extraits du Petit Prince, debout ou assis dans des positions fixées, avec tout de même d'inévitables mouvements de tête. Au total, les locuteurs occupent 14 positions autour de la table. Pour chaque position, environ 5 minutes de lecture sont enregistrées, soit un total de 71 minutes d'enregistrement. Dans tous les enregistrements, des bruits ambiants non maîtrisés sont présents par intermittence : bruits de pas, de pages tournées, et même une tondeuse à gazon venant de l'extérieur. Lorsque ces bruits sont présents, le SNR se situe entre 5 et 10 dB. Les mélanges à deux sources sont générés en sommant deux enregistrements avec un SIR compris entre 0 et 10 dB.

FIGURE 4.5. – Configuration d'enregistrement réel dans un salon. L'Eigenmike utilisé pour la capture est entouré en rouge.

4.2.5. Algorithmes de référence

Nous comparons notre système à plusieurs algorithmes de référence décrits dans la partie 3.3 :

— L'algorithme VVM (*velocity vector module*) mis au point par l'équipe audio TPS (Traitement Parole et Son) d'Orange Labs pour la localisation ambisonique. Il a été présenté à LOCATA [127], où il surpasse les performances de MUSIC. Il s'appuie sur les directions apparentes en chaque point temps-fréquence si l'on considère un encodage FOA parfait pour une onde plane, tout en pondérant la contribution des points en fonction de la validité de cette hypothèse. Cet algorithme n'utilise donc pas d'apprentissage.

— Pour tester l'intérêt de la paramétrisation par le vecteur d'intensité, nous comparons notre système à un CRNN identique en tout point, à l'exception près qu'il est entraîné à déduire les directions d'arrivée à partir des 8 canaux formés par l'amplitude et la phase de la TFCT des signaux FOA. Ce CRNN est similaire à celui utilisé par Adavanne et al. [147]. Contrairement au cas du CRNN-Intensité, la

base d'apprentissage est ici centrée et l'écart-type normalisé sur toutes les trames, indépendamment pour chaque bande de fréquence et chaque canal. Les moyennes et écarts-type de l'ensemble d'apprentissage sont conservés pour normaliser les ensembles de test. Ce réseau sera appelé par la suite CRNN-FOA.

Ces algorithmes sont testés dans les mêmes conditions que le CRNN-Intensité proposé. L'apprentissage du CRNN-FOA se fait également sur les mêmes bases de données.

4.2.6. Métriques

Pour la détection de pics donnant lieu à l'estimation finale, on définit dans l'équation (4.3) un voisinage angulaire de rayon $\Delta = 2\alpha$, où α est la résolution angulaire de la grille dans (4.1). Les performances sont mesurées grâce aux indicateurs suivants :

- La moyenne et la médiane de l'erreur angulaire (4.4) en degrés.
- La précision, c'est-à-dire la proportion de sources et séquences estimées avec une erreur angulaire inférieure à 5, 10 ou 15°.
- La performance de classification, c'est-à-dire la proportion de sources et séquences associées à la meilleure classe possible (celle pour laquelle la distance angulaire par rapport à la cible est la plus faible). Avec la grille choisie, certaines directions d'arrivée ne peuvent être estimées avec moins de 7° d'erreur angulaire. La précision de classification permet donc de mesurer les performances des systèmes indépendamment de cet effet de grille, contrairement à la précision à 5°.

4.3. Résultats

4.3.1. Résultats pour une source

Le Tableau 4.1 permet de comparer les résultats des trois systèmes sur les ensembles de test à une source décrits dans la partie 4.2.4. Ces résultats peuvent également être visualisés sur les diagrammes en violon de la Figure 4.6.

SRIRs simulées Sur cet ensemble proche de l'ensemble d'apprentissage, on voit dans le Tableau 4.1a que les CRNNs surpassent largement le VVM, qui ne repose pas sur un apprentissage supervisé. Avec 58,5% de séquences classifiées correctement, le CRNN-Intensité est plus performant que le CRNN-FOA, qui en classifie correctement 54,8%. Les deux réseaux de neurones présentent très peu de résultats aberrants (c'est-à-dire anormalement mauvais), avec une précision à 15° supérieure à 95%.
Les Figures 4.7a et 4.7b montrent la répartition des erreurs angulaires de chaque système en fonction du SNR et du TR60. Pour tous les systèmes, un SNR plus faible favorise l'apparition de résultats aberrants. En revanche, si l'on observe les trois premiers quartiles, on constate que tous les systèmes sont robustes aux conditions de SNR difficiles. Concernant le TR60, les performances du VVM se détériorent régulièrement avec l'augmentation du temps de réverbération, tandis que les performances des CRNNs en sont indépendantes.

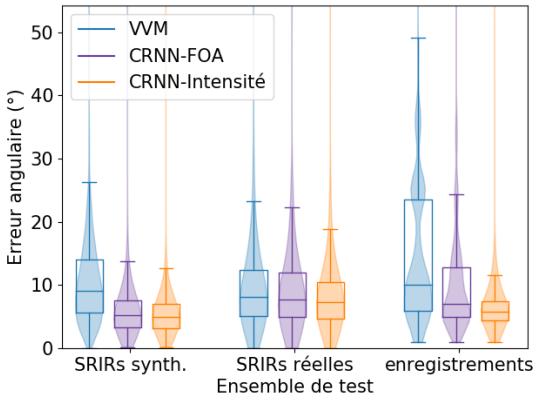
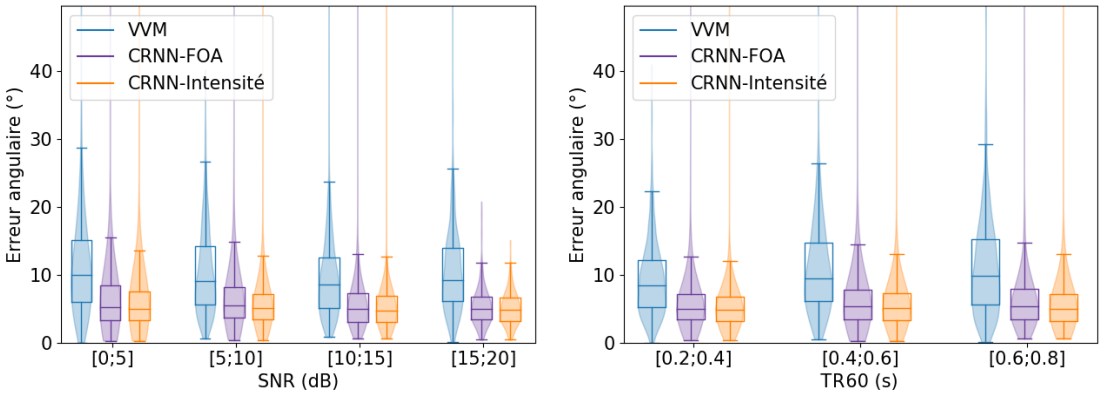


FIGURE 4.6. – Performances de localisation sur chaque ensemble de test contenant une seule source. Les boîtes montrent les premier et troisième quartiles, ainsi que la médiane. L'extrémité supérieure (respectivement inférieure) des moustaches correspond à la plus grande (respectivement la plus petite) valeur située à moins de 1,5 fois l'écart interquartile du quartile supérieur (respectivement inférieur).



(a) (b)

FIGURE 4.7. – Performances de localisation des différents systèmes sur l'ensemble de SRIRs simulées contenant une seule source (a) en fonction du SNR (b) en fonction du TR60.

SRIRs réelles Pour cet ensemble de test, les performances du VVM reportées dans le Tableau 4.1b sont comparables à celles du CRNN-FOA. Le CRNN-Intensité leur est supérieur, notamment en terme de précision fine, avec 28,0% des séquences localisées avec moins de 5° d'erreur, contre 24,9% pour le CRNN-FOA.
L'une des difficultés de cette salle est que les enceintes générant les sweeps sont directives, à la différence des sources simulées omnidirectionnelles vues lors de l'apprentissage par les CRNNs. Ici, l'antenne de microphones peut se situer derrière l'enceinte. On observe dans la Figure 4.8 que tous les systèmes ont des résultats satisfaisants lorsque le micro-

	Précision (%)				Err. ang. (°)	
Algo.	<5°	<10°	<15°	classif.	moy.	méd.
VVM [127]	20,6	55,4	78,9	24,7	10,6	9,1
CRNN-FOA	48,2	87,8	95,8	54,8	7,2	**5,2**
CRNN-Intensité	**51,2**	**93,3**	**98,1**	**58,5**	**6,2**	**4,9**

(a) SRIRs simulées

	Précision (%)				Err. ang. (°)	
Algorithme	<5°	<10°	<15°	classif.	moy.	méd.
VVM [127]	24,0	64,1	83,6	31,7	**10,5**	**8,1**
CRNN-FOA	24,9	66,4	85,3	31,3	**11,2**	**7,7**
CRNN-Intensité	**28,0**	**71,0**	**89,1**	**36,2**	**10,1**	**7,3**

(b) SRIRs réelles

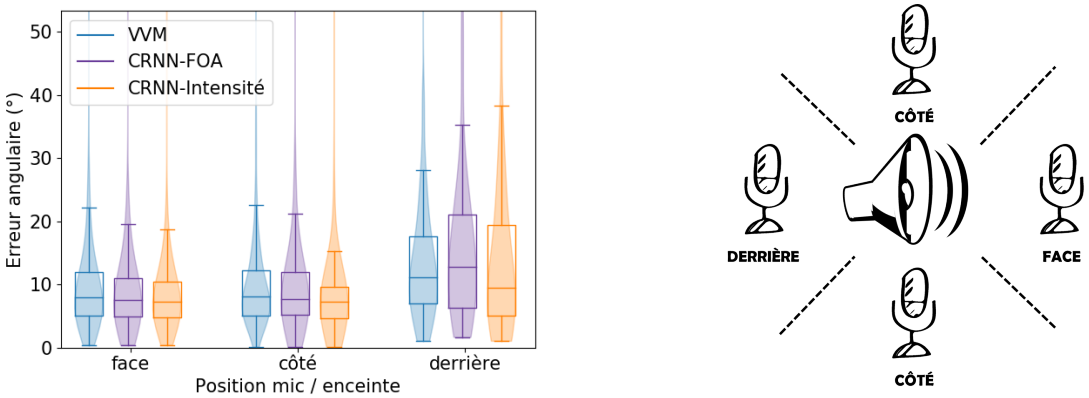
	Précision (%)				Err. ang. (°)	
Algorithme	<5°	<10°	<15°	classif.	moy.	méd.
VVM [127]	10,4	41,9	66,0	23,7	14,7	10,1
CRNN-FOA	22,3	60,3	81,0	36,3	11,3	7,1
CRNN-Intensité	**29,1**	**86,1**	**96,4**	**46,3**	**8,1**	**5,7**

(c) Enregistrements

TABLEAU 4.1. – Performances de localisation des algorithmes sur les différents ensembles de test contenant une seule source ponctuelle : (a) construit avec les SRIRs simulées, (b) construit avec les SRIRs réelles, (c) enregistrements réels. Les meilleures performances sont indiquées en gras. Lorsque plusieurs algorithmes présentent un résultat en gras, la différence entre ceux-ci n'est pas statistiquement significative. Les intervalles de confiance à 95% varient entre $\pm 0,4\%$ et $\pm 2,9\%$ pour la précision, et $\pm 0,8°$ et $\pm 1,7°$ pour l'erreur angulaire.

phone est situé en face ou sur le côté de l'enceinte. Les résultats se dégradent pour un microphone derrière l'enceinte, avec notamment beaucoup plus de résultats aberrants. On peut raisonnablement supposer que ceux-ci sont dûs au fait que le champ direct est faible voire inexistant par rapport à la réverbération ou aux premières réflexions qui sont dans cas très énergétiques. Le CRNN-Intensité reste plus performant au niveau des deux premiers quartiles.

Enregistrements Pour les enregistrements effectués dans un salon avec un microphone posé sur une table basse (voir Tableau 4.1c), le CRNN-Intensité surpasse le CRNN-FOA, avec une augmentation de la précision à 15° de 22,3 à 29,1%. Le CRNN-FOA est lui-même supérieur au VVM, qui ne classifie que 10,4% des séquences avec une précision inférieure à 5°. On observe sur la Figure 4.6 que le VVM généralise beaucoup moins bien



(a) (b)

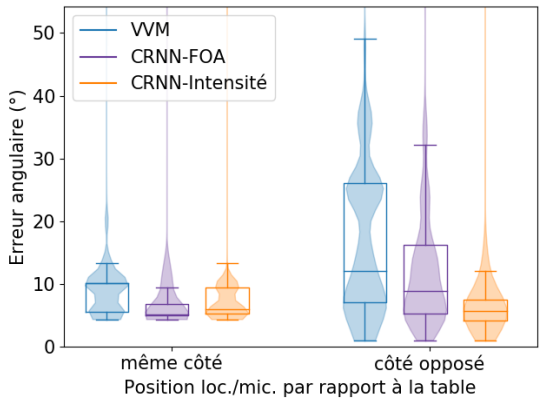
FIGURE 4.8. – (a) Performances de localisation des différents systèmes sur l'ensemble de SRIRs réelles contenant une seule source en fonction de l'orientation microphone/enceinte. (b) Séparation de l'espace entre les orientations « face », « côté » et « derrière ».

aux enregistrements réels que les CRNNs.
L'un des enjeux de cette situation est de distinguer le son direct de la première réflexion, précoce et énergétique, due à la table. Pour analyser les capacités des systèmes à faire cette distinction, on peut observer dans la Figure 4.9 la répartition des erreurs angulaires selon le côté de la table où a été émise la parole. Le microphone étant posé près d'un bord, cette première réflexion sera moins présente pour une locutrice située du même côté. Les performances du VVM et du CRNN-FOA se dégradent significativement lorsque, au contraire, le locuteur est de l'autre côté de la table, tandis que le CRNN-Intensité est robuste à cette situation. Les paramètres d'entrée dérivés du vecteur d'intensité acoustique lui permettent vraisemblablement de distinguer les sons directs des premières réflexions. On constate sur la Figure 4.9 que dans la situation où le locuteur et le microphone sont du même côté, l'erreur n'est jamais inférieure à 5°. Cela est un artefact dû aux positions des locuteurs et à la grille d'estimation : sur cet ensemble de test, 8 positions sur 14 sont telles qu'il est impossible d'avoir une erreur inférieure à 5°. C'est en particulier le cas de toutes les positions situées du même côté que le microphone.

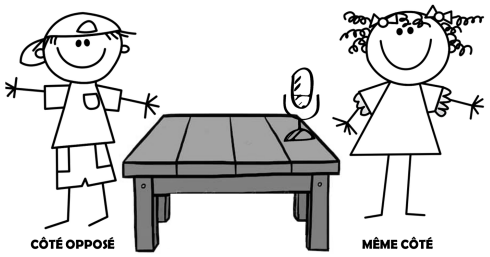
4.3.2. Résultats pour deux sources

Lorsque plusieurs sources sont présentes, on cherche à localiser chacune d'entre elle. Le Tableau 4.2 montre les résultats des trois systèmes sur les ensembles de test à deux sources décrits dans la partie 4.2.4. La Figure 4.10 représente la répartition des erreurs angulaires correspondantes.

SRIRs simulées Comme dans le cas avec une seule source ponctuelle, le Tableau 4.2a montre que les CRNNs sont beaucoup plus performants que le VVM sur les SRIRs simu-



(a)



(b)

FIGURE 4.9. – (a) Performances de localisation des différents systèmes sur les enregistrements réels contenant une seule source en fonction du côté de la table où elle est située. (b) Représentation de la situation où la source est du même côté de la table que le microphone et de la situation où elle est du côté opposé.

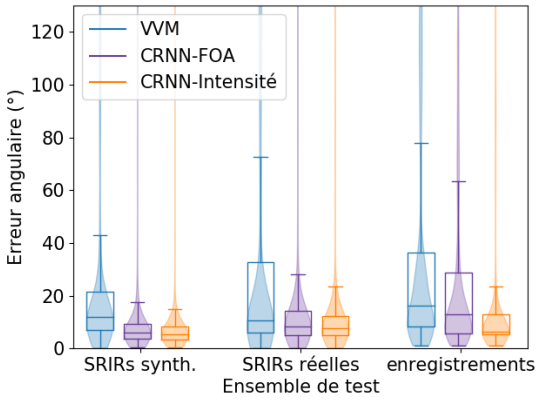


FIGURE 4.10. – Performances de localisation sur chaque ensemble de test contenant deux sources.

lées puisque leur apprentissage s'est déroulé dans des conditions similaires. Le CRNN-Intensité classifie correctement 50,8% de sources et séquences, tandis que le CRNN-FOA en classifie correctement 45,1% et le VVM 17,0%.
La Figure 4.11 montre la répartition des erreurs angulaires sur cet ensemble de test en fonction du SIR. Sans surprise, il est plus difficile pour tous les systèmes de localiser la source la moins énergétique. Les CRNNs y sont toutefois moins sensibles que le VVM.

SRIRs réelles Si les performances des CRNNs se dégradent sur les SRIRs réelles par rapport aux SRIRs simulées, elles restent significativement meilleures que celles du VVM

	Précision (%)				Err. ang. (°)	
Algorithme	<5°	<10°	<15°	classif.	moy.	méd.
VVM [127]	14,2	41,3	61,2	17,0	26,8	12,0
CRNN-FOA	39,1	77,7	88,0	45,1	11,6	**6,0**
CRNN-Intensité	**44,8**	**83,2**	**90,9**	**50,8**	**10,4**	**5,5**

(a) SRIRs simulées

	Précision (%)				Err. ang. (°)	
Algorithme	<5°	<10°	<15°	classif.	moy.	méd.
VVM [127]	18,4	48,2	61,2	22,3	34,0	10,6
CRNN-FOA	23,7	59,3	76,2	29,7	17,4	**8,4**
CRNN-Intensité	**27,6**	**64,8**	**81,7**	**34,8**	**14,9**	**7,7**

(b) SRIRs réelles

	Précision (%)				Err. ang. (°)	
Algorithme	<5°	<10°	<15°	classif.	moy.	méd.
VVM [127]	4,5	29,9	49,0	14,1	36,1	16,2
CRNN-FOA	13,1	40,3	57,9	23,3	27,8	12,8
CRNN-Intensité	**18,8**	**66,8**	**80,7**	**31,9**	**17,3**	**6,3**

(c) Enregistrements

TABLEAU 4.2. – Performances de localisation des algorithmes sur les différents ensembles de test contenant deux sources ponctuelles : (a) construit avec les SRIRs simulées, (b) construit avec les SRIRs réelles, (c) enregistrements réels. Les intervalles de confiance à 95% varient entre $\pm 0,8\%$ et $\pm 2,0\%$ pour la précision, et $\pm 0,5°$ et $\pm 1,5°$ pour l'erreur angulaire.

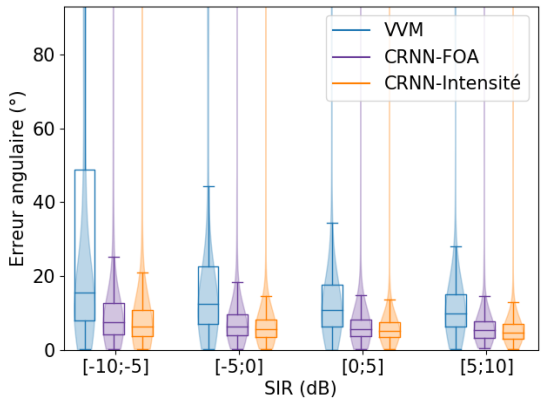


FIGURE 4.11. – Performances sur l'ensemble de SRIRs simulées contenant deux sources en fonction du SIR.

dont la précision à 15° vaut 61,2%, contre 76,2% pour le CRNN-FOA et de 81,7% pour

le CRNN-Intensité (Tableau 4.2b). D'autre part, les deux CRNNs ont une erreur angulaire médiane comparable, mais le CRNN-Intensité obtient de meilleures performances en moyenne avec 14,9° d'erreur moyenne contre 17,4° pour le CRNN-FOA. Cela prouve sa meilleure robustesse aux résultats aberrants.

Enregistrements Enfin, le CRNN-Intensité se révèle être le seul capable de généraliser aux enregistrements réels d'après les résultats du Tableau 4.2c. Le VVM localise moins de la moitié des sources avec moins de 15° d'erreur, tandis que le CRNN-FOA en localise correctement 58,0% et le CRNN-Intensité 81,0%, soit autant que pour les SRIRs réelles.

4.4. Analyse par *layerwise relevance propagation*

4.4.1. Présentation de la technique

Afin d'analyser le comportement des CRNNs, nous procédons à une analyse par LRP, une technique de visualisation qui indique la corrélation entre les paramètres d'entrée (que nous appellerons parfois « pixels ») et une sortie donnée du réseau, par le biais d'une grandeur que nous appellerons la *relevance*. Cette technique est rapidement présentée dans la partie 3.4.2 ; nous entrons ici dans les détails mathématiques.

Équations générales La LRP est conçue pour respecter trois contraintes [173] :

- la *relevance* est redistribuée depuis la couche supérieure (la couche de sortie) jusqu'à la couche inférieure (la couche d'entrée) en fonction des poids et des activations du réseau ;
- la rétropropagation est conservative, c'est-à-dire que la *relevance* totale est la même à chaque couche du réseau ;
- la *relevance* au niveau de la couche de sortie du réseau est égale à l'estimation du réseau.

Considérons le cas simple de deux couches FF linéaires successives i et j, illustré dans la Figure 4.12. Par construction du réseau de neurones, les activations dans la couche supérieure sont données par $z_j = \sum_i w_{ij} z_i + b_j$, où les z_i sont les activations dans la couche inférieure, les w_{ij} sont les poids entre les neurones et les b_j sont les biais. La *relevance* R_j correspondant au neurone d'activation z_j est redistribuée sur tous les neurones z_i de la couche inférieure avec les parts $R_{i \leftarrow j}$. Différentes formules peuvent être utilisées pour calculer $R_{i \leftarrow j}$; les principales sont exposées ci-dessous. Un neurone z_i de la couche inférieure reçoit des parts de *relevance* de tous les neurones de la couche supérieure auxquels il est connecté :

$$R_i = \sum_j R_{i \leftarrow j}. \tag{4.6}$$

La propriété de conservation impose que la somme de toutes les parts venant d'un neurone z_j de la couche supérieure soit égale à la *relevance* de ce neurone :

$$\sum_i R_{i \leftarrow j} = R_j. \tag{4.7}$$

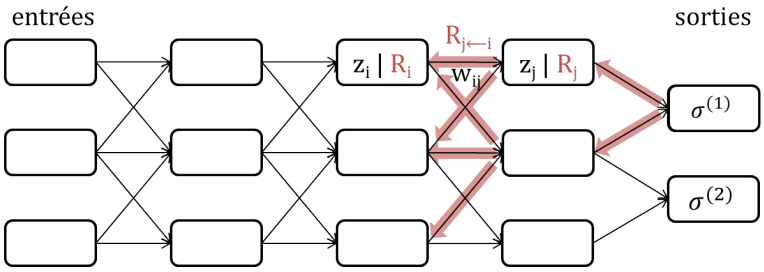


FIGURE 4.12. – Exemple de rétropopagation de LRP pour un réseau FF. La passe avant du réseau est représentée en noir, tandis que la rétropopagation de la *relevance* est en rouge. Les biais ne sont pas représentés pour plus de lisibilité. La *relevance* est rétropropagée depuis un des deux neurones de sortie du réseau.

La LRP a pour but de mettre en valeur les principaux chemins du réseau par lesquels l'information transite, rétropropageant ainsi la *relevance* jusqu'aux pixels d'entrée les plus significatifs. Ceci peut être effectué grâce à la règle de répartition suivante :

$$R_{i \leftarrow j} = \frac{w_{ij} z_i}{z_j} R_j. \tag{4.8}$$

Cependant, cette règle ne prend pas en compte les biais b_j, dans lesquels une partie de la *relevance* peut être bloquée. D'autre part, cette règle n'est pas stable lorsque le dénominateur est proche de zéro. Elle a donc été modifiée pour donner la règle dite « règle-ϵ » [182] :

$$R_{i \leftarrow j} = \frac{w_{ij} z_i + \frac{\epsilon \operatorname{sgn}(z_j) + b_j}{N}}{z_j + \epsilon \operatorname{sgn}(z_j)} R_j, \tag{4.9}$$

où ϵ est une petite constante positive qui fait office de régularisation, N est le nombre de neurones connectés à z_j dans la couche inférieure et la fonction sgn(.) renvoie le signe d'un nombre. À cause du facteur ϵ, la conservation de la *relevance* n'est pas parfaite. Si ϵ est trop grand, la rétropropagation de la *relevance* est altérée et n'a plus de signification. La « règle-$\alpha\beta$ » a été proposée comme alternative conservative et stable à la règle-ϵ [173, 161]. Elle traite séparément les activations positives et négatives :

$$R_{i \leftarrow j} = \left[\alpha \frac{(w_{ij} z_i)^+}{z_j^+} - \beta \frac{(w_{ij} z_i)^-}{z_j^-} \right] R_j \text{ avec}$$
$$z_j^+ = \sum_i (w_{ij} z_i)^+ + b_j^+ \text{ et } z_j^- = \sum_i (w_{ij} z_i)^- + b_j^-, \tag{4.10}$$

où $(.)^+$ et $(.)^-$ sont respectivement les parties positives et négatives d'un nombre réel. Les paramètres doivent être choisis tels que $\alpha = \beta + 1$.
Les formules ci-dessus s'appliquent aux couches linéaires, mais les neurones sont généralement suivis de fonctions d'activation $f(.)$ non-linéaires. Les formules restent valides à condition que cette fonction soit croissante. Les z_i dans les formules sont alors remplacées par les activations $y_i = f(z_i)$.
Ces règles ont initialement été conçues pour des couches FF mais restent valables pour les couches convolutives et de *pooling* [173]. Une adaptation aux couches LSTMs a également été proposée afin de gérer le mécanisme de crénelage [182]. Les activations dans la couche supérieure sont calculées lors de la propagation directe par $z_j = z_g z_s$, où z_g est une porte avec une activation comprise entre 0 et 1 et z_s est la source porteuse de l'information des couches inférieures ou précédentes. La règle de rétropopagation est alors simplement $R_g = 0$ et $R_s = R_j$. Il peut sembler que cette méthode ignore les valeurs de z_g et z_s, mais elles sont en réalité prises en compte dans R_j, qui dépend de z_j (par exemple selon l'équation (4.10) si la couche LSTM est suivie d'une couche FF).
Montavon et al. [161] montrent que la règle-$\alpha\beta$ pour une couche avec un biais négatif munie d'une activation ReLU équivaut à effectuer une décomposition de Taylor de la *relevance* en un point bien choisi. Ce résultat permet d'inclure la LRP dans un cadre mathématique rigoureux.

Bonnes pratiques empiriques Le formalisme ci-dessus ne permet toutefois pas d'assurer à la LRP une stabilité et une interprétabilité à toute épreuve. Un certain nombre d'astuces, exposées par Montavon et al. [161] ou découvertes lors de nos expérimentations, permettent de favoriser ces propriétés.

— Utiliser aussi peu de couches FF qu'il est possible sans compromettre la performance du réseau, et utiliser un *dropout* sur ces couches lors de l'apprentissage. Les couches FF ont tendance à diminuer l'interprétabilité de la *relevance* qui est rétropropagée. Le *dropout* permet de focaliser la *relevance* sur les neurones effectivement utiles.

— Utiliser les couches de *sum pooling* plutôt qu'un autre type de pooling, et de façon aussi abondante que possible. Ces couches sont intégrables au cadre de la décomposition de Taylor, et permettent de surcroît de concentrer le flux de *relevance* lors de la rétropropagation.

— Dans les couches FF ou convolutives, contraindre les biais à être négatifs ou nuls. Cela favorise la parcimonie des activations dans le réseau, ce qui est un moyen supplémentaire de canaliser la *relevance*.

— Par défaut, utiliser la règle-$\alpha\beta$ avec $\alpha = 1$ et $\beta = 0$ dans les couches cachées. Pour favoriser la présence de *relevance* négative, on peut choisir plutôt $\alpha = 2$ et $\beta = 1$.

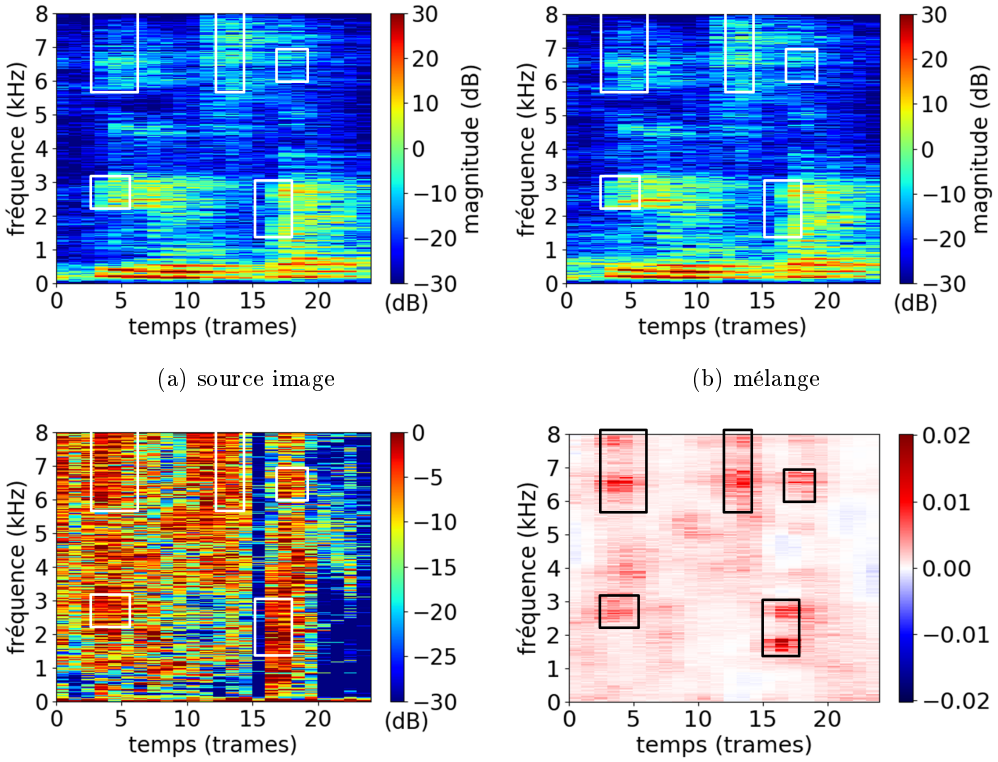
— Pour les couches FF, préférer la règle-ϵ à la règle-$\alpha\beta$ qui s'avère parfois instable.

Mise en pratique Par la suite, nous utilisons la règle-$\alpha\beta$ avec $\alpha = 1$ et $\beta = 0$ pour les couches BiLSTM et convolutives, et la règle-ϵ pour les couches FF. Le coefficient ϵ est fixé à 0,1, valeur qui s'est avérée stabiliser la rétropropagation tout en assurant une fuite

de *relevance* presque nulle. Nous avons également forcé les biais à être négatifs lors de l'apprentissage, sans impact sur les performances du réseau.
Nous adaptons la LRP au contexte d'estimation de la direction d'arrivée de la façon suivante. Pour une direction donnée, nous fixons la *relevance* de la dernière couche de chaque trame temporelle à la valeur du score $\sigma_t(\theta, \phi)$ pour la classe correspondant à la direction choisie, et à 0 pour les autres classes. La *relevance* est retropropagée séparément pour chaque trame. Les *relevances* sont ensuite sommées selon l'axe temporel, ainsi que selon les canaux, afin d'obtenir une unique carte temps-fréquence. Une *relevance* positive (respectivement négative) en un point temps-fréquence donné indique que la paramétrisation d'entrée en ce point porte une information utile (respectivement nuisible) à l'estimation.

4.4.2. LRP pour le CRNN-Intensité

Localisation d'une source Commençons par examiner la cartographie présentée sur la Figure 4.13 résultant de l'analyse par LRP du comportement du CRNN face à un cas simple : une seule source de parole dans une salle simulée similaire aux salles d'apprentissage, avec peu de bruit diffus (18 dB SNR). L'estimation du réseau est correcte pour ce signal. Les zones mises en valeur sur la Figure 4.13d correspondent à des points temps-fréquence où le rapport entre le champ direct de la source et le champ total incluant la réverbération et le bruit (DMR, *direct-to-mixture ratio*) est important. L'importance des points temps-fréquence correspondant au champ direct a déjà été utilisée pour la localisation de sources à travers différents indicateurs comme l'estimation du SNR [110, 113, 127], le ratio entre le son direct et l'écho [183], ou la cohérence interaurale [184]. On constate également sur le spectrogramme du signal incident (Figure 4.13b) que ces zones correspondent aux attaques des sons dans les moyennes et hautes fréquences. Cela correspond à un phénomène connu en psychoacoustique : l'effet de précédence [136] selon lequel les humains s'appuient fortement sur l'attaque des sons afin de les localiser.
L'Annexe C présente les cartographies de *relevance* canal par canal, à mettre en correspondance avec les six canaux du vecteur d'intensité active et réactive présentés à l'entrée du réseau (Annexe A). On constate que le canal Z est porteur de plus d'information que les autres canaux, ce qui est cohérent avec la direction d'arrivée de la source (139°, -61°). D'autre part, la *relevance* est certes moins importante sur la partie réactive du vecteur d'intensité que sur la partie active, mais elle tout de même significative. Nous étudions son importance dans la partie 4.5.2.
La Figure 4.14 montre l'analyse par LRP de l'estimation (correcte) du réseau pour le même signal de parole, mais dans une autre salle simulée et avec un bruit diffus à 0 dB SNR. On constate de nouveau que les points temps-fréquence les plus utilisés correspondent au DMR le plus élevé. Ici, puisque le bruit se concentre en basse fréquence, le réseau utilise principalement les hautes fréquences. Encore une fois, l'attaque des sons est particulièrement importante.
Étudions maintenant la situation de la Figure 4.15 où l'estimation du réseau est très éloignée de la véritable direction d'arrivée. Le signal de parole est le même que sur les exemples précédents, mais la salle simulée et le bruit sont différents. La *relevance* peut être rétropropagée soit à partir de la classe estimée (Figure 4.15d), soit à partir de la



(a) source image

(b) mélange

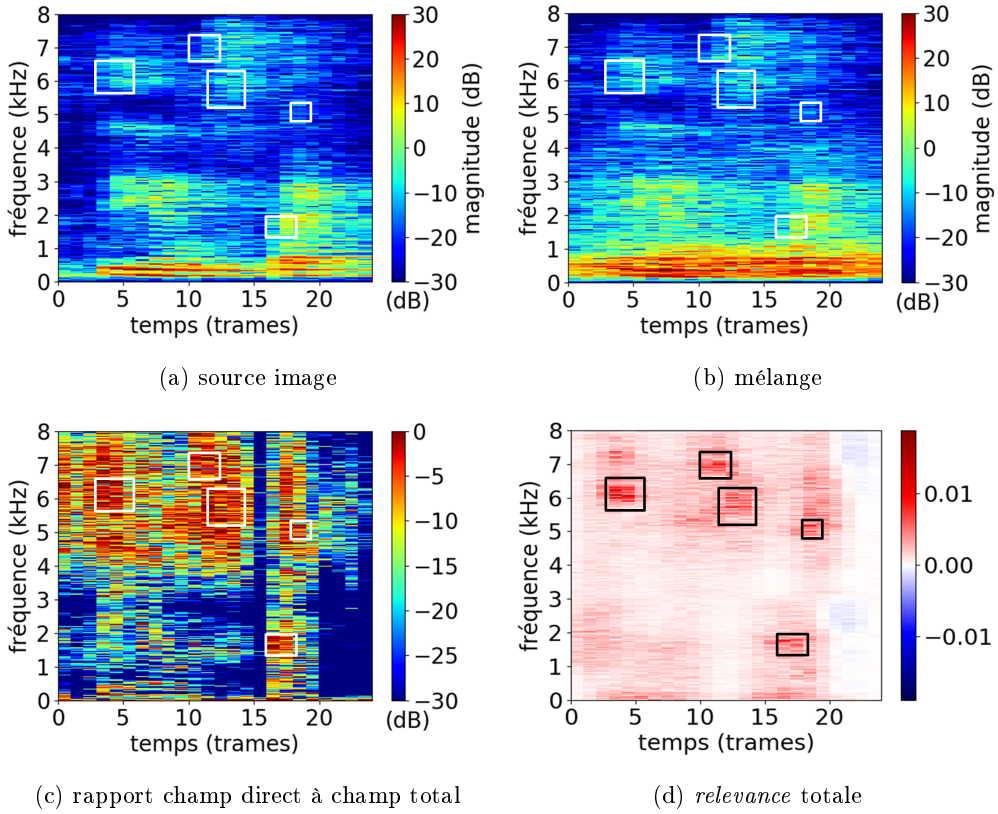
(c) rapport champ direct à champ total

(d) *relevance* totale

FIGURE 4.13. – LRP pour un signal contenant une source ponctuelle venant de (139°, -61°) et du bruit de foule diffus à 18 dB SNR, avec TR60 = 772 ms. Spectrogrammes (a) de l'image spatiale de la source (c) et du mélange. (c) Rapport champ direct à champ total. (d) *Relevance* sommée sur tous les canaux. Le détail des paramètres d'entrée (vecteur d'intensité actif et réactif pour chaque canal) et des *relevances* par canal est présenté en Annexes A et C.

classe cible (Figure 4.15e). Pour la classe estimée, le réseau s'appuie fortement sur les basses fréquences situées entre les trames 0 et 5 et entre les trames 15 et 25. Ces points correspondent à un DMR très bas, il n'est donc pas surprenant que l'estimation soit erronée. Les points temps-fréquence utiles pour estimer la classe cible (zones rouges sur la Figure 4.15e) plaident en défaveur de la classe estimée (zones bleues sur la Figure 4.15d). Cependant, leur poids ne suffit pas à compenser les points temps-fréquence favorables à la classe estimée : la somme de la *relevance* positive sur toutes les trames de la séquence vaut 25,7, tandis que celle de la *relevance* négative vaut -12,5. La *relevance* totale, égale par définition à la somme des sorties sur toute la séquence pour cette classe, vaut 13,2, contre 7,4 seulement pour la classe cible.

Localisation de deux sources La Figure 4.16 présente un résultat de LRP pour l'estimation du réseau étant donné un signal contenant deux sources ponctuelles. Dans ce cas,



(a) source image

(b) mélange

(c) rapport champ direct à champ total

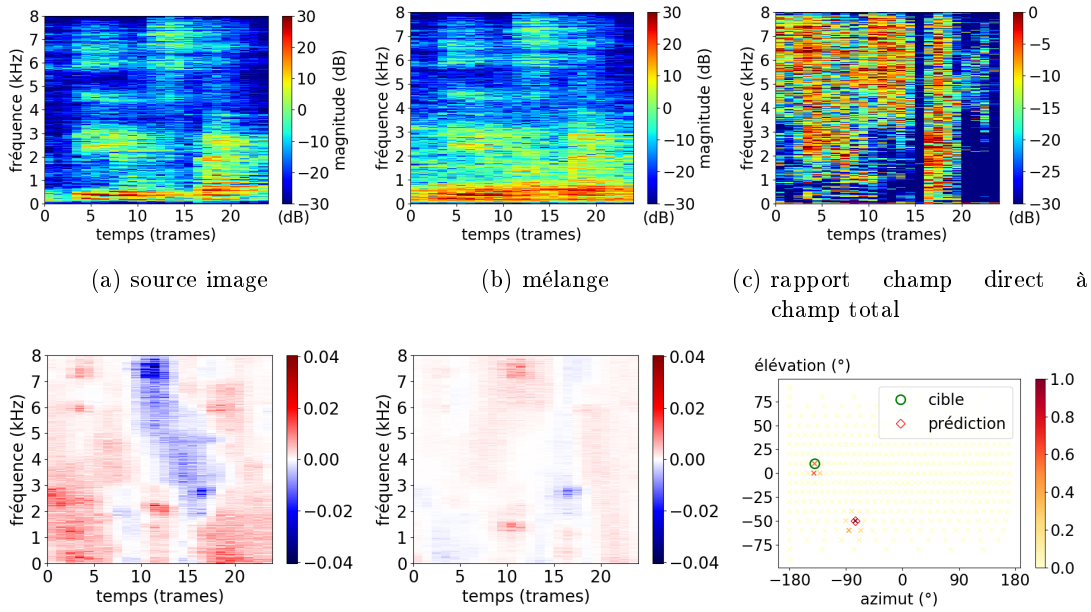
(d) *relevance* totale

FIGURE 4.14. – LRP pour un signal contenant une source ponctuelle venant de (115°, -68°) et du bruit de foule diffus à 0 dB SNR, avec TR60 = 750 ms. Spectrogrammes (a) de l'image spatiale de la source et (b) du mélange. (c) Rapport champ direct à champ total. (d) *Relevance* sommée sur tous les canaux.

l'estimation du réseau est correcte pour les deux sources. La *relevance* peut être retro-propagée à partir des classes correspondant à chacune des deux sources (Figures 4.16d et 4.16e). Les zones favorables à l'estimation de la première classe sont globalement défavorable à l'estimation de la seconde, et réciproquement. On constate sur les Figures 4.16a et 4.16b que la correspondance avec les attaques des sons est moins évidente que pour les scénarios à une seule source, en particulier pour la deuxième source. En effet, dans ce cas, le réseau doit aussi favoriser les zones temps-fréquence où la source correspondant à la classe considérée est prédominante par rapport à l'autre source.

4.4.3. LRP pour le CRNN-FOA

La Figure 4.17 présente les résultats de la LRP appliquée au CRNN-FOA pour le même signal que sur la Figure 4.14. Alors que le CRNN-Intensité s'appuie sur les zones où le DMR est important, qui correspondent ici aux attaques des sons en haute fréquence (où les interférences sont moins présentes), le CRNN-FOA privilégie les zones où le signal est



(a) source image (b) mélange (c) rapport champ direct à champ total

(d) *relevance* classe estimée (e) *relevance* classe cible (f) sorties du réseau

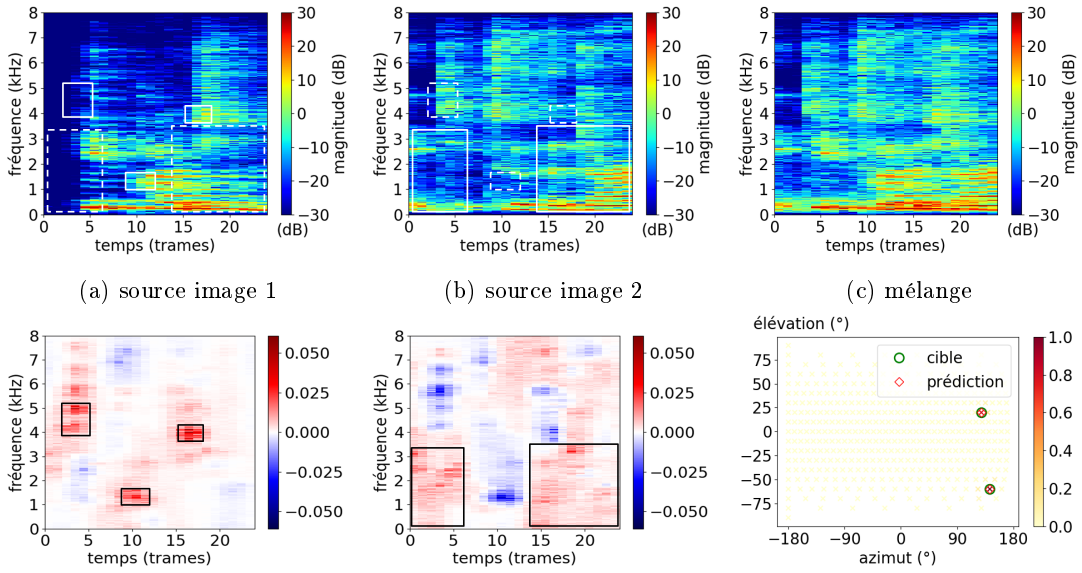
FIGURE 4.15. – LRP pour un signal contenant une source ponctuelle venant de (-141°, 8°) et du bruit de foule diffus à 2 dB SNR, avec TR60 = 688 ms. Spectrogrammes (a) de l'image spatiale de la source et (b) du mélange. (c) Rapport champ direct à champ total. (f) Estimation du réseau. (d) *Relevance* sommée sur tous les canaux pour la classe estimée et (e) pour la classe cible.

actif en général, sans lien avec le DMR (Figure 4.17c). Cela peut expliquer sa robustesse moindre au bruit et à la réverbération.

4.4.4. Limitations de la LRP

Les résultats précédents apportent un éclairage sur le fonctionnement des CRNNs utilisés pour la localisation. Cependant, ils doivent être considérés avec prudence. Les cartographies présentées correspondent à des exemples particuliers. Même si des comportements similaires ont été constatés sur la plupart des exemples, ils n'ont pas la légitimité d'une métrique générale qui permettrait par exemple de calculer la corrélation entre les attaques des sons et la *relevance*.

D'autre part, les résultats de LRP dépendent des règles choisies et des hyperparamètres α et ϵ, qui peuvent être différents pour chaque couche. Il n'existe pas de méthodologie générale pour les choisir, seule l'expérimentation permet de comparer différentes combinaisons. En particulier, nous avons également essayé de procéder à une LRP avec la règle-$\alpha\beta$ pour toutes les couches, y compris FF. Dans le cas multisources, les cartographies étaient alors les mêmes pour chacune des sources estimées, ce qui n'est pas cohérent avec le sens que doit porter la *relevance*.

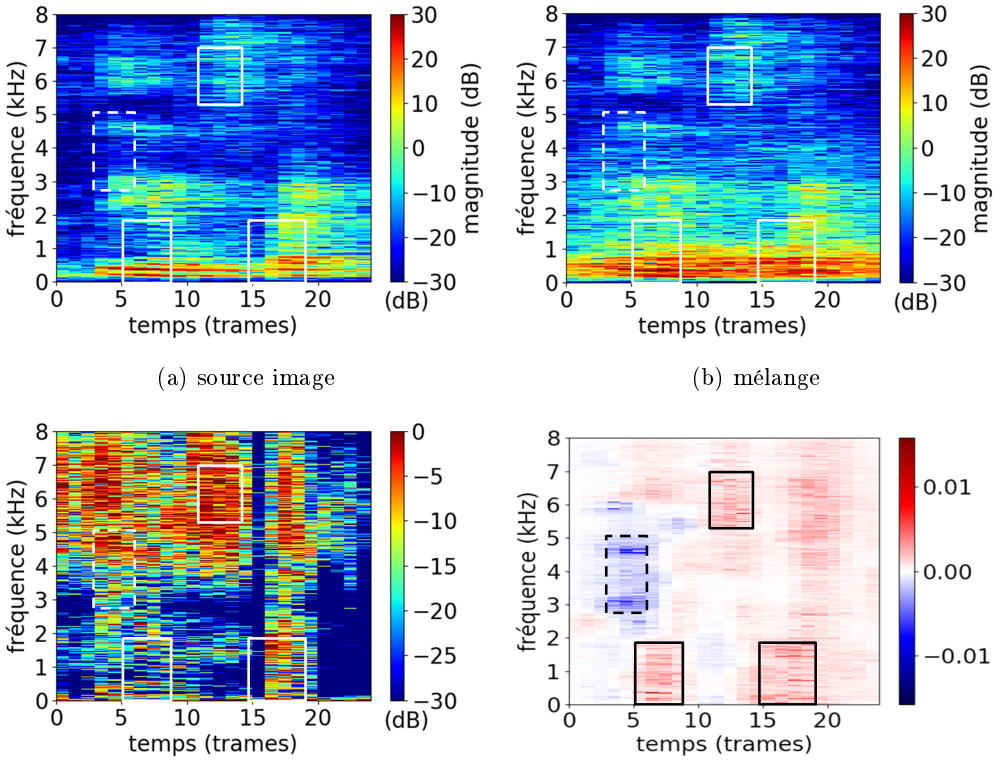


(a) source image 1 (b) source image 2 (c) mélange

(d) *relevance* classe 1 (e) *relevance* classe 2 (f) sorties du réseau

FIGURE 4.16. – LRP pour un signal contenant deux sources ponctuelles venant de (139°, -61°) et (134°, 18°), et du bruit de foule diffus à 20 dB SNR, avec TR60 = 772 ms. Le SIR entre la première et la deuxième source est de 3 dB. Spectrogrammes (a) de l'image spatiale de la première source, (b) de la deuxième source et (c) du mélange. (f) Estimations du réseau. (d) *Relevance* sommée sur tous les canaux pour la classe correspondant à la première source et (e) pour la classe correspondant à la deuxième source.

Enfin, la LRP est sensible au réseau utilisé. Pour deux réseaux appris dans des conditions proches et atteignant des performances équivalentes, les cartographies ne sont pas nécessairement semblables. Nous avons par exemple appris un modèle sur le même ensemble que celui présenté précédemment, mais où les signaux à une et deux sources étaient présentés dans un ordre aléatoire. Les performances sont très proches mais légèrement inférieures à celles du réseau présenté dans la partie 4.1.2, qui est d'abord appris sur l'ensemble à une source et affiné sur l'ensemble à deux sources. L'exemple de la Figure 4.18 montre les sorties des deux réseaux et la *relevance* pour le même signal que celui de la Figure 4.13. On observe que les cartographies des deux réseaux sont assez différentes, bien que certaines zones à *relevance* fortement positive concordent. Cette constatation est à rapprocher de l'affirmation de [174] selon laquelle la LRP ne respecte pas l'invariance d'implémentation, même si dans notre cas, les réseaux ne sont pas strictement équivalents. La différence de cartographie peut donc également être due à une différence de fonctionnement réelle des réseaux, et non à un défaut de la LRP.



(a) source image

(b) mélange

(c) rapport champ direct à champ total

(d) *relevance* totale

FIGURE 4.17. – LRP pour le CRNN-FOA avec le signal de la Figure 4.14. Spectrogrammes (a) de l'image spatiale de la source et (b) du mélange. (c) Rapport champ direct à champ total. (d) *Relevance* sommée sur tous les canaux.

4.5. Influence des paramètres sur l'apprentissage

Les résultats ci-dessus ont été obtenus avec le CRNN-Intensité canonique décrit dans la partie 4.1. Dans cette partie, nous étudions l'impact des paramètres d'apprentissage, de la base de données à la fonction de coût, en passant par l'architecture du réseau.

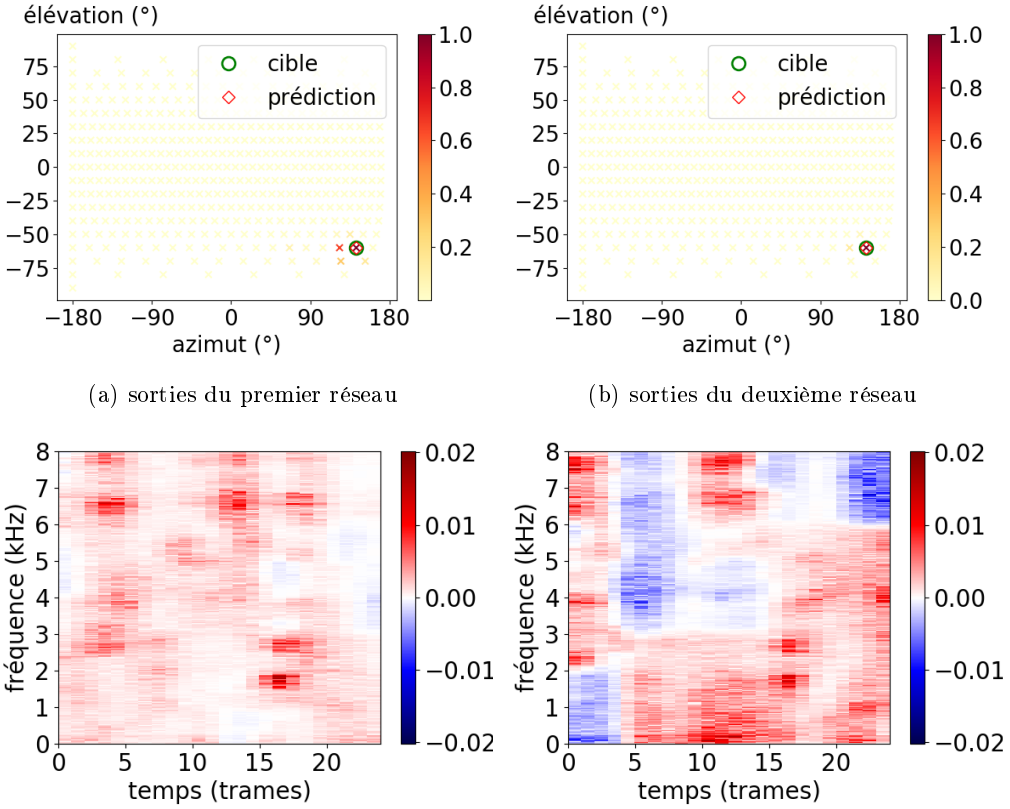
4.5.1. Base d'apprentissage

Nombre de sources dans les mélanges Nous comparons ici différentes façon d'apprendre aux réseaux à traiter des mélanges avec un nombre variable de sources.

— Un premier réseau est appris sur le sous-ensemble à une seule source créé à partir des SRIRs simulées, présenté dans la partie 4.2.3 (« appr. 1 src »).

— Un deuxième réseau est appris uniquement sur le sous-ensemble à deux sources (« appr. 2 src »).

— Le CRNN-Intensité utilisé jusqu'ici est d'abord appris sur le sous-ensemble à une



(a) sorties du premier réseau

(b) sorties du deuxième réseau

(c) *relevance* pour le premier réseau

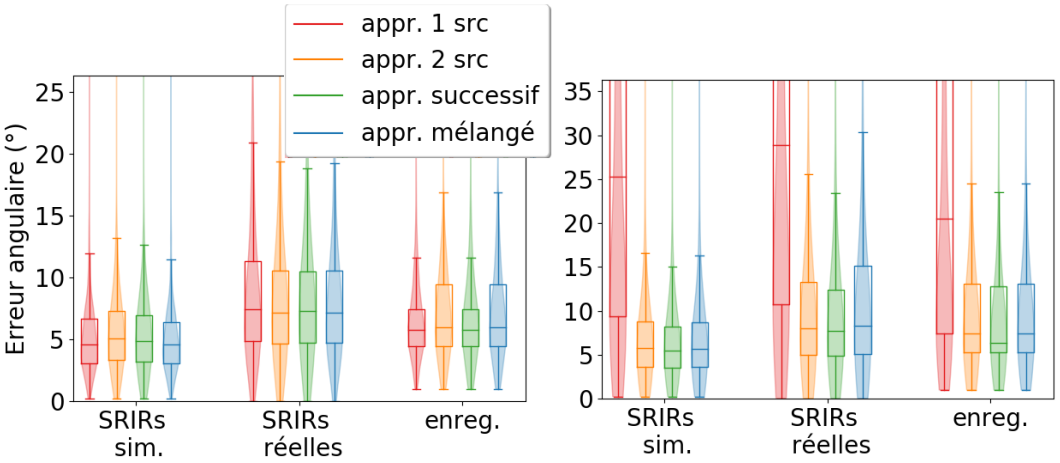
(d) *relevance* pour le deuxième réseau

FIGURE 4.18. – LRP pour un signal contenant une source ponctuelle venant de (139°, -61°) et du bruit de foule diffus à 18 dB SNR, avec TR60 = 772 ms, comme dans la Figure 4.13. (a) et (b) : estimations des deux réseaux. (c) et (d) *Relevances* pour la classe estimée pour les deux réseaux.

source, puis ajusté sur l'apprentissage à deux sources, selon le protocole présenté dans la partie 4.2.3 (« appr. successif »).

— Enfin, on considère le réseau appris sur l'ensemble complet, où les mélanges à une et deux sources sont présentés dans un ordre aléatoire, comme sur la Figure 4.18 (« appr. mélangé »).

Lors du post-traitement qui consiste à sélectionner les pics correspondant aux sources dans les estimations des réseaux, on considère le nombre de sources connu.
Sans surprise, la Figure 4.19 montre que le CRNN appris sur des mélanges à une seule source peine à généraliser à deux sources. Même la source prédominante n'est pas localisée correctement, puisque l'erreur médiane est beaucoup plus élevée que pour les autres réseaux. A l'inverse, le CRNN appris sur des mélanges à deux sources obtient des performances très satisfaisantes sur les mélanges à une seule source. Elles sont comparables à celles des CRNN entraînés sur tous les mélanges ; cependant, le CRNN d'abord appris



(a) Test sur une seule source

(b) Test sur deux sources

FIGURE 4.19. – Performances de localisation des CRNNs sur des mélanges de test à une (a) et deux (b) sources selon l'ensemble d'apprentissage : à une source, à deux sources, à une puis deux sources, ou enfin sur tous les exemples présentés dans un ordre aléatoire.

sur l'ensemble à une source puis ajusté sur l'ensemble à deux sources est le meilleur de tous, en particulier sur les enregistrements réels.

Diversité des SRIRs d'apprentissage Nous avons opté dans la partie 4.2.3 pour un apprentissage sur une base de données constituée à partir de SRIRs simulées grâce à la méthode image. Nous comparons ici les performances du réseau appris de cette manière avec un réseau appris grâce à des SRIRs réelles, ainsi qu'avec un réseau appris sur une base plus restreinte de SRIRs simulées. Dans cette expérience, toutes les bases d'apprentissage sont restreintes aux mélanges à une source. Nous considérons donc les trois bases suivantes :

— La base de SRIRs simulées complète est celle présentée dans la partie 4.2.3.

— La base de SRIRs simulées restreinte est créée de la même manière, mais seul un dixième des SRIRs est utilisé, soit 12 870 SRIRs. La diversité des salles et des directions d'arrivée est cependant préservée. Afin d'avoir un nombre de signaux audio comparable à celui de l'ensemble précédent malgré le nombre restreint de SRIRs, nous utilisons des extraits audio de 10 s au lieu d'1 s. La régularisation par *dropout* est fixée à 50%, ce qui s'est avéré être la valeur optimale.

— Les 576 SRIRs réelles présentées dans la partie 4.2.4 sont utilisées pour créer une base d'apprentissage similaire aux précédentes. Nous utilisons des extraits audio de 75 s pour obtenir la même quantité de données. Plusieurs taux de *dropout* sont testés, sans influence significative sur les performances.

Le réseau appris sur la base de SRIRs simulées entière atteint une précision de classification de 61% sur l'ensemble d'apprentissage et de 64% sur l'ensemble de validation. Sur la

base simulée réduite d'un facteur 10, le réseau atteint 63% sur l'ensemble d'apprentissage mais seulement 52% sur l'ensemble de validation. Cette dégradation des performances se ressent de la même manière sur les ensembles de test. Passer d'une moyenne de 300 SRIRs par classe à 30 change donc significativement la capacité du réseau à apprendre le lien entre la paramétrisation d'entrée et les directions d'arrivée. Enfin, le réseau appris sur les SRIRs réelles atteint 99% de précision de classification sur la base d'apprentissage au bout de 15 époques seulement, mais la précision de validation stagne à 10%. Le réseau sur-apprend sans être capable de généraliser. Les résultats restent très faibles en diminuant le nombre de paramètres du réseau ou en augmentant la régularisation. En conclusion, étant donnée la complexité du problème, il est important de simuler un grand nombre de SRIRs incluant une diversité importante pour apprendre le réseau.

4.5.2. Paramétrisation des données d'entrée

Normalisation Nous étudions dans ce paragraphe l'impact de différents modes de normalisation des données d'entrée.

- La normalisation par l'énergie correspond à l'équation (4.5). Elle est appliquée en chaque point temps-fréquence.
- La normalisation par rapport aux statistiques d'apprentissage consiste à calculer les moyennes et écarts-type de chaque bande de fréquence et canal sur toutes les trames de la base d'apprentissage, puis de centrer et réduire les entrées du réseau pour chaque trame.
- Les moyennes et écarts-types de la base d'apprentissage peuvent également être calculés sur l'ensemble des trames et canaux, indépendamment pour chaque bande de fréquence. Contrairement à l'option précédente, cela permet de garder les différences de niveaux entre les canaux qui contiennent une information sur la direction d'arrivée. Dans la Figure 4.20, nous appelons cette option « stats appr. / canal ».

Les normalisations dépendant des moyennes et écart-types présentent l'inconvénient d'être dépendantes de l'ensemble d'apprentissage. Si l'ensemble de test s'en éloigne, le changement d'échelle ne permet pas de centrer et normaliser les données. Idéalement, il faudrait avoir accès aux moyennes et écarts-type de l'ensemble de test complet, ce qui n'est pas réalisable avec une latence réduite. Nous ajoutons donc une dernière option.

- La normalisation par séquence, similaire à la normalisation précédente, mais où l'ensemble de test est centré et normalisé sur toutes les trames de la séquence en cours, indépendamment pour toutes les bandes de fréquence et tous les canaux.

Pour ces expériences, les modèles sont appris et testés sur les sous-ensembles à une seule source.
Dans tous les scénarios de test, la normalisation la moins robuste est celle par séquence. Le contexte est probablement trop court pour que la paramétrisation qui en résulte soit homogène entre les différentes séquences. La normalisation par l'énergie fournit les meilleurs résultats quelque soit l'ensemble de test, dépassant les performances des normalisations selon les statistiques des ensembles d'apprentissage communément utilisées dans le cadre des réseaux de neurones.

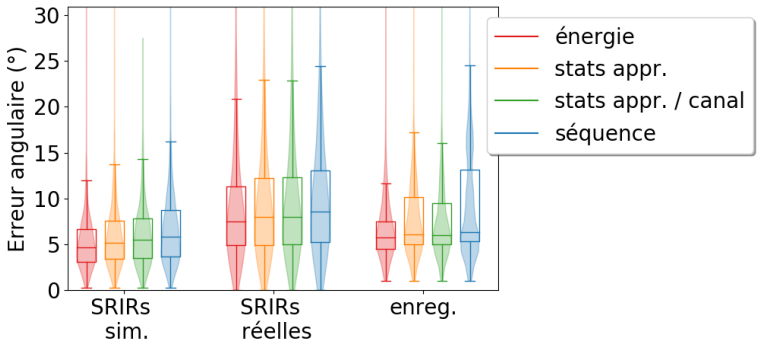


FIGURE 4.20. – Performances de localisation du CRNN-Intensité appris avec différentes normalisations de la paramétrisation d'entrée, sur les ensembles de test à une seule source.

Intérêt de l'intensité réactive Nous évaluons l'intérêt de fournir au CRNN la partie réactive du vecteur d'intensité acoustique en comparant deux modèles :

— Le CRNN-Intensité présenté dans la partie 4.2 reçoit en entrée les parties active et réactive du vecteur d'intensité acoustique.

— un CRNN-Intensité-active qui ne prend en compte que la partie active du vecteur d'intensité acoustique.

Ces deux réseaux sont d'abord appris sur le sous-ensemble à une source créé à partir des SRIRs simulées, puis l'apprentissage est ajusté sur le sous-ensemble à deux sources (voir partie 4.2.3).
Les résultats sur les ensembles de test à une source sont présentés dans le Tableau 4.3. On constate que le gain dû à la prise en compte de la partie réactive du vecteur d'intensité est important pour les SRIRs simulées. Dans la plupart d'entre elles, il y a en effet une réverbération importante ou bien la source ou le microphone sont proches des murs, donnant naissance à des réflexions qui peuvent être identifiées par le vecteur d'intensité réactive. De même, pour les SRIRs réelles, l'erreur à 15° est diminuée de 40% en relatif lorsque l'on utilise l'intensité réactive. Enfin, pour les enregistrements réels où le micro est posé sur une table basse, connaître la partie réactive est un avantage crucial qui permet de réduire l'erreur à 15° d'un facteur 5, passant d'une précision de 80,7% à 96,4%. Dans cette situation en effet, la réflexion due à la table basse peut induire une confusion qui génère une erreur de localisation ; mais les points temps-fréquence dominés par cette réflexion sont probablement identifiés par le CRNN-Intensité grâce à l'intensité réactive.
Sur les mélanges à deux sources présentés dans le Tableau 4.4, l'utilisation de l'intensité réactive apporte une amélioration significative pour tous les ensembles de test. Cela laisse supposer que l'intensité réactive permet vraisemblablement au réseau de différencier les points temps-fréquence ne contenant qu'une source (et donc porteurs d'information) de ceux contenant deux sources ou contenant essentiellement des réflexions. Cette hypothèse mériterait toutefois de faire l'objet d'une analyse théorique complète.

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN-Intensité-active	38,7	82,5	94,6	7,8	6,1
CRNN-Intensité	**51,2**	**93,3**	**98,1**	**6,2**	**4,9**

(a) SRIRs simulées

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN-Intensité-active	**27,6**	64,8	81,7	14,9	**7,7**
CRNN-Intensité	**28,0**	**71,0**	**89,1**	**10,1**	**7,3**

(b) SRIRs réelles

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN-Intensité-active	18,8	66,8	80,7	17,3	**6,3**
CRNN-Intensité	**29,1**	**86,1**	**96,4**	**8,1**	**5,7**

(c) Enregistrements

TABLEAU 4.3. – Précisions et erreurs angulaires sur des mélanges à une seule source des CRNNs de localisation prenant en compte les parties active et réactive du vecteur d'intensité (CRNN-Intensité), ou uniquement la partie active (CRNN-Intensité-active). Les intervalles de confiance à 95% varient entre 0,5 et 2,6% pour la précision et 0,6 et 1,7° pour l'erreur angulaire.

4.5.3. Structure du réseau

Nous présentons dans cette partie une comparaison de différentes architectures de réseau de neurones.

Prise en compte des contextes temporel et fréquentiel Nous testons plusieurs noyaux de convolution en plus du noyau 3x3 utilisé pour le CRNN-Intensité proposé, afin de mettre en valeur l'importance de la dimension fréquentielle et de la dimension temporelle de l'analyse. Nous testons également un CNN avec des filtres 3x3 mais sans couches récursives. Enfin, nous avons testé des réseaux purement biLSTMs, sans succès : l'apprentissage est interrompu en raison des performances dégradées sur l'ensemble de validation avant que le réseau n'atteigne des performances correctes sur l'ensemble d'apprentissage. Pour ces expériences, les réseaux sont appris et testés sur des mélanges à une source uniquement. Les résultats sont présentés sur la Figure 4.21. Les tailles des noyaux de convolution ont finalement un impact assez limité sur les performances. On constate globalement que l'aspect temporel n'est pas crucial, puisque le CRNN 1x3 est aussi performant, voire légèrement meilleur, que le CRNN 3x3. Le CNN 3x3, qui ne contient aucune couche récursive, obtient des performances légèrement moins bonnes, mais tout de même très correctes. La prise en compte des rapports entre les bandes de fréquence est plus important : le CRNN 3x1 obtient plus de résultats aberrants sur l'ensemble de

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN-Intensité-active	35,5	74,6	87,0	12,8	6,6
CRNN-Intensité	**44,8**	**83,2**	**90,9**	**10,4**	**5,5**

(a) SRIRs simulées

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN-Intensité-active	22,6	57,9	75,2	18,3	**8,5**
CRNN-Intensité	**27,6**	**64,8**	**81,7**	**14,9**	**7,7**

(b) SRIRs réelles

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN-Intensité-active	15,0	50,7	65,0	21,1	9,6
CRNN-Intensité	**18,8**	**66,8**	**80,7**	**17,3**	**6,3**

(c) Enregistrements

TABLEAU 4.4. – Précisions et erreurs angulaires sur des mélanges à deux sources des CRNNs de localisation prenant en compte les parties active et réactive du vecteur d'intensité (CRNN-Intensité), ou uniquement la partie active (CRNN-Intensité-active). Les intervalles de confiance à 95% sont compris entre 0,8 et 1,9% pour la précision et 0,5 et 1,5° pour l'erreur angulaire.

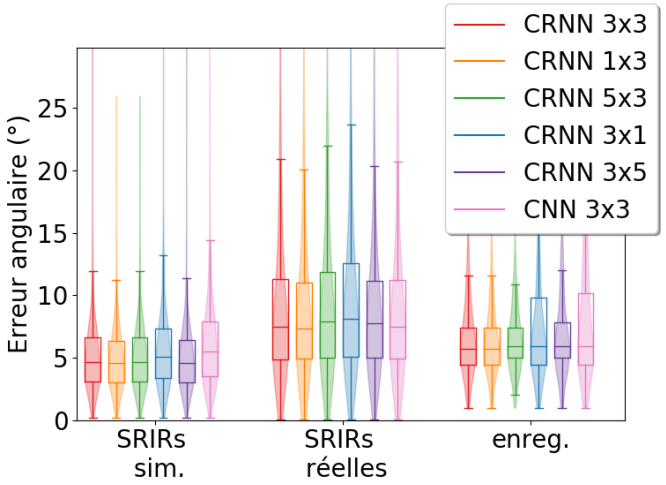


FIGURE 4.21. – Performances de localisation des CRNNs-Intensité sur les ensembles à une source selon de la taille des filtres de convolution et de la présence des couches bi-LSTM. Les tailles des filtres de convolution sont indiquées par trames temporelles x bandes de fréquence.

test constitué d'enregistrements réels que les autres CRNNs. En revanche, augmenter le champ réceptif fréquentiel avec un CRNN 3x5 n'améliore pas les performances, ce qui peut être dû au fait que l'apprentissage d'un plus grand nombre de paramètres est plus difficile.

Couches LSTM unidirectionnelles On teste également des CRNNs semblables à ceux de la Figure 4.2, en utilisant des couches LSTMs au lieu des biLSTMs qui ne sont pas utilisables dans un contexte quasi-temps-réel. Pour ces expériences, les réseaux sont appris et testés sur des mélanges à une source uniquement. Les résultats sont présentés dans le Tableau 4.5. Sur la précision à moins de 5 et 10° pour les enregistrements réels, les performances sont significativement moins bonnes avec des couches unidirectionnelles qu'avec des couches bidirectionnelles. Cependant, pour tous les autres ensembles de test et toutes les métriques, l'utilisation de couches LSTM dégrade à peine les performances, ce qui valide la possibilité d'utiliser ce travail dans un contexte quasi-temps-réel.

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN avec uniLSTM	**54,6**	**93,9**	98,8	**5,2**	**4,7**
CRNN avec biLSTM	**55,9**	**94,5**	**99,3**	**5,1**	**4,6**

(a) SRIRs simulées

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN avec uniLSTM	**24,8**	**66,0**	**85,1**	**10,1**	**7,9**
CRNN avec biLSTM	**26,5**	**67,3**	**84,4**	**9,7**	**7,5**

(b) SRIRs réelles

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN avec uniLSTM	23,4	74,2	**97,3**	**7,7**	**5,9**
CRNN avec biLSTM	**25,8**	**82,2**	**97,4**	**7,3**	**5,7**

(c) Enregistrements

TABLEAU 4.5. – Précisions et erreurs angulaires sur des mélanges à une seule source des CRNNs de localisation dont les couches récurrentes sont des LSTMs bidirectionnelles ou unidirectionnelles. Les intervalles de confiance à 95% varient entre 0,3 et 2,7% pour la précision et 0,6 et 1,7° pour l'erreur angulaire.

4.5.4. Cible et fonction de coût

Discrétisation de la sphère unité Pour formuler la localisation comme un problème de classification, nous utilisons la discrétisation de la sphère unité (4.1). D'autres grilles sont possibles. Nous étudions l'influence du choix de la grille en comparant les trois options suivantes :

— La discrétisation « rectangulaire » (4.1) utilisée jusqu'à présent, avec un pas angulaire $\alpha = 10°$ contenant 429 points. La distance maximale entre une direction de l'espace et un point de la grille est 7,0°.

— La grille de Lebedev[3] contenant 434 points. Le répartition des points est plus uniforme que sur la grille précédente. La distance angulaire maximale entre un point de l'espace et le point le plus proche est 6,5°.

— Afin d'étudier l'influence de la résolution de la grille, nous considérons également une grille de Lebedev à 974 points. Dans ce cas, la distance maximale entre une direction de l'espace et un point de la grille est 4,2°.

Ces différents systèmes sont testés sur les ensembles contenant une seule source ponctuelle par mélange.

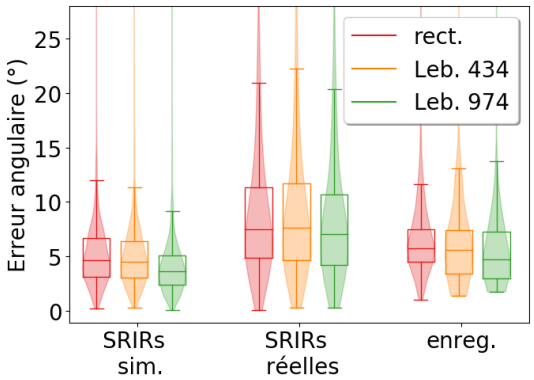


FIGURE 4.22. – Performances de localisation des CRNNs-Intensité selon de la discrétisation de la sphère unité utilisée.

On constate sur la Figure 4.22 qu'utiliser la grille Lebedev 434, plus régulière que la grille rectangulaire, améliore les performances sur les enregistrements réels pour le premier quartile. En effet, avec la grille rectangulaire, environ la moitié des directions d'arrivée des sources sont particulièrement éloignées de leur classe de référence (plus de 5°), ce qui n'est pas le cas avec les grilles de Lebedev. En dehors de cet artefact particulier, les grilles rectangulaires et de Lebedev ont des performances similaires. La grille de Lebedev 974, plus fine, permet d'améliorer la précision de l'estimation, notamment pour les deux premiers quartiles. Cependant, les trois réseaux ont des performances très satisfaisantes et comparables sur tous les ensembles de test, avec notamment très peu de résultats aberrants.
Les grilles de Lebedev (que nous n'avons pas adoptées) sont finalement légèrement préférables, mais ce choix n'est pas crucial.

Régression ou classification ? Nous avons discuté dans la partie 3.3.4 des avantages et inconvénients de formuler la localisation comme un problème de classification ou de

3. https://en.wikipedia.org/wiki/Lebedev_quadrature

régression. Dans le cas multi-sources, la régression s'accompagne notamment d'un certain nombre de difficultés techniques.
Dans cette partie, nous nous intéressons à l'impact de la formulation sur la capacité d'apprentissage du réseau, en mettant de côté ces difficultés. Nous nous limitons donc à la localisation d'une seule source. Nous comparons trois systèmes de régression et trois systèmes de classification :

— un réseau de régression ciblant les coordonnées sphériques. Ce réseau est appris à estimer deux sorties, θ et ϕ, en minimisant l'erreur aux moindres carrés. Cette fonction de coût ne prend pas en compte la géométrie sphérique de l'espace de sortie, en particulier la périodicité de l'azimut.

— un réseau de régression ciblant les coordonnées sphériques θ et ϕ, appris cette fois en utilisant l'erreur angulaire (4.4) comme fonction de coût. Cela permet de prendre en compte la géométrie sphérique de l'espace des directions d'arrivée, mais la projection de θ et ϕ sur la sphère n'est pas régulière, en particulier à proximité des pôles.

— un réseau de régression ciblant les coordonnées cartésiennes sur la sphère correspondant à la direction d'arrivée de la source. Ce réseau possède donc trois sorties, x, y et z. Un réseau semblable a été proposé par Adavanne et al. [148]. Nous utilisons la fonction de coût aux moindres carrés, qui représente cette fois la distance cartésienne entre les points sur la sphère correspondant à la direction estimée et à la vraie direction.

Pour ces trois réseaux de régression, la fonction d'activation après la dernière couche est la fonction sigmoïde. Nous avons également testé une activation linéaire, avec ou sans seuillage, avec des résultats légèrement moins bons.
Les réseaux de classification sont similaires au CRNN-Intensité présenté sur la Figure 4.2. En particulier, ils possèdent tous n_{DOA} sorties correspondant à la discrétisation rectangulaire de la grille (4.1).

— Un réseau de classification appris avec la fonction d'entropie croisée et un encodage *one-hot* pour les cibles. La classe correspondant à la direction d'arrivée la plus proche de la vraie direction se voit attribuer une valeur de 1, tandis que les autres sont assignées à 0. Nous utilisons ici la fonction softmax en sortie du réseau, contrairement au CRNN-Intensité présenté précédemment. En effet, c'est l'option la plus répandue pour les réseaux de classification car cela permet d'interpréter les sorties du réseau comme des probabilités, mais cela ne permet pas la généralisation à plusieurs sources. Dans ce cas, la structure de l'espace des directions d'arrivée n'est pas prise en compte.

— Un réseau de classification appris avec la fonction aux moindres carrés, avec une cible encodée par une distribution de Gibbs. Cette cible, moins discriminante qu'un encodage *one-hot*, permet de prendre en compte la structure de l'espace des directions d'arrivée, puisque les points au voisinage de la vraie direction d'arrivée se voient accorder un poids non nul. D'autre part, cela permet plus de réalisme dans le cas où la direction d'arrivée est équidistante de plusieurs classes, en attribuant

un poids équivalent à ces classes. Pour une source venant de la direction $\psi = (\theta, \phi)$, la valeur cible associée à la classe $\psi_{ij} = (\theta^i_j, \phi_i)$ est

$$\mathcal{G}(\psi_{ij}) = e^{-\delta[\psi_{ij},\psi]^2/\beta^2} \tag{4.11}$$

où δ est la distance angulaire (4.4) et β définit un voisinage angulaire. Pour ce réseau, nous utilisons la sigmoïde comme fonction d'activation pour la dernière couche associée à une erreur aux moindres carrés, de façon similaire à He et al. [185]. La sortie ne peut donc pas être interprétée comme une distribution de probabilité.

— Pour conserver la structure de l'espace de sortie tout en permettant une interprétation probabiliste de la sortie, nous proposons d'intégrer la distribution de Gibbs dans le cadre de l'entropie croisée, avec la fonction de coût suivante :

$$\text{coût} = -\log(\sigma_{ij}) - \sum_{\substack{(i',j') \\ \neq (i,j)}} (1 - \mathcal{G}(\psi_{i'j'})) \log(1 - \sigma_{i'j'}) \tag{4.12}$$

où ψ_{ij} est la classe la plus proche de la vraie direction d'arrivée, et σ_{ij} est la sortie du réseau pour la classe ψ_{ij} après une fonction d'activation sigmoïde. Pour $\beta = 0$, si la vraie direction d'arrivée est sur la grille, on retrouve la fonction de coût d'entropie croisée avec un encodage *one-hot* des cibles.

Le tableau 4.6 montre les résultats des différents réseaux sur les trois ensembles de test à une seule source. Sur l'ensemble créé à partir des SRIRs simulées (4.6a), les trois réseaux de classification obtiennent des performances également bonnes, et meilleures que le réseau de régression appris avec les moindres carrés sur les coordonnées sphériques. La régression sur les coordonnées cartésiennes obtient les meilleurs résultats. Sur les SRIRs réelles (4.6b), les résultats sont généralement moins bons que sur les SRIRs simulées, mais l'ordre de performance des réseaux est inchangé. Sur les enregistrements réels (4.6c), en revanche, le réseau de régression sphérique avec l'erreur angulaire obtient les meilleures performances en terme d'erreur angulaire médiane et moyenne. En raison de la construction de la grille, les réseaux de classification sont désavantagés pour la précision à moins de 5° d'erreur. En revanche, ils génèrent moins de résultats aberrants que les réseaux de régression, avec plus de 98% des sources localisées avec moins de 15° d'erreur.
En conclusion, bien que la régression soit écartée dans la plupart des travaux de localisation, celle-ci est tout à fait légitime en terme de performances. Cependant, des questions techniques se posent dès que plusieurs sources sont présentes (voir les discussions dans la partie 3.3.4). Ceci peut suffire à justifier la formulation par classification. Dans ce cas, il peut être avantageux de choisir une fonction de coût adaptée telle que celle que nous avons proposée utilisant la distribution de Gibbs.

4.6. Résumé

Dans ce chapitre, nous avons présenté un système de localisation de sources sonores dans des contenus ambisoniques par CRNN. Nous proposons une paramétrisation de la

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
Régression sphérique moindres carrés	53,5	88,1	96,2	5,9	4,7
Régression sphérique erreur angulaire	62,0	92,0	97,2	5,3	4,0
Régression cartésienne	**82,8**	**97,5**	**99,3**	**3,5**	**2,8**
Classification one-hot	56,1	95,0	**99,0**	5,1	4,6
Classification cible Gibbs	57,8	**96,7**	**99,5**	4,8	4,5
Classification coût Gibbs	57,2	**96,0**	**99,5**	4,9	4,6

(a) SRIRs simulées

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
Régression sphérique moindres carrés	28,5	60,2	81,2	11,4	8,1
Régression sphérique erreur angulaire	29,7	64,3	80,0	11,4	7,9
Régression cartésienne	**37,1**	**72,9**	**88,3**	**8,5**	**6,4**
Classification one-hot	26,6	66,8	85,3	**9,7**	8,0
Classification cible Gibbs	26,0	67,7	**86,4**	**9,3**	**7,5**
Classification coût Gibbs	27,6	67,7	84,5	**9,7**	**7,2**

(b) SRIRs réelles

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
Régression sphérique moindres carrés	25,4	71,7	87,6	11,4	7,1
Régression sphérique erreur angulaire	**49,8**	**90,8**	95,9	**7,1**	**5,0**
Régression cartésienne	46,8	87,7	93,3	9,1	**5,3**
Classification one-hot	22,6	85,5	**98,0**	**7,5**	5,9
Classification cible Gibbs	25,0	83,0	**98,3**	**7,3**	5,9
Classification coût Gibbs	27,3	88,8	**98,0**	**6,9**	5,7

(c) Enregistrements

TABLEAU 4.6. – Précisions et erreurs angulaires des réseaux de régression et de classification (a) sur les SRIRs simulées, (b) sur les SRIRs réelles, (c) sur les enregistrements réels. Les intervalles de confiance à 95% varient entre $\pm 0,3\%$ et $\pm 2,8\%$ pour la précision, et $\pm 0,6°$ et $\pm 1,6°$ pour l'erreur angulaire.

captation qui s'appuie sur le vecteur d'intensité acoustique normalisé, et nous montrons que cela permet un gain notable de robustesse en situation réelle. La visualisation du comportement des CRNNs par LRP permet d'observer que les attaques des sons sont particulièrement utiles pour la localisation. Puisqu'elles sont mises en valeur par le vecteur d'intensité acoustique, cela explique le gain de performance du CRNN-Intensité par rapport au CRNN utilisant une simple paramétrisation FOA. Plus précisément, nous montrons que l'utilisation de la partie réactive du vecteur d'intensité permet une amélioration notable : si elle n'est pas prise en compte, les performances du CRNN-Intensité sont

comparables à celle du CRNN-FOA. Cela semble indiquer que le CRNN-FOA n'est pas capable d'extraire les information présente dans la partie réactive du vecteur d'intensité. Nous étudions l'impact des conditions d'apprentissage sur les performances du CRNN, mettant en valeur l'importance de la diversité et de la taille de la base d'apprentissage, l'impact de la normalisation des données d'entrée, l'influence de l'architecture du réseau et en particulier des convolutions selon l'axe fréquentiel. En revanche, les convolutions selon l'axe temporel et le choix de la discrétisation de la sphère unité pour la classification n'ont qu'une importance limitée. Enfin, nous présentons différentes stratégies pour formuler la localisation comme un problème de classification ou de régression, montrant qu'avec des fonctions de coût adaptées la régression est tout aussi puissante que la classification, bien que peu adaptée au cas multi-sources.

5. Rehaussement de la parole par des filtres déduits de masques temps-fréquence

Dans ce chapitre, nous présentons le système mis au point pour extraire une source de parole d'un mélange ambisonique multi-locuteurs. La direction d'arrivée de chacun des locuteurs est considérée connue. Il peut s'agir de la véritable direction d'arrivée ou d'une estimation. Un réseau de neurones est utilisé afin d'estimer un masque temps-fréquence correspondant à la locutrice cible. Ce masque est ensuite utilisé pour estimer un filtre spatial et extraire la voix. La technique proposée est adaptée au format ambisonique, mais pourrait être appliquée pour d'autres types de contenus. Nous évaluons les performances de ce système en terme de reconnaissance vocale en fonction des données d'entrée du réseau, des conditions d'apprentissage et du filtre choisi. Nous vérifions également la robustesse du système de rehaussement vis-à-vis des erreurs de localisation, ainsi que son efficacité sur des enregistrements réels.

5.1. Structure de la solution

5.1.1. Obtention des masques par un réseau de neurones

Dans un premier temps, un réseau de neurones est utilisé afin d'estimer un masque temps-fréquence correspondant à la source de parole cible à partir du signal au format FOA contenant plusieurs sources de parole et du bruit.

Structure du réseau Le réseau de neurones utilisé pour estimer les masques est présenté sur la Figure 5.1. Les signaux d'entrée sont présentés au réseau sous forme de spectrogrammes. Seule l'amplitude est prise en compte. Plusieurs trames de signal sont regroupées en séquences. Afin de permettre au réseau d'identifier la source cible dans le mélange, on lui présente plusieurs versions du signal capté ayant subi différents prétraitements : grâce aux directions d'arrivée des sources cible et interférente supposées connues, on utilise des filtres de formation de voie FOA contraints (2.33) afin d'obtenir des estimations monocanales $\hat{s}(t,f)$ et $\hat{n}(t,f)$ de chaque source. On présente au réseau de neurones les spectrogrammes d'amplitude de $\hat{s}(t,f)$, $\hat{n}(t,f)$ et $x_W(t,f)$, où $x_W(t,f)$ représente le module du canal omnidirectionnel du mélange $\mathbf{x}(t,f)$ défini dans l'équation (3.12). L'influence du choix des spectrogrammes d'entrée est présenté dans la partie 5.3.1. Pour être traités par le réseau, ces signaux sont concaténés selon l'axe fréquentiel. Le réseau est constitué d'une couche LSTM suivie d'une couche FF distribuée

selon l'axe temporel.

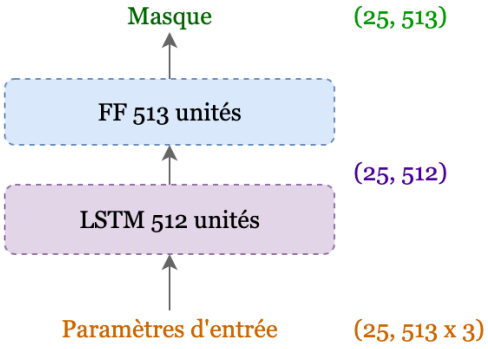


FIGURE 5.1. – Architecture du réseau de neurones d'estimation de masque de parole.

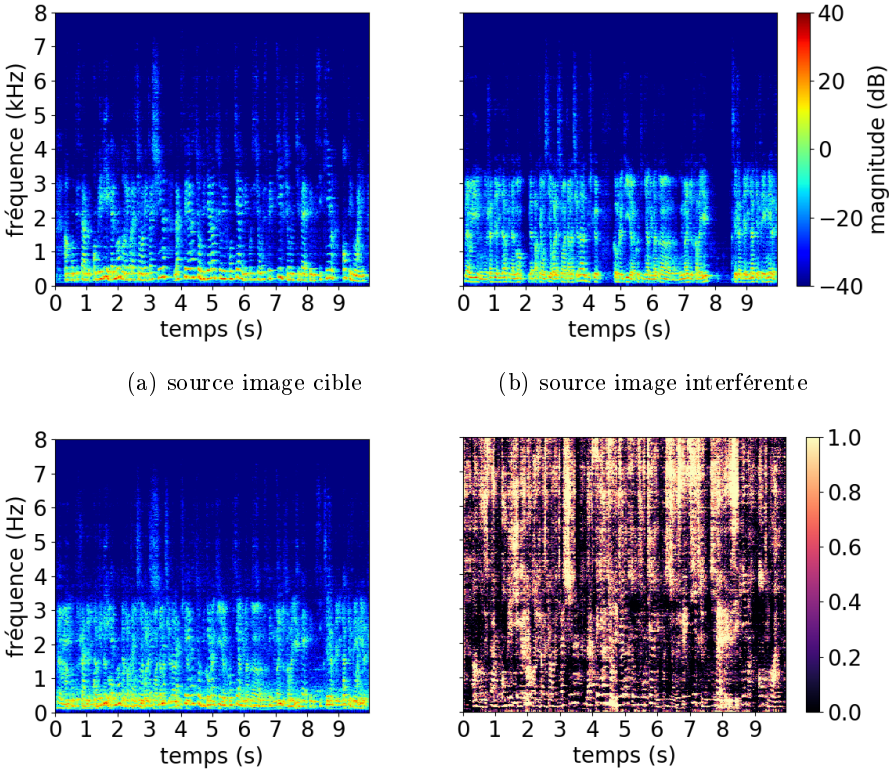
La cible du réseau lors de l'apprentissage est le masque de Wiener idéal. Il est calculé à partir de $\mathbf{s}(t,f)$ et de $\mathbf{n}(t,f)$, constitué des signaux de parole interférents et du bruit ambiant introduits dans l'équation (3.12). La formule du masque de Wiener idéal pour le canal omnidirectionnel W du signal FOA est la suivante :

$$M_s^{(\mathrm{id})}(t,f) = \frac{s_{\mathrm{W}}(t,f)^2}{s_{\mathrm{W}}(t,f)^2 + n_{\mathrm{W}}(t,f)^2}. \tag{5.1}$$

où $s_{\mathrm{W}}(t,f)$ est le canal omnidirectionnel de $\mathbf{s}(t,f)$. Un exemple de masque idéal est présenté sur la Figure 5.2d.

5.1.2. Système complet

Le réseau de neurones d'estimation de masque est intégré dans un système plus complexe présenté sur la Figure 5.3.
Dans un premier temps, on calcule les estimations monocanales $\hat{s}(t,f)$ et $\hat{n}(t,f)$ de chaque source, dont les spectrogrammes d'amplitude seront donnés en entrée au réseau conjointement avec celui de $x_W(t,f)$. Le réseau renvoie ensuite un masque correspondant à la parole cible, $M_s(t,f)$. La Figure 5.4 montre un exemple de masque estimé correspondant au masque idéal de la Figure 5.2d. Le masque prédit permet d'estimer une version multicanale de la parole et des interférences, $\tilde{\mathbf{s}}(t,f)$ et $\tilde{\mathbf{n}}(t,f)$, grâce à l'équation (3.33). Les estimations des matrices de covariance $\mathbf{R_{ss}}(t,f)$ et $\mathbf{R_{nn}}(t,f)$ calculées d'après (3.34) permettent de dériver un filtre de Wiener multicanal (voir partie 3.2.2) fournissant une estimation monocanale $y(t,f)$ de la parole cible. Sauf mention contraire, le filtre choisi est le MWF-r1 (5.5). L'influence du filtre est discutée dans la partie 5.3.3.
Ce système entre dans le cadre des techniques de filtrage multicanal utilisant un masque temps-fréquence présentées dans la partie 3.2.4.1. Par rapport aux techniques existant dans la littérature, nous proposons additionnellement d'utiliser les estimations $\hat{s}(t,f)$ et $\hat{n}(t,f)$ calculées à partir du signal FOA en entrée du réseau de neurones, afin de pouvoir traiter le cas multi-sources.



(a) source image cible (b) source image interférente

(c) mélange (d) masque de Wiener idéal

FIGURE 5.2. – Spectrogrammes des canaux W des sources images cible (a) et interférente (b) à 45° d'écart angulaire, ainsi que du canal $x_W(t, f)$ de leur mélange avec du bruit diffus (c). (d) Masque de Wiener idéal correspondant, ciblé par le réseau.

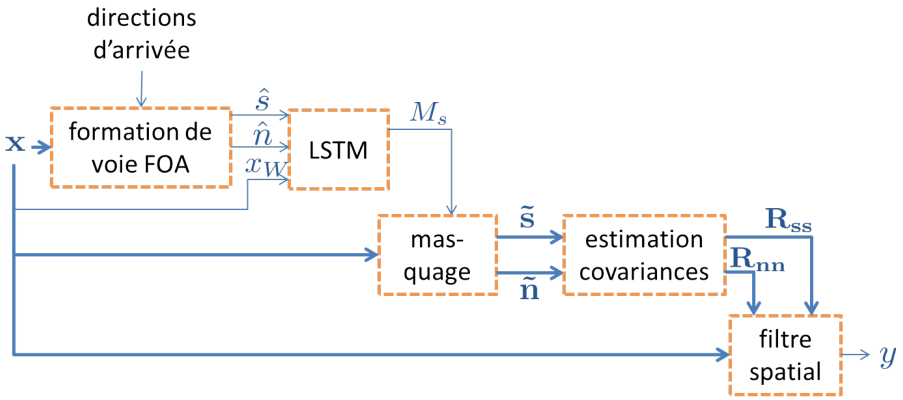


FIGURE 5.3. – Système de rehaussement de la parole proposé.

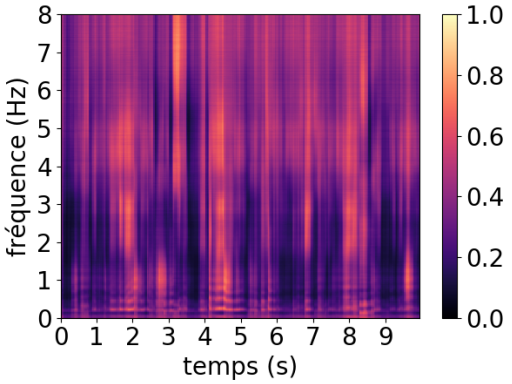


FIGURE 5.4. – Exemple de masque estimé par le réseau LSTM.

Après le rehaussement proposé, nous appliquons un module de déreverbération par WPE (voir partie 3.2.2) [56] sur le signal monocanal $y(t, f)$. Nous utilisons 50 coefficients de filtrage et un délai d'estimation de 3, ce qui permet d'obtenir les meilleurs résultats lors de l'évaluation.

5.2. Protocole expérimental

5.2.1. Paramètres audio

Tous les signaux audio sont échantillonnés à 16 kHz. La TFCT est calculée avec une fenêtre d'analyse sinusoïdale de 1024 échantillons et un recouvrement de 50%. Une fenêtre de synthèse, identique à la fenêtre d'analyse, est utilisée pour la reconstruction.

5.2.2. Paramètres d'apprentissage

Les signaux sont présentés au réseau par séquence de 25 trames. Le réseau renvoie une trame de masque par trame présentée en entrée. Les séquences sont extraites des signaux d'apprentissage en utilisant un recouvrement de 12 trames entre deux séquences, afin d'obtenir plus de données d'apprentissage que s'il n'y avait aucun recouvrement. La couche LSTM du réseau contient 512 unités cachées. La couche FF contient 513 unités, correspondant aux 513 bandes de fréquence du masque. Elle est suivie d'une activation sigmoïdale qui permet d'assurer que le masque estimé est compris entre 0 et 1. Le réseau est optimisé pour la fonction de coût aux moindres carrés grâce à l'optimiseur Nadam [178]. Le pas d'apprentissage est initialement fixé à 10^{-3}. On utilise une régularisation L2 de 10^{-4} sur la fonction de coût et un *dropout* de 50% sur les poids, à la fois pour les poids récurrents et les poids non-récurrents. L'apprentissage est interrompu lorsque l'erreur cesse de diminuer sur l'ensemble de validation pendant 5 époques consécutives, donnant lieu à un apprentissage d'environ 20 époques.

5.2.3. Ensembles d'apprentissage et de validation

La base d'apprentissage est constituée de 1801 extraits de 10 s issus d'un sous-ensemble du corpus Bref [180] contenant 44 locuteurs et locutrices différents. Chaque extrait est convolué avec une SRIR simulée choisie au hasard dans la base décrite dans la partie 4.2.3. Un second extrait de parole et une seconde SRIR sont piochés au hasard afin de synthétiser le signal de parole interférent. La distance angulaire minimale entre les SRIRs est de 25° ; le SIR entre la source cible et la source interférente est fixé à 0 dB, sauf mention du contraire. Chacun des 1801 extraits apparaît donc une fois exactement en tant que signal cible, et potentiellement plusieurs fois comme signal interférent. Un bruit de foule spatialement diffus est synthétisé comme dans la partie 4.2.3 et ajouté au mélange de manière à obtenir un SNR de 20 dB.
Un ensemble de validation est constitué de la même manière, avec un autre sous-ensemble de Bref contenant 684 signaux de 10 s prononcés par 17 nouveaux locuteurs et convolué avec de nouvelles SRIRs simulées. Les bruits de foule ajoutés à l'ensemble de validation sont également différents de ceux de l'ensemble d'apprentissage.

5.2.4. Ensembles de test

Tous les ensembles de test sont constitués des mêmes 20 extraits d'environ une minute issus du corpus Ester [186], avec certains paramètres variant en fonction de l'expérience menée. Chaque ensemble de test contient un total de 4043 mots pour le signal cible. Chacun des 20 locuteurs est utilisé exactement une fois pour le signal cible, et peut être pioché aléatoirement en tant que signal interférent. Les signaux de parole sont convolués avec des SRIRs piochées aléatoirement parmi les SRIRs réelles présentées dans la partie 4.2.4. Selon les expériences, la distance angulaire entre les sources peut être fixe ou variable. Le SIR est fixé à 0 dB, sauf mention contraire. Du bruit de foule spatialement diffus non vu en apprentissage, est ajouté avec un SNR de 20 dB par rapport à la première source.

5.2.5. Mesure de performance

Pour évaluer les performances de nos algorithmes, nous utilisons un moteur de RAP français, Cobalt, mis au point par les équipes de recherche d'Orange Labs. Il s'appuie sur un modèle acoustique neuronal développé en Kaldi [187], entraîné sur plus de 2000 h de signaux bruités et réverbérés, mais sans pré-traitement pouvant induire des distorsions. Le modèle de langage s'appuie sur des 5-grammes et comprend un lexique de 1.7 millions de mots. L'apprentissage s'effectue sur plus de 3 milliards de mots. Nous utilisons par la suite ce système de RAP comme une boîte noire.
La métrique que nous utiliserons est le WER, calculé de la façon suivante :

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100 \tag{5.2}$$

avec S le nombre de substitutions de mots, D le nombre de suppressions, I le nombre d'insertions et N le nombre total de mots de la vérité terrain (voir Figure 5.5). Notons

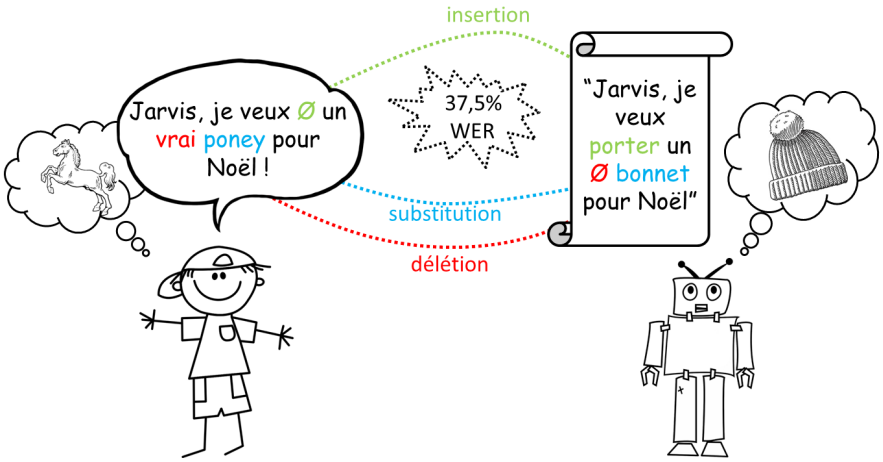


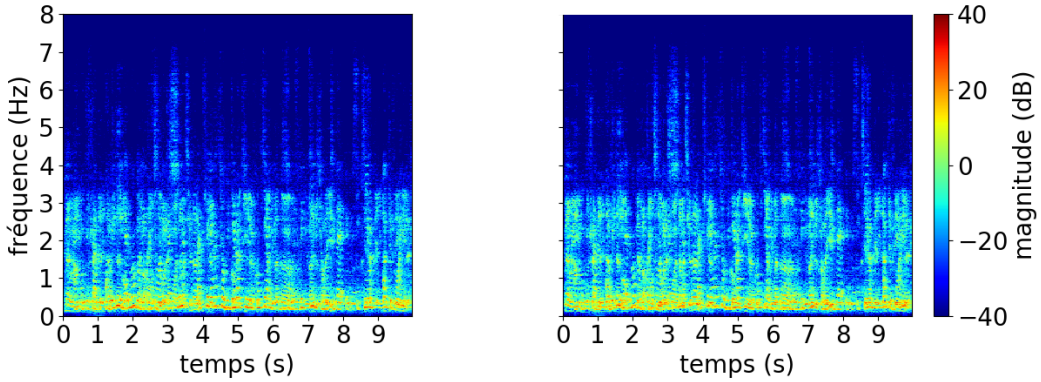
FIGURE 5.5. – Exemple de calcul de WER.

que le WER n'est pas strictement un pourcentage ; sa valeur minimale est 0 pour une transcription parfaite, mais il ne possède pas de maximum en raison du terme I non borné.

5.3. Résultats

5.3.1. Influence du choix des entrées

Dans une première expérience, nous mesurons l'impact des signaux choisis comme entrée pour le réseau parmi $x_W(t,f)$, $\hat{s}(t,f)$ et $\hat{n}(t,f)$. Un exemple de ces signaux est présenté sur les Figures 5.2c et 5.6. On constate que les estimations $\hat{s}(t,f)$ et $\hat{n}(t,f)$ (Figures 5.6a et 5.6b) sont très éloignées des canaux omnidirectionnels des deux sources de parole (Figures 5.2a et 5.2b). Un réseau est appris pour chaque jeu de données d'entrée.
Des ensembles de test sont constitués pour 25° et 45° d'écart angulaire entre les sources, avec ±2° de tolérance. Les SRIRs sont donc différentes pour les deux ensembles de test. En revanche, les extraits de parole sont identiques. Les résultats de WER sont présentés dans le Tableau 5.1. Tous les signaux soumis au système de RAP sont déréverbérés par WPE en dehors des signaux de parole initiaux non réverbérés.
Les performances pour le corpus non réverbéré constituent le WER optimal qu'il est possible d'obtenir avec le système de RAP utilisé. Le WER sur les signaux images $s_W(t,f)$ indique la performance de RAP pour des sources parfaitement séparées et débruitées, mais il est théoriquement possible de surpasser ce résultat puisque le filtrage multicanal proposé induit une légère déréverbération. Pour compléter la comparaison, nous présentons les résultats de WER sur le signal mélangé $x_W(t,f)$ et sur l'estimation du signal cible par formation de voie pleine bande $\hat{s}(t,f)$, qui permettent de mesurer l'amélioration du système proposé par rapport à une absence de rehaussement ou un rehaussement purement spatial de référence. Nous présentons également les performances obtenues



(a) Estimation $\hat{s}(t,f)$ (b) Estimation $\hat{n}(t,f)$

FIGURE 5.6. – Spectrogrammes des estimations par formation de voie des signaux cible $\hat{s}(t,f)$ (a) et interférent $\hat{n}(t,f)$ (b).

Écart angulaire			**25°**	**45°**
Parole non réverbérée			7,6	
Source image s_W			10,2	10,6
Mélange x_W			83,0	73,6
Formation voie FOA $\hat{s}$			79,3	53,8
Masquage par $M_s^{(\mathrm{id})}$			14,2	13,9
Filtre issu de $M_s^{(\mathrm{id})}$			14,7	13,6
Entrées du réseau	x_W	masque M_s	82,9	71,5
		filtre $\mathbf{w}_{\text{MWF-r1}}$	79,5	71,8
	$\hat{s}$	masque M_s	84,0	70,2
		filtre $\mathbf{w}_{\text{MWF-r1}}$	65,7	45,2
	x_W, $\hat{s}$	masque M_s	82,5	65,9
		filtre $\mathbf{w}_{\text{MWF-r1}}$	60,1	25,9
	x_W, $\hat{s}$, $\hat{n}$	masque M_s	73,5	59,1
		filtre $\mathbf{w}_{\text{MWF-r1}}$	**19,0**	**17,0**

TABLEAU 5.1. – WER en fonction des paramètres d'entrée sur des ensembles de test avec des écarts angulaires entre les sources égaux à 25° ou 45°. Les résultats sont indiqués avec un intervalle de confiance à 95% de $\pm 1,0\%$.

par masquage temps-fréquence ou par filtrage multicanal à partir des masques idéaux $M_s^{(id)}(t,f)$.

Lorsque le réseau ne voit que le canal omnidirectionnel $x_W(t,f)$ du signal FOA capturé, les masques estimés et les filtres qui en découlent ne permettent pas d'améliorer significativement les performances du système de RAP par rapport à une reconnaissance effectuée sur le signal $x_W(t,f)$ lui-même, ou à peine. Le WER qui en résulte est d'ailleurs

plus élevé qu'avec le signal $\hat{s}(t,f)$ rehaussé par le filtre de formation de voie FOA pleine bande pointant vers la source d'intérêt. En effet, le réseau ne possède aucune information lui indiquant quelle source de parole doit être rehaussée parmi les deux sources présentes. En fournissant uniquement le signal $\hat{s}(t,f)$ au réseau, le filtre $\mathbf{w}_{\text{MWF-r1}}(f)$ issu du système global permet d'obtenir un WER de 45,2% pour les sources situées à 45° d'écart, au lieu de 53,8% lorsque la RAP est effectuée sur le signal $\hat{s}(t,f)$. Lorsque les sources sont à 25° d'écart, le WER passe de 79,3% pour le signal rehaussé $\hat{s}(t,f)$ à 65,7% pour le signal filtré par $\mathbf{w}_{\text{MWF-r1}}(f)$. C'est une amélioration notable, mais le WER reste trop élevé pour que la transcription puisse être utilisée en conditions réelles.
Combiner $x_W(t,f)$ et $\hat{s}(t,f)$ en entrée du réseau permet une nette amélioration pour les sources écartées de 45°, avec un WER de 25,9% seulement. En revanche, pour les sources à 25°, l'information reçue par le réseau n'est pas suffisante pour que le filtre obtenu soit satisfaisant, avec encore 60,1% de WER.
En revanche, si l'on ajoute à $x_W(t,f)$ et $\hat{s}(t,f)$ le signal $\hat{n}(t,f)$ résultant du filtre de formation de voie pleine bande pointant vers la source concurrente, les WERs sur les signaux issus du filtre $\mathbf{w}_{\text{MWF-r1}}(f)$ sont de 17,0% et 19,0% sur les signaux à 45° et 25°, respectivement. Cela représente des diminutions relatives du WER de 68% et 76% par rapport aux signaux rehaussés par formation de voie pleine bande et permet de se rapprocher des performances obtenues avec des masques idéaux.
Dans les cas les plus difficiles, l'ajout de $\hat{n}(t,f)$ apporte donc une amélioration majeure, bien que les signaux $\hat{s}(t,f)$ et $\hat{n}(t,f)$ soient très bruités, comme le montrent les spectrogrammes de la Figure 5.6. Nous considérons par la suite des réseaux de neurones appris et testés avec $x_W(t,f)$, $\hat{s}(t,f)$ et $\hat{n}(t,f)$ en entrée.

5.3.2. Influence des conditions d'apprentissage

Dans cette expérience, nous testons la robustesse du système de rehaussement proposé aux variations du SIR entre les sources. Nous considérons trois réseaux. Le premier est le réseau de référence présenté dans la partie 5.1.1, pour lequel tous les mélanges de l'ensemble d'apprentissage sont formés avec un SIR de 0 dB. Pour les deuxième et troisième réseaux, les SIRs des mélanges d'apprentissage sont tirés aléatoirement dans [0 5] et [-5 5] dB, respectivement. Les réseaux sont testés sur trois ensembles de test contenant des mélanges dont les SIRs sont fixés à -5, 0 et 5 dB. Pour tous les ensembles de test, l'écart angulaire entre les sources est tiré aléatoirement et est supérieur à 25°.
Les résultats sont présentés dans le Tableau 5.2. Quels que soient les ensembles de test, les réseaux de neurones présentent des résultats équivalents au masque idéal, à l'exception du réseau appris sur des ensembles à [0 5] dB qui est légèrement moins performant sur les mélanges à -5 dB. Le filtre idéal reste légèrement meilleur. En particulier, cette expérience montre qu'il est suffisant d'apprendre le réseau sur un SIR fixe de 0 dB pour qu'il puisse généraliser à d'autres SIRs suffisamment proches.

SIR test			-5 dB	0 dB	5 dB
Parole non réverbérée				7,6	
Source image s_W				11,8	
Mélange x_W			100,8	79,6	42,5
Formation voie FOA $\hat{s}$			72,1	39,1	20,6
Masquage par $M_s^{(\mathrm{id})}$			19,6	16,1	14,3
Filtre issu de $M_s^{(\mathrm{id})}$			16,8	14,0	12,8
SIR appr.	0 dB	masque M_s	89,3	59,3	32,7
		filtre $\mathbf{w}_{\text{MWF-r1}}$	**20,4**	**15,9**	**15,0**
	[0 5] dB	masque M_s	89,7	57,7	31,0
		filtre $\mathbf{w}_{\text{MWF-r1}}$	22,1	**16,6**	**15,2**
	[-5 5] dB	masque M_s	90,3	59,3	31,9
		filtre $\mathbf{w}_{\text{MWF-r1}}$	**20,8**	**15,9**	**15,3**

TABLEAU 5.2. – WER pour des modèles appris avec des mélanges à différents SIRs, pour des ensembles de test avec un écart angulaire entre les sources supérieur à 25°. Les résultats sont indiqués avec un intervalle de confiance à 95% de $\pm 1,0\%$.

5.3.3. Influence du filtre

Différents filtres multicanaux peuvent être calculés à partir des masques estimés par le réseau de neurones (voir la partie 3.2.2). En particulier, nous comparons le filtre MWF-r1 utilisé dans le système proposé et les filtres max-SNR et MWF. Nous rappelons leurs équations :

$$\mathbf{w}_{\text{MWF}}(t,f) = [\mathbf{R_{ss}}(t,f) + \mathbf{R_{nn}}(t,f)]^{-1}\mathbf{R_{ss}}(t,f)\mathbf{u}_0, \tag{5.3}$$

$$\mathbf{w}_{\text{max-SNR}}(t,f) = \mathcal{P}[\mathbf{R_{nn}}^{-1}(t,f)\mathbf{R_{ss}}(t,f)] \tag{5.4}$$

et

$$\mathbf{w}_{\text{MWF-r1}}(t,f) = [\mathbf{R_{ss-r1}}(t,f) + \mathbf{R_{nn}}(t,f)]^{-1}\mathbf{R_{ss-r1}}(t,f)\mathbf{u}_0. \tag{5.5}$$

Les filtres MWF-r1 et max-SNR peuvent être associés ou non à une normalisation BAN (3.26) afin de limiter les distorsions. Le filtre plébiscité par la plupart des auteurs travaillant sur le filtrage multicanal par réseau de neurones est le max-SNR avec une normalisation BAN [49]. Les ensembles de tests utilisés contiennent des signaux où les sources sont écartées de 25° ou 45°.

Les résultats sont présentés dans le Tableau 5.3. Lorsque les filtres sont calculés à partir des masques idéaux $M_s^{(\mathrm{id})}$, le MWF est significativement moins bon que les autres filtres. Pour des sources à 25°d'écart, le filtre max-SNR avec normalisation BAN préconisé par Heymann et al. [49] obtient 19,1% de WER. Le filtre MWF-r1 sans normalisation BAN est le plus performant, avec 14,7% de WER, soit une diminution relative du WER de 23% par rapport au max-SNR.

En calculant les filtres à partir des masques estimés par le réseau de neurones, les performances sont globalement moins bonnes, mais la relation d'ordre entre les performances des différents filtres est globalement inchangée. En particulier, sur les sources à 25°

	Écart angulaire	**25°**	**45°**
Parole non réverbérée		7,6	
Source image s_W		10,2	10,6
Mélange x_W		83,0	73,6
Formation voie FOA $\hat{s}$		79,3	53,8
Masquage par $M_s^{(\mathrm{id})}$		14,2	13,9
Filtres issus de $M_s^{(\mathrm{id})}$	MWF $\mathbf{w}_{\mathrm{MWF}}$	30,4	24,4
	max-SNR $\mathbf{w}_{\text{max-SNR}}$	20,1	19,4
	max-SNR-BAN $w_{\mathrm{BAN}}\mathbf{w}_{\text{max-SNR}}$	19,1	18,2
	MWF rang 1 $\mathbf{w}_{\text{MWF-r1}}$	**14,7**	**13,6**
	MWF rang 1 $w_{\mathrm{BAN}}\mathbf{w}_{\text{MWF-r1}}$	19,7	17,7
Masquage par M_s		73,5	59,1
Filtres issus de M_s	MWF $\mathbf{w}_{\mathrm{MWF}}$	52,0	31,7
	max-SNR $\mathbf{w}_{\text{max-SNR}}$	29,6	27,6
	max-SNR-BAN $w_{\mathrm{BAN}}\mathbf{w}_{\text{max-SNR}}$	30,3	26,5
	MWF rang 1 $\mathbf{w}_{\text{MWF-r1}}$	**19,0**	**17,0**
	MWF rang 1 $w_{\mathrm{BAN}}\mathbf{w}_{\text{MWF-r1}}$	27,5	22,5

TABLEAU 5.3. – WER pour différents filtres multicanaux sur des ensembles de test créés à partir de SRIRs réelles, avec un écart angulaire entre les sources de 25° ou 45°. Les résultats sont indiqués avec un intervalle de confiance à 95% de $\pm 1,0\%$.

d'écart, les filtres max-SNR avec et sans normalisation BAN sont équivalents, avec 29,6% de WER sans la normalisation. Le filtre MWF-r1 obtient quant à lui un WER de 19,0%, soit une diminution relative de 36% par rapport au max-SNR.
Nous avons testé différents post-traitements appliqués au masque avant de calculer les filtres, ce qui est également conseillé par Heymann et al. [31]. Mettre le masque estimé à l'échelle entre 0 et 1, afin de le forcer une amplitude maximale, n'a pas d'impact sur le WER, qui vaut alors 19,5% pour le MWF-r1 sur l'ensemble à 25° d'écart angulaire au lieu de 19,0%. Nous avons également utilisé un seuil (fixé manuellement) pour faire apparaître les points temps-fréquence correspondant à la parole cible, rendant ainsi le masque binaire. Dans ce cas, les performances sont significativement moins bonnes, avec un WER qui atteint cette fois 32,0%. Par la suite, nous n'appliquerons donc aucun post-traitement sur le masque.

5.3.4. Robustesse aux erreurs de localisation

En vue d'utiliser le système de rehaussement proposé dans une situation réelle où les directions d'arrivée ne sont pas parfaitement connues, nous testons la robustesse du système en introduisant artificiellement une erreur sur les directions d'arrivée. Les filtres de formation de voie pleine bande appliqués pour obtenir les signaux d'entrée du réseau, $\hat{s}$ et $\hat{n}$, ne pointent alors plus tout à fait vers les sources.
En pratique, pour chaque signal de test, nous ajoutons un biais tiré aléatoirement dans

$[-\beta, \beta]$ à l'azimut, et un autre biais tiré dans le même intervalle à l'élévation, où β désigne le biais maximal. Le réseau reste celui présenté précédemment ; il n'est pas réappris avec erreurs.
On voit dans le Tableau 5.4 que lorsque les sources sont à 25° d'écart, l'impact d'une erreur de localisation n'est pas significatif pour $\beta = 5°$. Pour $\beta = 10°$, le WER sur le signal rehaussé est de 22,0%, contre 19,0% sans erreur. Lorsque les sources sont écartées de 45°, le WER passe de 17,0% à 22,2% lorsque β passe de 0 à 10°. Dans tous les cas, la perte de performance reste raisonnable et laisse penser que ce système de rehaussement peut être couplé à un système de localisation réel, ce que nous faisons dans la partie 5.4.

	25°			**45°**		
Biais max. β	±0°	±5°	±10°	±0°	±5°	±10°
Parole non réverbérée		7,6			7,6	
Source image s_W		10,2			10,6	
Mélange x_W		83,0			73,6	
Formation voie FOA $\hat{s}$	79,3	79,2	78,9	53,8	54,7	56,3
Masquage par $M_s^{(\mathrm{id})}$		14,2			13,9	
Filtre issu de $M_s^{(\mathrm{id})}$		14,7			13,6	
masque M_s	73,5	73,8	75,3	59,1	58,9	59,4
filtre $\mathbf{w}_{\text{MWF-r1}}$	19,0	19,0	22,0	17,0	18,3	22,2

TABLEAU 5.4. – WER lorsque les azimuts et élévations des sources sont fournis avec une erreur variable, pour des écarts angulaires entre les sources égaux à 25° ou 45°. Les résultats sont indiqués avec un intervalle de confiance à 95% de ±1,0%.

5.4. Intégration des modules de localisation et de rehaussement en situation réelle

Les résultats présentés jusqu'ici sont obtenus sur des signaux créés à partir de SRIRs mesurées dans un environnement maîtrisé, à partir des directions d'arrivée exactes des sources. Nous testons dans cette partie la généralisation du système de rehaussement proposé à des enregistrements réels, en obtenant les directions d'arrivée avec le module de localisation mis au point dans le chapitre 4.

5.4.1. Ensemble de test

Les tests sont effectués sur les enregistrements réels présentés dans la partie 4.2.4. Ils contiennent des extraits du Petit Prince lus par trois lecteurs placés autour d'une table basse, au-dessus de laquelle est situé l'antenne de microphones. Pour chaque scène, les lecteurs restent à la même position, bien que des mouvements de têtes soient inévitables. Les lectures d'environ 5 minutes chacune sont enregistrées séparément et constituent au total 71 minutes de signal, soit 13 096 mots. Contrairement aux tests de la partie 4.2, aucune détection vocale n'est appliquée pour supprimer les passages de silence.

Chaque session de lecture est mélangée aléatoirement avec une lecture effectuée par un autre locuteur ou une autre locutrice, à niveau égal en moyenne sur l'extrait. L'écart angulaire entre les sources est compris entre 25 et 90°. Des bruits non maîtrisés sont présents entre 5 et 10 dB SNR : pages qui se tournent, pas, bruits provenant de la rue... Une télévision a également été enregistrée avec différents contenus (musique, publicité, série). Compte tenue de la taille de l'écran et de la distance au microphone, c'est une source interférente directionnelle mais plus spatialement étendue qu'une locutrice, ce qui n'a pas été vu lors de l'apprentissage où le bruit interférent est tout à fait diffus. Elle est ajoutée au mélange avec un SNR de 20 dB par rapport à chacune des sources de parole.

5.4.2. Suivi de sources

Si l'on considère que les directions d'arrivée réelles des sources ne sont pas connues, même si celles-ci sont fixes, il est nécessaire de rattacher les sorties du système de localisation à chaque trame aux sources correspondantes, afin fournir au module de rehaussement les entrées dans le bon ordre, avec le signal cible en premier. Nous utilisons pour cela un module supplémentaire de suivi de sources, en considérant qu'au début de la scène sonore le signal cible est identifié (cela peut se faire, par exemple, à l'aide d'un mot-clé). Dans le cadre des signaux audio, une partie des tâches du challenge LOCATA nécessitait de procéder au suivi de sources audio immobiles ou en mouvement [188]. Ce challenge a permis de mettre en lumière la difficulté intrinsèque de ce problème, mais aussi de la constitution de bases de données annotées de sources en mouvement et de l'évaluation des systèmes de suivi.
Le suivi de sources est un domaine complexe sur lequel nous n'avons pas travaillé dans le cadre de cette thèse. Nous utilisons donc un système existant de suivi par filtrage particulaire proposé par Kitić et al. dans leur solution soumise au challenge LOCATA, le VVM [127]. Le module de filtrage particulaire prend en entrée les sorties brutes du CRNN de localisation, les scores pour toutes les classes correspondant aux directions d'arrivée. Chaque source est modélisée par un ensemble de particules caractérisées par leur position et leur vecteur vélocité. La position et la vélocité de la source est déduite de celles des particules par une moyenne pondérée. La probabilité qu'une observation issue du CRNN soit due à une source en mémoire est calculée en fonction des positions, vélocités et poids des particules. De plus, des formules sont proposées pour calculer la probabilité d'apparition et de disparition de sources en fonction des observations à un instant donné par rapport aux sources et aux particules en mémoire.

5.4.3. Résultats de localisation

Les performances de localisation sont présentées en termes de précision à 5, 10 et 15°, ainsi que par le biais des moyenne et médiane de l'erreur angulaire. Aucune détection vocale n'est appliquée. La vérité terrain est considérée égale à la direction d'arrivée tout au long de la scène. Dans les séquences où la source de parole n'est pas présente, il est impossible pour le réseau d'estimer une direction d'arrivée en lien avec la vérité terrain. Cependant, le module de suivi doit permettre d'extrapoler la position de la source y

compris dans les périodes de silence.
Les résultats de localisation avec le CRNN seul ou bien couplé au module de filtrage particulaire sont présentés dans le Tableau 5.5.

	Précision (%)			Err. ang. (°)	
Algorithme	<5°	<10°	<15°	moy.	méd.
CRNN-Intensité	15,7	63,7	73,6	22,3	7,1
CRNN-Intensité + suivi	**25,6**	**65,2**	**78,5**	**14,5**	**7,5**

TABLEAU 5.5. – Précisions et erreurs angulaires du CRNN-Intensité avec et sans suivi de source par filtrage particulaire sur les enregistrements réels avec télévision. Les intervalles de confiance à 95% varient entre $\pm 1,1\%$ et $\pm 1,3\%$ pour la précision, et $\pm 0,7°$ et $\pm 1,0°$ pour l'erreur angulaire.

La baisse de performance du CRNN-Intensité par rapport aux résultats du Tableau 4.2 est due à la présence supplémentaire de la télévision, ainsi qu'à l'absence de détection vocale. Conformément à nos attente, on observe que le module de filtrage particulaire permet de gagner en précision. En particulier, le nombre de sources localisées avec moins de 5° d'erreur passe de 15,7 à 25,6%. L'erreur moyenne, elle, passe de 22,3 à 14,5° avec l'ajout du suivi de sources. Cependant, l'erreur médiane reste similaire. Cela traduit le fait que le module de filtrage particulaire permet principalement d'éviter les résultats aberrants, par exemple la détection de la télévision ou une estimation sans fondement lorsque la source est inactive. Un exemple de suivi de sources est présenté sur la Figure 5.7.

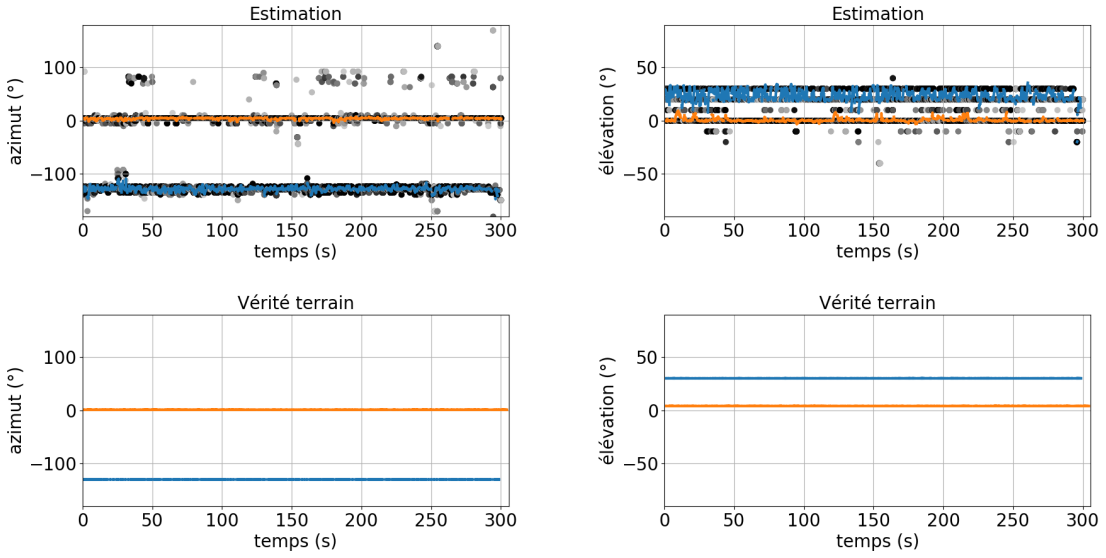


FIGURE 5.7. – Exemple de suivi de sources par filtrage particulaire sur une scène sonore. Gauche/droite : azimut/élévation. Haut/bas : estimation/vérité terrain. L'intensité des points gris est proportionnelle au score renvoyé par le réseau de neurones.

Le réseau de neurones attribue des probabilités non nulles à la télévision. Cependant, elles

ne sont pas assez élevées et surtout leur présence est trop parcimonieuse dans le temps pour que le module de filtrage particulaire y associe une source. On constate également l'effet de la grille d'estimation du réseau de neurones sur les estimations, en particulier pour l'élévation qui saute incessamment entre 20 et 30° pour la source bleue.

5.4.4. Résultats de reconnaissance de la parole

Les résultats de RAP sont présentés dans le Tableau 5.6. Dans cette situation, nous n'avons pas accès au signal de parole non réverbéré, et par conséquent pas au masque idéal $M_s^{(\mathrm{id})}$ ni au filtre qui en est issu.

Localisation	Vérité terrain	CRNN + suivi
Source image s_W	9,4	
Mélange x_W	89,8	
Formation voie FOA $\hat{s}$	50,3	48,8
masque M_s	79,7	81,3
filtre $\mathbf{w}_{\mathrm{MWF\text{-}r1}}$	14,9	19,3

TABLEAU 5.6. – WER pour des enregistrements réels en utilisant les directions d'arrivées réelles ou estimées. Les écarts angulaires entre les sources sont compris entre 25 et 90°. Les résultats sont indiqués avec un intervalle de confiance à 95% de $\pm 1,0\%$.

Lorsque le système de rehaussement utilise la véritable direction d'arrivée des sources, le WER de 14,9% est comparable à celui de 15,9% obtenu sur le corpus créé à partir de SRIRs mesurées, pour des écarts entre les sources similaires (voir Tableau 5.2). Il faut toutefois garder en mémoire que les corpus de parole sont différents, cette comparaison est donc indicative.
Par ailleurs, lorsque les directions d'arrivée sont estimées avec le CRNN de localisation couplé au suivi par filtrage particulaire, le WER est de 19,3%. L'augmentation est notable, ce qui met en lumière l'importance d'une localisation précise des sources. Cependant, l'amélioration par rapport à une simple formation de voie est significative. Au lieu d'une transcription incompréhensible, ce système permet de se rapprocher d'une transcription permettant de reconstituer des phrases et d'interpréter des ordres.

5.5. Résumé

Dans ce chapitre, nous proposons un système de rehaussement de la parole d'un locuteur cible adapté au cas multi-locuteurs. Il s'appuie sur une estimation préalable des signaux de parole de chaque locuteur grâce à un filtre de formation de voie pleine bande, ce qui nécessite de connaître les directions d'arrivée des sources. Un réseau LSTM estime ensuite un masque temps-fréquence correspondant au locuteur cible. Enfin, afin d'éviter les distorsions inhérentes à l'application du masque (qui s'avèrent très pénalisantes pour la RAP), nous utilisons ces masques pour calculer les matrices de covariance des signaux, dont nous dérivons un filtre de Wiener multicanal de rang 1, qui est supérieur aux autres

filtres spatiaux testés. Ce système est utilisé pour des signaux au format ambisonique, mais le principe peut être adapté à d'autres types de signaux.
Nous montrons que dans les cas difficiles, lorsque les deux locuteurs sont proches et parlent à volume égal, il est indispensable de fournir au réseau l'estimation de la locutrice concurrente en plus de celle du locuteur cible. Nous montrons également qu'il suffit d'apprendre le réseau sur des mélanges où les deux sources sont à niveau égal pour que celui-ci soit en mesure de généraliser à des configurations où les sources ont des niveaux différents. Nous vérifions la robustesse de ce système aux erreurs d'estimation des directions d'arrivée des sources, condition indispensable pour son utilisation en conditions réelles. Enfin, nous assemblons localisation, suivi de sources et rehaussement sur des enregistrements réels en conditions difficiles, et nous montrons que le module de localisation est suffisamment précis pour que les résultats du système de rehaussement soient exploitables pour la RAP.

6. Conclusion et perspectives

Dans ce chapitre, nous résumons les contributions proposées dans le cadre de cette thèse et proposons des suites à ce travail de recherche.

6.1. Bilan

6.1.1. Contexte et prise de position

Cette thèse s'inscrit à l'intersection de deux domaines d'application en forte expansion. D'une part, les assistants vocaux changent de forme : après le téléphone, ils s'incarnent aujourd'hui dans les enceintes connectées. Cela soulève nombre de défis techniques liés à la RAP en champ lointain qui implique plus de réverbération, plus de bruits parasites, et potentiellement la présence de locuteurs concurrents parlant à un volume comparable voire supérieur à celui de l'utilisatrice. Les dispositifs existants ne sont pas robustes à ces conditions ; cette thèse s'intéresse donc aux moyens permettant de faciliter la reconnaissance vocale d'un locuteur cible. D'autre part, profitant de l'engouement pour l'audio spatialisée, le format ambisonique est passé d'un sujet de recherche de laboratoire à un format utilisé dans l'industrie, en particulier dans sa version d'ordre 1, le FOA. Ce format met en valeur l'aspect spatial du champ sonore. Nous avons donc cherché à faciliter la reconnaissance vocale à partir de contenus FOA.
D'un point de vue technique, nous avons mis au point un système de rehaussement de la parole indépendant du système de reconnaissance. Comme nous le montrons dans l'état de l'art (voir la partie 3.1), c'est actuellement la façon la plus efficace de faciliter la reconnaissance en conditions adverses. Cela permet également de concevoir un système de rehaussement adapté au format ambisonique, tout en utilisant n'importe quel système de reconnaissance (qui peut, par exemple, être adapté en fonction du contexte sémantique ou de la langue utilisée).
Nous nous plaçons dans le cadre de rehaussement de la parole le plus performant à l'heure actuelle (voir la partie 3.2.4.1) : le filtrage multicanal utilisant un masque temps-fréquence. Ce masque est estimé par un réseau de neurones [49]. Comme les travaux précédents, nous utilisons pour cela un réseau LSTM. Cependant, les systèmes proposés jusqu'ici dans ce cadre ne s'appliquent qu'au cas d'un locuteur dont la parole est corrompue par du bruit ambiant très différent de la parole. En présence de plusieurs locuteurs, il est impossible au réseau d'identifier le locuteur cible sans information supplémentaire. Nous proposons d'utiliser l'information spatiale présente dans le format ambisonique, sous la forme de la direction d'arrivée de chaque source, pour lever cette ambiguïté.
Nous nous sommes donc également intéressés au cours de ce travail à l'estimation de directions d'arrivée de sources sonores à partir de contenus ambisoniques. Les systèmes les plus

performants s'appuient aujourd'hui sur des CRNNs. Ceux-ci permettent d'extraire une information de localisation à partir d'une paramétrisation bas niveau du champ acoustique, en général une représentation temps-fréquence des signaux capturés par l'antenne de microphones. Afin de traiter simplement le cas multi-locuteurs, la plupart des travaux formulent la localisation comme un problème de classification, en considérant une discrétisation de la sphère unité (voir la partie 3.3.4). Nous avons également proposé et/ou évalué plusieurs formulations par régression, mais la classification permet de traiter plus simplement le cas multi-sources. Nous utilisons donc également cette formulation dans la plus grande partie de ce travail. Les techniques de localisation mises au point jusqu'ici ne sont pas robustes aux situations réelles, notamment lorsque les sources sont directives ou que des meubles génèrent des réflexions précoces importantes qui perturbent la localisation. La solution que nous proposons s'appuie sur une large base d'apprentissage simulée et une paramétrisation des données d'entrée adaptée permettant de surmonter ces difficultés.

6.1.2. Contributions

Les contributions de cette thèse consistent principalement à avoir mis au point des modules de pré-traitement avant la RAP : tout d'abord un module de localisation de sources sonores, puis un module de rehaussement de la parole cible. Les deux modules sont adaptés aux contenus FOA.
Dans le chapitre 4, nous présentons le module d'estimation de directions d'arrivée mis au point pour une à deux sources. Nous utilisons un CRNN de classification qui estime les directions d'arrivée sur une discrétisation de la sphère unité. L'architecture convolutif de ce réseau permet d'extraire des informations haut niveau à partir d'une paramétrisation d'entrée très différente de la cible, tandis que l'aspect récurrent permet de traiter la dimension séquentiel des signaux. Afin de permettre la généralisation de ce réseau à une grande variété de salles, de positions des sources et de l'antenne de microphones, nous constituons une base de données à partir d'un grand nombre de SRIRs synthétisées par la méthode image, selon la procédure proposée dans l'algorithme 1. Au total, la base contient en moyenne 100 salles différentes et 300 SRIRs par classe (c'est-à-dire direction d'arrivée), sachant que les directions d'arrivée des SRIRs ne correspondent pas nécessairement exactement à celles des classes. Nous montrons dans la partie 4.5.1 qu'une telle taille et variété de la base de données est une condition nécessaire à la généralisation. D'autre part, afin de permettre au réseau de traiter les conditions acoustiques difficiles précédemment citées, nous proposons une nouvelle paramétrisation des données d'entrée. Elle utilise le vecteur d'intensité acoustique, grandeur dont la partie réelle (appelée partie active) est directement liée à la direction de propagation du flux d'énergie. De par la simplicité de sa formulation à partir de contenus FOA, le vecteur d'intensité acoustique actif a déjà été utilisé pour la localisation de sources au format ambisonique [121], mais jamais dans le contexte des réseaux de neurones. D'autre part, la partie imaginaire (dite réactive) de cette grandeur n'a jamais été utilisée jusqu'à présent, certainement en raison de son aspect très aléatoire, trop bruité pour permettre l'extraction d'information. Néanmoins cette composante est liée au caractère diffus du champ acoustique. Aussi, nous

proposons de fournir au réseau les représentations temps-fréquence des parties active et réactive du vecteur d'intensité. Nous montrons que cela permet d'obtenir un réseau bien plus performant que s'il avait été appris uniquement sur les spectrogrammes des signaux FOA (voir le Tableau 4.2). Nous montrons également dans les Tableaux 4.3 et 4.4 que la partie réactive du vecteur d'intensité apporte un véritable gain de performance, en particulier sur les ensembles de test réels.
Nous testons également diverses architectures de réseau et diverses normalisations de la paramétrisation d'entrée. Dans le cas où une seule source est présente, nous formulons la localisation comme un problème de régression et comparons différentes fonctions de coût pour la régression et pour la classification. En particulier, nous proposons pour la classification une paramétrisation de la sortie et une fonction de coût qui prennent en compte l'aspect ordonné de l'ensemble des directions d'arrivée. Nous montrons qu'en estimant les coordonnées cartésiennes de la source sur la sphère unité, ce qui permet d'éviter la cyclicité des coordonnées sphériques, les performances de localisation sont parfois meilleures qu'avec un réseau de classification. Cependant, cette solution est difficile à étendre à un nombre variable de sources, puisque le nombre de sorties du réseau de régression dépend du nombre de sources.
Dans l'optique de mieux comprendre le comportement du réseau de neurones et pour s'assurer de sa fiabilité et du bien-fondé de ses estimations, nous utilisons une technique de visualisation appelée LRP. D'autres techniques sont présentées dans la partie 3.4. Nous y justifions également l'importance de procéder à de telles analyses sur les réseaux de neurones. Dans la partie 4.4, la LRP nous permet de constater que le réseau de localisation s'appuie sur les attaques des sons, mises en valeur par le vecteur d'intensité acoustique. Cela est cohérent avec le phénomène de précédence déjà observé pour l'audition humaine. D'autre part, on observe de façon cohérente que le réseau s'appuie sur les zones temps-fréquence où la source à localiser est plus puissante que les autres sources.
Le chapitre 5, quant à lui, présente le module de rehaussement de la parole proposé pour les signaux ambisoniques dans le cas où deux locuteurs parlent simultanément. Nous utilisons un réseau LSTM afin d'estimer le masque temps-fréquence correspondant à la parole cible. Afin que le réseau puisse différencier la cible du signal interférent, nous proposons de lui fournir des estimations monocanales de chacun de ces signaux en plus du premier canal (omnidirectionnel) de la captation FOA. Les estimations résultent d'un filtre de formation de voie pleine bande ambisonique dont le calcul est simple dès lors que les directions d'arrivée des sources sont connues. Nous montrons que dans les cas les plus difficiles, lorsque les sources sont proches spatialement (à 25° d'écart environ), il est indispensable de fournir au réseau les estimations des deux signaux, et pas seulement de la cible (voir partie 5.3.1). À partir des masques temps-fréquence, il est possible d'estimer les matrices de covariance du signal cible et des interférences, et donc d'en obtenir des estimations multicanales. Celles-ci permettent de calculer un filtre de Wiener multicanal. En particulier, nous en choisissons une approximation de rang 1, le MWF-r1 (voir la partie 3.2.2). Nous montrons dans la partie 5.3.3 que ce filtre est plus adapté à la reconnaissance vocale que le max-SNR largement plébiscité [49]. Nous vérifions également la robustesse du système à des variations de niveau entre les sources ou à des erreurs sur

les directions d'arrivée estimées.
Enfin, dans une dernière expérience (partie 5.4), nous appliquons toute la chaîne de traitement à des signaux enregistrés en conditions réelles. Afin d'assurer que la source cible reste identifiée au cours du temps, nous couplons le réseau de localisation à un module de suivi de sources pré-existant. Nous appliquons ensuite le module de rehaussement en utilisant soit la vraie direction d'arrivée des sources soit la direction estimée. Nous observons qu'il existe encore une différence de performance de RAP entre les deux, mais les résultats obtenus en considérant les directions d'arrivée estimées sont tout de même très encourageants.

6.1.3. Publications

Les travaux de cette thèse ont donné lieu aux publications suivantes.
Article de revue :

— L. Perotin, R. Serizel, E. Vincent et A. Guérin, « CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings », IEEE JSTSP, vol. 13, no. 1, pp. 22–33, 2019.

Articles de conférence :

— L. Perotin, R. Serizel, E. Vincent et A. Guérin, « Multichannel speech separation with recurrent neural networks from high-order Ambisonics recordings », in Proc. of ICASSP, 2018, pp. 36–40.

— L. Perotin, R. Serizel, E. Vincent et A. Guérin, « CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector », in Proc. of IWAENC, 2018, pp. 241–245.

— L. Perotin, A. Défossez, E. Vincent R. Serizel, et A. Guérin, « Regression versus classification for neural network based audio source localization », à paraître, in Proc. of WASPAA, 2019.

Présentations en congrès :

— « Speech separation by neural network », Forum des jeunes mathématicie·ne·s, Nancy, France, 2017.

— « Multichannel RNN-based separation of overlapping speech », LISTEN Workshop, Bonn, Allemagne, 2018.

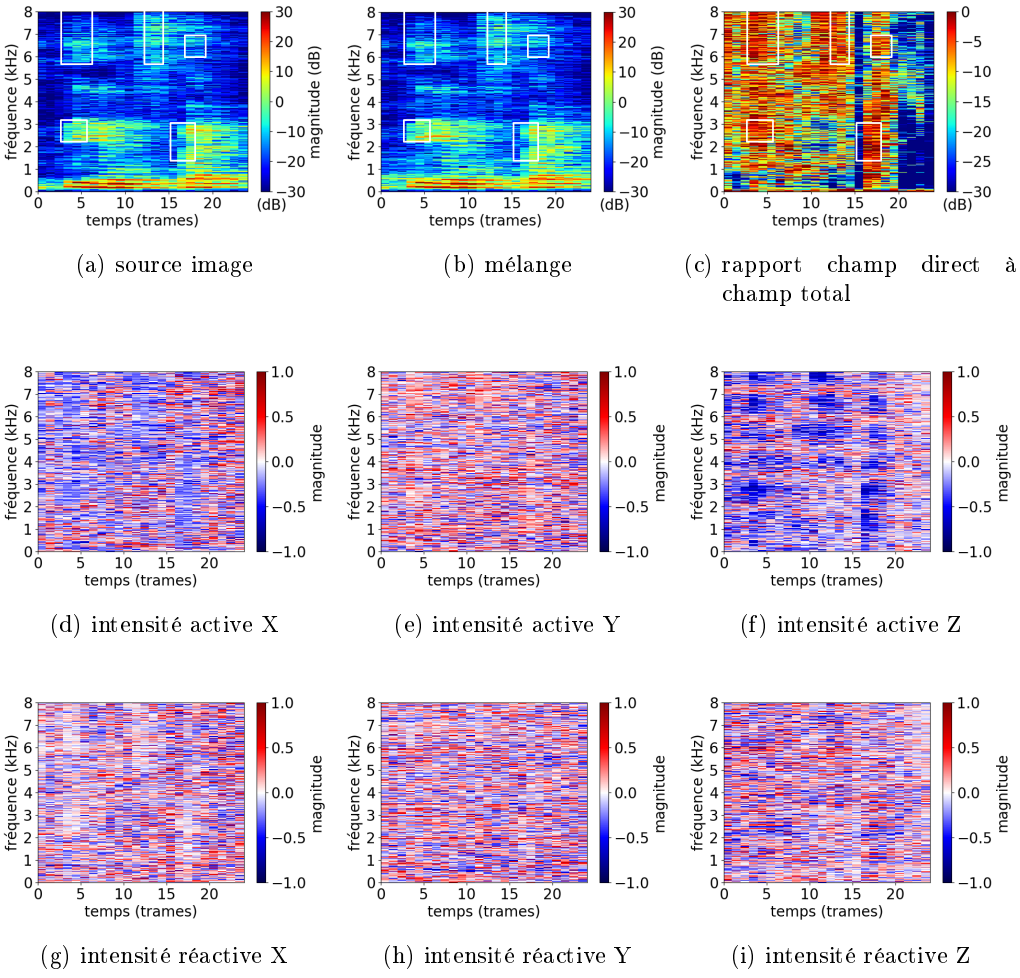
6.2. Pistes de poursuite

Les travaux présentés dans cette thèse et les limitations rencontrées peuvent appeler à divers développements.
Le réseau de localisation présenté fonctionne théoriquement pour un nombre quelconque de sources. Sans module de suivi de sources, il est nécessaire que ce nombre soit connu, ou bien de définir un hyper-paramètre de seuillage permettant de définir si un score est suffisamment élevé pour correspondre à une source. Ce paramètre est implicitement inclus

dans les réglages du module de suivi, qui décide de l'apparition ou de la disparition de sources en fonction des sorties du réseau et du contexte. Par la suite, on pourra tester le comportement du réseau de localisation avec ou sans suivi lorsque le nombre de sources varie et dépasse potentiellement 2, et chercher à l'améliorer.
Par ailleurs, les façons d'exploiter la régression dans le cas multi-sources méritent d'être explorées. En considérant connu le nombre maximal de sources, il est possible d'utiliser une seule architecture quel que soit le nombre de sources actives à un instant donné, même si celui-ci est inconnu. On peut par exemple renvoyer les coordonnées (0,0,0) s'il y a moins de sources détectées que le nombre maximal. Dans ce cas, il faut traiter avec soin le problème de permutation d'étiquettes lors de l'apprentissage, en s'inspirant par exemple de solutions proposées pour le rehaussement de la parole comme l'apprentissage invariant par permutation [99]. L'intégration du module de localisation avec un module de suivi de sources sera également particulièrement importante.
En ce qui concerne le rehaussement, traiter le cas avec plus de deux sources soulève de nouvelles questions. Il est possible d'entraîner des réseaux différents pour chaque nombre de sources, avec un nombre de paramètres d'entrées variant en fonction du nombre de sources. En effet, si l'on considère acquises la localisation et le suivi des sources, le nombre de sources est connu à cet étape du processus. Cependant, pour éviter de stocker différents réseaux, on peut également envisager un seul réseau de rehaussement. Pour plusieurs locuteurs interférents, il faut alors combiner en entrée les estimations effectuées par formation de voie pleine bande pour chacun d'entre eux, par exemple en les sommant. Des tests sont nécessaires afin de savoir si l'architecture proposée est suffisamment puissante pour traiter ces situations intrinsèquement plus difficiles. Si ce n'est pas le cas, il sera possible d'explorer de nouvelles architectures de réseau, à la fois capable de traiter des données temporelles et d'extraire des informations à partir de données bruitées. Les réseaux convolutionnels, en particulier s'ils utilisent des convolutions dilatées, se sont parfois montrés plus adaptés que les LSTMs pour traiter des signaux temporels [189]. Puisque la sortie du réseau est un masque temps-fréquence, comparable aux spectrogrammes d'entrée, une architecture de type U-net pourrait s'avérer particulièrement adaptée. Elle extrait des informations de haut niveau en réduisant la dimension de la représentation, puis étend de nouveau la représentation jusqu'à des dimensions comparables à celles d'entrée. Grâce à des ponts reliant les étapes d'encodage aux étapes de décodage, le U-net permet d'estimer une représentation du signal avec une précision fine. On peut également envisager d'utiliser une architecture U-net incluant des convolutions dilatées en temps ou en fréquence afin de prendre un compte un contexte temporel plus large ou la structure harmonique de la parole.
Jusqu'ici, nous considérons des sources immobiles. Dans la réalité, les locuteurs peuvent tout à fait se déplacer en énonçant un ordre. Le système proposé traitant le signal par séquences de 832 ms qui peuvent se chevaucher, rien ne s'oppose en théorie à traiter le cas de sources mouvantes. Cependant, le filtre MWF-r1 utilisé pour des sources immobiles est lissé temporellement sur des durées longues (de l'ordre de la minute), ce qui permet de limiter de manière cruciale les distorsions du signal rehaussé. Si les sources bougent, il est nécessaire de réduire drastiquement les périodes de lissage. Dans ce cas,

nous avons constaté que le moteur de reconnaissance fait face à d'importantes difficultés pour comprendre le signal rehaussé. Il s'agirait donc de mettre au point une méthode pour optimiser le lissage temporel du filtre multicanal, voir le filtre lui-même, en fonction du signal [95], mais aussi de l'objectif final, à savoir la transcription.
Dans l'optique de pouvoir utiliser les solutions proposées dans un cas réel, il est nécessaire de définir une méthode pour identifier le locuteur cible. Cela peut être fait par l'intermédiaire d'un mot-clé prononcé au début de la scène sonore ou par l'enrôlement préalable du locuteur. Cela implique d'utiliser un nouveau réseau capable de détecter ce mot-clé y compris en conditions bruitées et de localiser la source correspondante. Il est aussi possible de procéder au traitement entier sans savoir quelle source est la cible, et donc d'appliquer les systèmes de rehaussement et de RAP tour à tour pour chaque source détectée. Le mot-clé peut ensuite être reconnu dans les transcriptions finales, mais cette méthode est coûteuse en temps de calcul. Enfin, une troisième solution consisterait à intégrer la détection du mot-clé au module de suivi de sources, qui renverrait alors toujours la source cible en premier. Pour cela, une solution envisageable serait de procéder au suivi de sources avec un réseau de neurones. Cela soulève de nouveaux défis en terme de constitution de bases de données de sources mobiles, de formulation du problème et d'évaluation de la performance d'un module de suivi.

A. Paramétrisation pour la localisation



(a) source image (b) mélange (c) rapport champ direct à champ total

(d) intensité active X (e) intensité active Y (f) intensité active Z

(g) intensité réactive X (h) intensité réactive Y (i) intensité réactive Z

FIGURE A.1. – Exemple de paramétrisation selon (4.5) d'un signal contenant une source ponctuelle venant de (139°, -61°) et du bruit de foule diffus à 18 dB SNR. (a) spectrogramme de l'image spatiale de la source. (b) spectrogramme du mélange. (c) Rapport champ direct à champ total. (d), (e) et (f) : vecteur d'intensité active normalisé. (g), (h) et (i) : vecteur d'intensité réactive normalisé.

B. Tirages aléatoires pour la génération de SRIRs

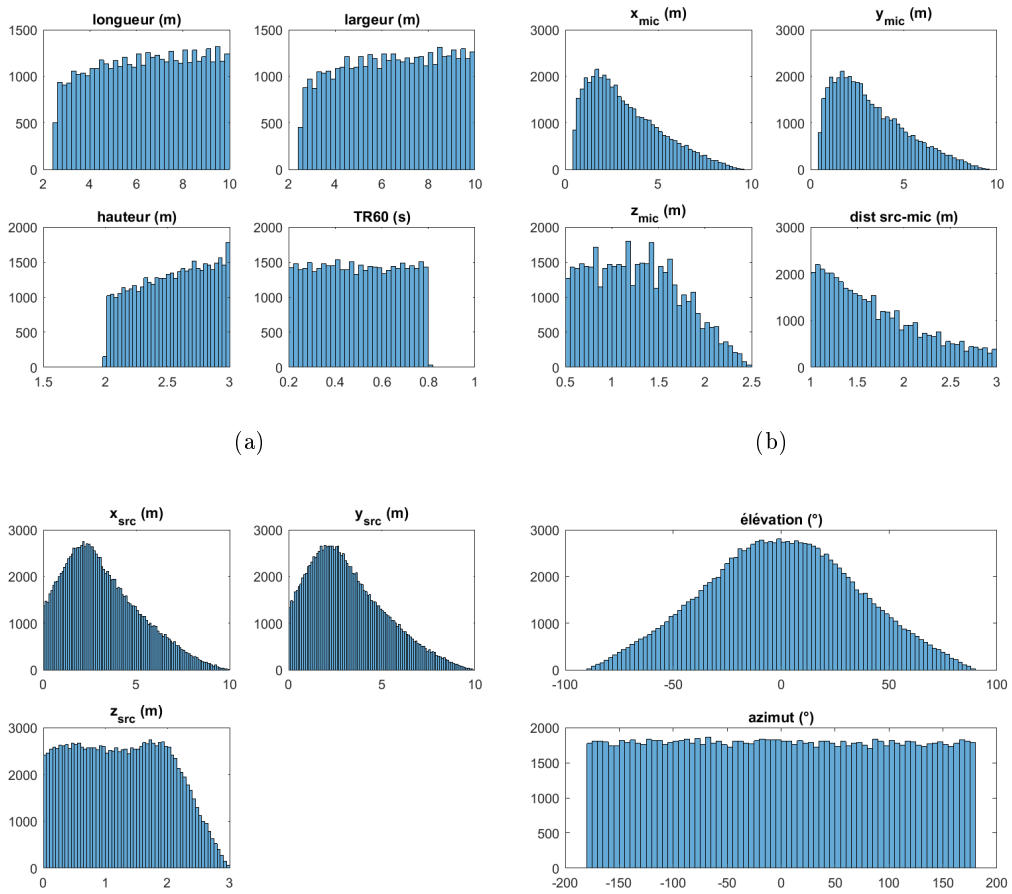
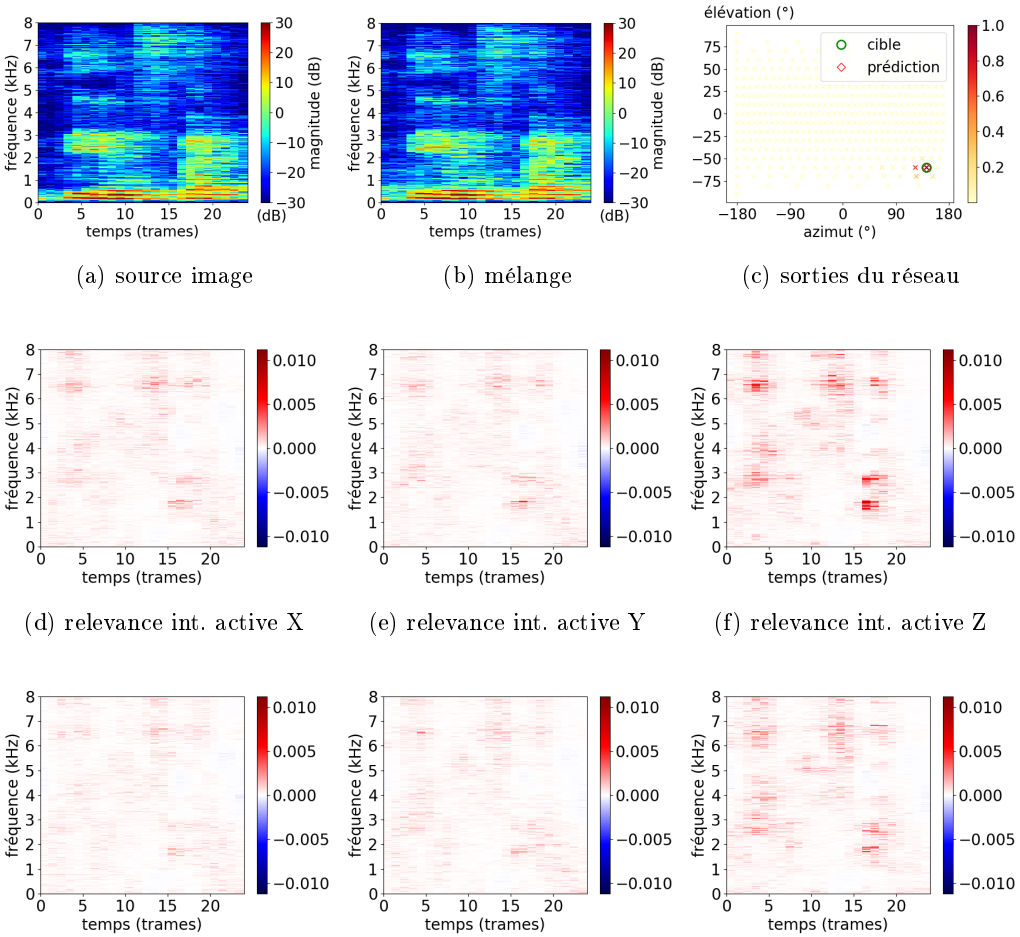


FIGURE B.1. – Histogrammes des tirages aléatoires pour la génération de la base de SRIRs pour l'apprentissage. (a) Dimensions des salles, (b) positions des antennes de microphones, (c) positions des sources, (d) directions d'arrivée des sources.

C. Résultats de la LRP canal par canal



(a) source image (b) mélange (c) sorties du réseau

(d) relevance int. active X (e) relevance int. active Y (f) relevance int. active Z

(g) relevance int. réactive X (h) relevance int. réactive Y (i) relevance int. réactive Z

FIGURE C.1. – LRP détaillée par canaux pour la prédiction de la direction d'arrivée d'un signal contenant une source ponctuelle venant de (139°, -61°) et du bruit de foule diffus à 18 dB SNR (voir Annexe A). (a) Spectrogramme de l'image spatiale de la source. (b) spectrogramme du mélange. (c) Prédiction du réseau pour ce signal. (d), (e) et (f) : relevances pour les canaux X, Y et Z du vecteur intensité actif normalisé. (g), (h) et (i) : relevances pour les canaux X, Y et Z du vecteur intensité réactif normalisé.

Bibliographie

[1] L. Deng and D. Yu. *Deep learning : methods and applications.* NOW, 2014.

[2] J. Heymann, M. Bacchiani, and T. N. Sainath. Performance of mask based statistical beamforming in a smart home scenario. In *Proc. of ICASSP*, pages 6722–6726, 2018.

[3] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties. MPEG-H 3d audio - The new standard for coding of immersive spatial audio. *IEEE JSTSP*, 9(5) :770–779, 2015.

[4] E. G. Williams and J. A. Mann. *Fourier acoustics : sound radiation and nearfield acoustical holography*, volume 108. Academic Press, 2000.

[5] P. M. Morse and K. U. Ingard. *Theoretical acoustics.* Princeton University Press, 1986.

[6] J. Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia.* Thèse de doctorat, Univ. Paris VI, 2000.

[7] B. Rafaely. *Fundamentals of spherical array processing*, volume 8 of *Springer Topics in Signal Processing.* Springer Berlin Heidelberg, 2015.

[8] S. Moreau. *Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore HOA.* Thèse de doctorat, Univ. du Maine, 2006.

[9] M. A. Gerzon. Periphony : with-height sound reproduction. *JAES*, 21(1) :2–10, 1973.

[10] Mh Acoustics. EM32 Eigenmike microphone array release notes (v17. 0). Technical report, 2013.

[11] V. Pulkki, S. Delikaris-Manias, and A. Politis. *Parametric time-frequency domain spatial audio.* Wiley, 2017.

[12] M. Baqué. *Analyse de scène sonore multi-capteurs.* Thèse de doctorat, Univ. du Maine, 2017.

[13] N. Epain and J. Daniel. Improving spherical microphone arrays. In *Proc. of AES*, page 9, 2008.

[14] F. Jacobsen. A note on instantaneous and time-averaged active and reactive sound intensity. *J. of Sound and Vibration*, 147(3) :489–496, 1991.

[15] D. Yu and L. Deng. *Automatic speech recognition : a deep learning approach.* Signals and communication technology. Springer, London, 2015. OCLC : 876005536.

[16] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE TASLP*, 22(4) :745–777, 2014.

[17] G. Hinton, L. Deng, D. Yu, G. E. Dahl, et al. Deep neural networks for acoustic modeling in speech recognition : the shared views of four research groups. *IEEE Sig. Proc. Mag.*, 29(6) :82–97, 2012.

[18] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. Recurrent neural network based language model. In *Proc. Interspeech*, pages 1045–1048, 2010.

[19] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proc. of ICML*, pages 1764–1772, 2014.

[20] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux. End-to-end speech recognition from the raw waveform. In *Proc. of Interspeech*, 2018.

[21] J. Barker, S. Watanabe, E. Vincent, and J. Trmal. The fifth 'CHiME' speech separation and recognition challenge : dataset, task and baselines. In *Proc. Interspeech*, pages 1561–1565, 2018.

[22] M. Vacher, E. Vincent, M.-E. Bobillier Chaumon, T. Joubert, et al. The VocADom project : speech interaction for well-being and reliance improvement. In *Proc. of MobileHCI*, 2018.

[23] M. L. Seltzer, D. Yu, and Y. Wang. An investigation of deep neural networks for noise robust speech recognition. In *Proc. of ICASSP*, pages 7398–7402, 2013.

[24] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, et al. A summary of the REVERB challenge : state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP J. Adv. Signal Process.*, 2016(1) :7, 2016.

[25] D. Yu, L. Deng, J. Droppo, J. Wu, et al. A minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition. In *Proc. of ICASSP*, pages 4041–4044, 2008.

[26] J Barker, R. Marxer, E. Vincent, and S. Watanabe. The third 'CHiME' speech separation and recognition challenge : analysis and outcomes. *Computer Speech & Language*, 46 :605–626, 2017.

[27] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46 :535–557, 2017.

[28] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. Morales-Cordovilla, et al. Robust ASR using neural network based speech enhancement and feature simulation. In *Proc. of ASRU*, pages 482–489, 2015.

[29] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani. Context adaptive deep neural networks for fast acoustic model adaptation. In *Proc. of ICASSP*, pages 4535–4539. IEEE, 2015.

[30] B. King, I-F. Chen, Y. Vaizman, Y. Liu, et al. Robust speech recognition via anchor word representations. In *Proc. of Interspeech*, pages 2471–2475. ISCA, 2017.

[31] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach. Beamnet : end-to-end training of a beamformer-supported multi-channel ASR system. In *Proc. of ICASSP*, pages 5325–5329, 2017.

[32] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao. Unified architecture for multichannel end-to-end speech recognition with neural beamforming. *IEEE JSTSP*, 11(8) :1274–1288, 2017.

[33] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, et al. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE TASLP*, 25(5) :965–979, 2017.

[34] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, et al. The NTT CHiME-3 system : advances in speech enhancement and recognition for mobile multi-microphone devices. In *Proc. of ASRU*, pages 436–443, 2015.

[35] J. Barker, R. Marxer, E. Vincent, and S. Watanabe. The CHiME challenges : robust speech recognition in everyday environments. In *New ere for robust speech recognition - Exploiting deep learning*, pages 327–344. Springer, 2017.

[36] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov. A consolidated perspective on multi-microphone speech enhancement and source separation. *IEEE TASLP*, 25(4) :692–730, 2017.

[37] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE TASLP*, 14(4) :1462–1469, 2006.

[38] E. Vincent, T. Virtanen, and S. Gannot. *Audio source separation and speech enhancement*. Wiley, 2018.

[39] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Sig. Proc.*, 52(7) :1830–1847, 2004.

[40] T. Virtanen, E. Vincent, and S. Gannot. Time-frequency processing - Spectral properties. In *Audio source separation and speech enhancement*. Wiley, 2018.

[41] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Trans. Speech and Audio Proc.*, 11(2) :109–116, 2003.

[42] D.-T. Pham, Z. El-Chami, A. Guérin, and C. Servière. Modeling the short time Fourier transform ratio and application to underdetermined audio source separation. In *Proc. of ICA*, pages 98–105, 2009.

[43] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode. Adaptive antenna systems. *Proceedings of the IEEE*, 55(12) :2143–2159, 1967.

[44] S. Markovich-Golan, W. Kellermann, and S. Gannot. Spatial filtering. In *Audio source separation and speech enhancement*. Wiley, 2018.

[45] S. Doclo, A. Spriet, J. Wouters, and M. Moonen. Speech distortion weighted multichannel Wiener filtering techniques for noise reduction. In *Speech Enhancement*, Signals and Communication Technology, pages 199–228. Springer, 2005.

[46] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proc. of the IEEE*, 57(8) :1408–1418, 1969.

[47] S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. on Sig. Proc.*, 49(8) :1614–1626, 2001.

[48] E. Warsitz and R. Haeb-Umbach. Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE TASLP*, 15(5) :1529–1539, 2007.

[49] J. Heymann, L. Drude, and R. Haeb-Umbach. Neural network based spectral mask estimation for acoustic beamforming. In *Proc. of ICASSP*, pages 196–200, 2016.

[50] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters. Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE TASLP*, 22(4) :785–799, 2014.

[51] Z. Wang, E. Vincent, R. Serizel, and Y. Yan. Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments. *Computer Speech & Language*, 49 :37–51, 2018.

[52] E. A. P. Habets and P. A. Naylor. Dereverberation. In *Audio source separation and speech enhancement*, pages 317–343. Wiley, 2018.

[53] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, et al. Making machines understand us in reverberant rooms : robustness against reverberation for automatic speech recognition. *IEEE Sig. Proc. Mag.*, 29(6) :114–126, 2012.

[54] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor. Robust multichannel dereverberation using relaxed multichannel least squares. *IEEE TASLP*, 22(9) :1379–1390, 2014.

[55] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin. Variational bayesian inference for multichannel dereverberation and noise reduction. *IEEE TASLP*, 22(8) :1320–1335, 2014.

[56] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE TASLP*, 18(7) :1717–1731, 2010.

[57] K. Lebart, J. M. Boucher, and P. N. Denbigh. A new method based on spectral subtraction for speech dereverberation. *Acta Acustica united with Acustica*, 87(3) :359–366, 2001.

[58] E. A. P. Habets, S. Gannot, and I. Cohen. Late reverberant spectral variance estimation based on a statistical model. *IEEE Sig. Proc. Letters*, 16(9) :770–773, 2009.

[59] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets. An informed spatial filter for dereverberation in the spherical harmonic domain. In *Proc. of ICASSP*, pages 669–673, 2013.

[60] J. Allen. Synthesis of pure speech from a reverberant signal, 1974.

[61] K. Han, Y. Wang, D. Wang, W. S. Woods, et al. Learning spectral mapping for speech dereverberation and denoising. *IEEE TASLP*, 23(6) :982–992, 2015.

[62] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings F (Radar and Signal Processing)*, 140(6) :362–370, 1993.

[63] A. Hyvärinen and E. Oja. Independent component analysis : algorithms and applications. *Neural Networks*, 13(4–5) :411–430, 2000.

[64] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Trans. on Sig. Proc.*, 45(2) :434–444, 1997.

[65] D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Trans. on Sig. Proc.*, 49(9) :1837–1848, 2001.

[66] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1-3) :21–34, 1998.

[67] S. Ikeda and N. Murata. A method of blind separation on temporal structure of signals. In *Proc. of ICONIP*, volume 98, pages 737–742. IOS Press, 1998.

[68] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. Speech and Audio Proc.*, 12(5) :530–538, 2004.

[69] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis. Model-based expectation-maximization source separation and localization. *IEEE TASLP*, 18(2) :382–394, 2010.

[70] C. Fevotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models. In *Proc. of WASPAA*, pages 78–81, 2005.

[71] E. Vincent, S. Arberet, and R. Gribonval. Underdetermined instantaneous audio source separation via local Gaussian modeling. In *SpringerLink*, pages 775–782, 2009.

[72] N. Q. K. Duong, E. Vincent, and R. Gribonval. Spatial covariance models for under-determined reverberant audio source separation. In *Proc. of WASPAA*, pages 129–132, 2009.

[73] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE TASLP*, 18(7) :1830–1840, 2010.

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

[75] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. of WASPAA*, pages 177–180, 2003.

[76] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE TASLP*, 15(1) :1–12, 2007.

[77] G. J. Mysore and P. Smaragdis. A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. In *Proc. of ICASSP*, pages 17–20, 2011.

[78] A. Ozerov and C. Fevotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE TASLP*, 18(3) :550–563, 2010.

[79] Y. Wang and D. Wang. Towards scaling up classification-based speech separation. *IEEE TASLP*, 21(7) :1381–1390, 2013.

[80] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *Proc. of ICASSP*, pages 1562–1566, 2014.

[81] F. Weninger, F. Eyben, and B. Schuller. Single-channel speech separation with memory-enhanced recurrent neural networks. In *Proc. of ICASSP*, pages 3709–3713, 2014.

[82] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller. Discriminatively trained recurrent neural networks for single-channel speech separation. In *Proc. of GlobalSIP*, pages 577–581, 2014.

[83] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proc. of ICASSP*, pages 708–712, 2015.

[84] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, et al. Exploring multi-channel features for denoising-autoencoder-based speech enhancement. In *Proc. of ICASSP*, pages 116–120, 2015.

[85] S. Pascual, A. Bonafonte, and J. Serrà. SEGAN : speech enhancement generative adversarial network. In *Proc. Interspeech*, pages 3642–3646, 2017.

[86] A. Narayanan and D. Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proc. of ICASSP*, pages 7092–7096, 2013.

[87] Y. Jiang, D. Wang, R. Liu, and Z. Feng. Binaural classification for reverberant speech segregation using deep neural networks. *IEEE TASLP*, 22(12) :2112–2121, 2014.

[88] P. Pertilä and J. Nikunen. Microphone array post-filtering using supervised machine learning for speech enhancement. In *Proc. of Interspeech*, pages 2675–2679, 2014.

[89] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann. Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments. In *Proc. of ICASSP*, pages 4380–4384, 2015.

[90] Y. Wang, A. Narayanan, and D. Wang. On training targets for supervised speech separation. *IEEE TASLP*, 22(12) :1849–1858, 2014.

[91] L. Perotin, R. Serizel, E. Vincent, and A. Guérin. Multichannel speech separation with recurrent neural networks from high-order Ambisonics recordings. In *Proc. of ICASSP*, pages 36–40, 2018.

[92] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In *Proc. of ICASSP*, pages 5210–5214, 2016.

[93] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, et al. Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In *Proc. of Interspeech*, pages 2655–2659, 2017.

[94] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, et al. Deep beamforming networks for multi-channel speech recognition. In *Proc. of ICASSP*, pages 5745–5749, 2016.

[95] A. A. Nugraha, A. Liutkus, and E. Vincent. Multichannel audio source separation with deep neural networks. *IEEE TASLP*, 24(9) :1652–1664, 2016.

[96] J. M. Festen and R. Plomp. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *JASA*, 88(4) :1725–1736, 1990.

[97] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE TASLP*, 23(12) :2136–2147, 2015.

[98] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo. Deep neural networks for single-channel multi-talker speech recognition. *IEEE TASLP*, 23(10) :1670–1679, 2015.

[99] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *Proc. of ICASSP*, pages 241–245, 2017.

[100] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva. Multi-microphone neural speech separation for far-field multi-talker speech recognition. In *Proc. of ICASSP*, pages 5739–5743, 2018.

[101] Y. C. Subakan and P. Smaragdis. Generative adversarial source separation. In *Proc. of ICASSP*, pages 26–30, 2018.

[102] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering : discriminative embeddings for segmentation and separation. In *Proc. of ICASSP*, pages 31–35, 2016.

[103] Z. Chen, Y. Luo, and N. Mesgarani. Deep attractor network for single-microphone speaker separation. In *Proc. of ICASSP*, pages 246–250, 2017.

[104] Z.-Q. Wang, J. L. Roux, and J. R. Hershey. Multi-channel deep clustering : discriminative spectral and spatial embeddings for speaker-independant speech separation. In *Proc. of ICASSP*, pages 1–5, 2018.

[105] P. Pertilä and J. Nikunen. Distant speech separation using predicted time-frequency masks from spatial features. *Speech Communication*, 68 :97–106, 2015.

[106] P. Pertilä, A. Brutti, P. Svaizer, and M. Omologo. Multichannel source activity detection, localization, and tracking. In *Audio source separation and speech enhancement*, pages 47–64. Wiley, 2018.

[107] N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *JASA*, 114(4) :2236–2252, 2003.

[108] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, et al. The LOCATA challenge data corpus for acoustic source localization and tracking. In *Proc. of SAM*, pages 410–414, 2018.

[109] J. DiBiase, H. F. Silverman, and M. S. Brandstein. Robust localization in reverberant rooms. *SpringerLink*, pages 157–180, 2001.

[110] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoustics, Speech, Signal Process.*, 24(4) :320–327, 1976.

[111] M. S. Brandstein and H. F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proc. of ICASSP*, pages 375–378, 1997.

[112] R. Lebarbenchon, E. Camberlein, D. di Carlo, C. Gaultier, et al. Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge. In *arXiv :1812.05901*, 2018.

[113] C. Blandin, A. Ozerov, and E. Vincent. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. *Signal Processing*, 92(8) :1950–1960, 2012.

[114] R. Roy and T. Kailath. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoustics, Speech, Sig. Proc.*, 37(7) :984–995, 1989.

[115] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.*, 34(3) :276–280, 1986.

[116] O. Nadiri and B. Rafaeli. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE TASLP*, 22(10) :1494–1505, 2014.

[117] L. Madmoni, H. Beit-On, H. Morgenstern, and B. Rafaely. Description of algorithms for Ben-Gurion University submission to the LOCATA challenge. In *arXiv :1812.04942*, 2018.

[118] H. Sawada, R. Mukai, and S. Makino. Direction of arrival estimation for multiple source signals using independent component analysis. In *Proc. of Int. Symp. Sig. Proc. App.*, volume 2, pages 411–414, 2003.

[119] B. Loesch, S. Uhlich, and B. Yang. Multidimensional localization of multiple sound sources using frequency domain ICA and an extended state coherence transform. In *IEEE Workshop on Stat. Sig. Proc.*, pages 677–680, 2009.

[120] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann. TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis. *IEEE TASLP*, 19(6) :1490–1503, 2011.

[121] S. Tervo. Direction estimation based on sound intensity vectors. In *Proc. EUSIPCO*, pages 700–704, 2009.

[122] S. Delikaris-Manias, D. Pavlidi, A. Mouchtaris, and V. Pulkki. DOA estimation with histogram analysis of spatially constrained active intensity vectors. In *Proc. of ICASSP*, pages 526–530, 2017.

[123] M. Hawkes and A. Nehorai. Wideband source localization using a distributed acoustic vector-sensor array. *IEEE Trans. on Sig. Proc.*, 51(6) :1479–1491, 2003.

[124] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz. Acoustic source separation of convolutive mixtures based on intensity vector statistics. *IEEE Trans. Speech and Audio Proc.*, 16(4) :748–756, 2008.

[125] D. Levin, E. A. P. Habets, and S. Gannot. On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields. *JASA*, 128(4) :1800–1811, 2010.

[126] C. Dimoulas, G. Kalliris, K. Avdelidis, and G. Papanikolaou. Improved localization of sound sources using multi-band processing of ambisonic components. In *Proc. of AES Conv. 126*, pages 1–11, 2009.

[127] S. Kitić and A. Guérin. TRAMP : Tracking by a Real-time AMbisonic-based Particle filter. In *arXiv :1810.04080*, 2018.

[128] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor. 3d source localization in the spherical harmonic domain using a pseudointensity vector. In *Proc. of EUSIPCO*, pages 442–446, 2010.

[129] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann. Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays. In *Proc. of ICASSP*, pages 117–120, 2011.

[130] N. Epain and C. Jin. Independent component analysis using spherical microphones arrays. *Acta Acustica united with Acustica*, 1(98) :91–102, 2012.

[131] J. Merimaa and V. Pulkki. Spatial impulse response rendering. In *Proc. of DAFx*, pages 139–144, 2004.

[132] A. H. Moore, C. Evers, and P. A. Naylor. Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors. *IEEE TASLP*, 25(1) :178–192, 2017.

[133] N. Roman, D. Wang, and G. J. Brown. A classification-based cocktail-party processor. In *Proc. of NIPS*, pages 1425–1432, 2004.

[134] K. W. Wilson and T. Darrell. Learning a precedence effect-like weighting function for the generalized cross-correlation framework. *IEEE TASLP*, 14(6) :2156–2164, 2006.

[135] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *JASA*, 65(4) :943–950, 1979.

[136] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *JASA*, 106(4) :1633–1654, 1999.

[137] T. Nishino and K. Takeda. Binaural sound localization for untrained directions based on a Gaussian mixture model. In *Proc. of EUSIPCO*, pages 1–5, 2008.

[138] T. May, S. Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE TASLP*, 19(1) :1–13, 2011.

[139] H. Kayser and J. Anemüller. A discriminative learning approach to probabilistic acoustic source localization. In *Proc. of IWAENC*, pages 99–103, 2014.

[140] A. Deleforge, F. Forbes, and R. Horaud. Acoustic space learning for sound-source separation and localization on binaural manifolds. *Int. J. Neur. Syst.*, 25(01) :1440003, 2014.

[141] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman. Diffusion maps for signal processing : a deeper look at manifold-learning techniques based on kernels and graphs. *IEEE Sig. Proc. Mag.*, 30(4) :75–86, 2013.

[142] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, et al. A learning-based approach to direction of arrival estimation in noisy and reverberant environments. In *Proc. of ICASSP*, pages 2814–2818, 2015.

[143] R. Takeda and K. Komatani. Discriminative multiple sound source localization based on deep neural networks using independent location model. In *IEEE SLT Work.*, pages 603–609, 2016.

[144] S. Chakrabarty and E. A. P. Habets. Broadband DOA estimation using convolutional neural networks trained with noise signals. In *Proc. of WASPAA*, pages 136–140, 2017.

[145] S. Chakrabarty and E. A. P. Habets. Multi-speaker localization using convolutional neural network trained with noise. In *ML4Audio Worskhop at NIPS*, 2017.

[146] N. Ma, T. May, and G. J. Brown. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE TASLP*, 25(12) :2444–2453, 2017.

[147] S. Adavanne, A. Politis, and T. Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *Proc. EUSIPCO*, 2018.

[148] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE JSTSP*, 13(1) :34–48, 2019.

[149] S. Chakrabarty and E. A. P. Habets. Multi-scale aggregation of phase information for reducing computational cost of CNN based DOA estimation. In *arXiv :1811.08552*, 2018.

[150] D. Comminiello, M. Lella, S. Scardapane, and A. Uncini. Quaternion convolutional neural networks for detection and localization of 3d sound events. In *Proc. of ICASSP*, pages 8533–8537, 2019.

[151] W. He, P. Motlicek, and J.-M. Odobez. Joint localization and classification of multiple sound sources using a multi-task neural network. In *Proc. Interspeech*, pages 312–316, 2018.

[152] Q. Nguyen, L. Girin, G. Bailly, F. Elisei, and D.-C. Nguyen. Autonomous sensorimotor learning for sound source localization by a humanoid robot. In *Proc. of IROS*, 2018.

[153] L. Perotin, R. Serizel, E. Vincent, and A. Guérin. CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector. In *Proc. of IWAENC*, pages 241–245, 2018.

[154] D. Salvati, C. Drioli, and G. L. Foresti. Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions. *IEEE Trans. Em. Topics Comput. Intell.*, 2(2) :103–116, 2018.

[155] S. Sivasankaran, E. Vincent, and D. Fohr. Keyword based speaker localization : Localizing a target speaker in a multi-speaker environment. In *Proc. Interspeech*, 2018.

[156] D. Suvorov, G. Dong, and R. Zhukov. Deep residual network for sound source localization in the time domain. *arXiv :1808.06429*, 2018.

[157] L. Perotin, R. Serizel, E. Vincent, and A. Guérin. CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings. *IEEE JSTSP*, 13(1) :22–33, 2019.

[158] S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2) :76–99, 2017.

[159] Z. C. Lipton. The mythos of model interpretability. *arXiv :1606.03490*, 2016. arXiv : 1606.03490.

[160] D. Erhan, T. Bengio, A. Courville, P. Vincent, and P. O. Box. Visualizing higher-layer features of a deep network. *Univ. of Montreal*, 1341(3), 2009.

[161] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73 :1–15, 2018.

[162] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. Analyzing classifiers : Fisher vectors and deep neural networks. In *Proc. CVPR*, pages 2912–2920, 2016.

[163] E. Thuillier, H. Gamper, and I. J. Tashev. Spatial audio feature discovery with convolutional neural networks. In *Proc. of ICASSP*, pages 6797–6801, 2018.

[164] F. Stöter, S. Chakrabarty, B. Edler, and E. A. P. Habets. CountNet : estimating the number of concurrent speakers using supervised learning. *IEEE TASLP*, 27(2) :268–282, 2019.

[165] J. Schlüter. Learning to pinpoint singing voice from weakly labeled examples. In *Proc. of ISMIR*, pages 44–50, 2016.

[166] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. on Neur. Net. and Learn. Syst.*, 28(11) :2660–2673, 2017.

[167] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks : visualising image classification models and saliency maps. *arXiv :1312.6034*, 2013. arXiv : 1312.6034.

[168] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proc. CVPR*, pages 5188–5196. IEEE, 2015.

[169] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, et al. Generative adversarial nets. In *Proc. of NIPS*, pages 2672–2680, 2014.

[170] J. M. Zurada, A. Malinowski, and I. Cloete. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Proc. of ISCAS*, pages 447–450, 1994.

[171] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, et al. Grad-CAM : visual explanations from deep networks via gradient-based localization. In *Proc. of ICCV*, pages 618–626, 2017.

[172] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. of ECCV*, pages 818–833, 2014.

[173] S. Bach, A. Binder, G. Montavon, F. Klauschen, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7) :e0130140, 2015.

[174] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proc. of ICML*, pages 3319–3328, 2017.

[175] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, et al. The (un)reliability of saliency methods. *arXiv :1711.00867*, 2017.

[176] S. Ioffe and C. Szegedy. Batch normalization : accelerating deep network training by reducing internal covariate shift. In *Proc. of ICML*, pages 448–456, 2015.

[177] D. Stanzial, N. Prodi, and G. Schiffrer. Reactive acoustic intensity for general fields and energy polarization. *JASA*, 99(4) :1868–1876, 1996.

[178] T. Dozat. Incorporating Nesterov momentum into Adam. Technical report, Univ. of Stanford, 2015.

[179] E. A. P. Habets. Room impulse response generator. Technical report, Technische Universiteit Eindhoven, 2006.

[180] L. F. Lamel, J.-L. Gauvain, and M. Eskénazi. BREF, a large vocabulary spoken corpus for French. In *Proc. of Eurospeech*, pages 505–508, 1991.

[181] E. Vincent, S. Araki, and P. Bofill. The 2008 signal separation evaluation campaign : a community-based approach to large-scale evaluation. In *Proc. of ICA*, pages 734–741, 2009.

[182] L. Arras, G. Montavon, K.-R. Müller, and W. Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proc. of WASSA*, pages 159–168, 2017.

[183] J. Huang, N. Ohnishi, and N. Sugie. Sound localization in reverberant environment based on the model of the precedence effect. *IEEE Trans. Instrum. Meas.*, 46(4) :842–846, 1997.

[184] C. Faller and J. Merimaa. Source localization in complex listening situations : selection of binaural cues based on interaural coherence. *JASA*, 116(5) :3075–3089, 2004.

[185] W. He, P. Motlicek, and J.-M. Odobez. Deep neural networks for multiple speaker detection and localization. In *Proc. of ICRA*, pages 74–79, 2018.

[186] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, et al. The ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proc. of LREC*, pages 885–888, 2004.

[187] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, et al. The Kaldi speech recognition toolkit. 2011.

[188] C. Evers, H. W. Lollmann, H. Mellmann, A. Schmidt, et al. LOCATA challenge - Evaluation tasks and measures. In *Proc. of IWAENC*, pages 565–569, 2018.

[189] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv :1803.01271 [cs]*, 2018. arXiv : 1803.01271.