Codage ambisonique pour les communications immersives

Pierre Mahé

UNIVERSITÉ DE LA ROCHELLE

ÉCOLE DOCTORALE N°618 EUCLIDE

LABORATOIRE : L3i - Laboratoire informatique Image et Interaction

THÈSE présentée par :

Pierre MAHÉ

soutenue le : **03 février 2022**
pour obtenir le grade de : **Docteur de l'Université de La Rochelle**
Discipline : **Informatique**

Codage Ambisonique pour les Communications Immersives

JURY :

Roland BADEAU	Professeur, Télécom Paris	Rapporteur
Mathieu LAGRANGE	Chargé de Recherche, CNRS, Centrale Nantes	Rapporteur
Laurent GIRIN	Professeur, Grenoble-INP	Examinateur
Gaël RICHARD	Professeur, Télécom Paris	Examinateur
Sylvain MARCHAND	Professeur, Université de La Rochelle	Directeur de thèse
Stéphane RAGOT	Ingénieur de Recherche, Orange Labs Lannion	Encadrant de thèse

Remerciements

En premier lieu, je tiens à remercier mes deux encadrants de thèse Stéphane Ragot et Sylvain Marchand pour avoir pris le temps de suivre mes travaux en apportant, entre autres, leurs expertises dans le domaine de l'audio. Je voudrais également les remercier pour m'avoir fait confiance et épaulé pendant ces trois années et de m'avoir ouvert les portes de la recherche. Stéphane, merci de m'avoir offert l'opportunité d'effectuer cette thèse au sein des laboratoires d'Orange Labs et d'avoir partagé avec moi toutes tes connaissances sur les rouages et les fonctionnements de la compression audio. Sylvain, merci d'avoir partagé ton expérience et ta vision sur le domaine. Le regard que tu as su apporter sur mes recherches m'a permis d'améliorer mes capacités de mise en perspective et d'abstraction qui seront m'être utile dans ma vie future de chercheur. Je ne peux que regretter que les contraintes sanitaires n'aient pas permis de nous voir plus souvent.

J'aimerais exprimer ma gratitude aux membres du jury à Roland Badeau et Mathieu Lagrange pour avoir accepté d'être rapporteurs de ce manuscrit de thèse et à Laurent Girin et Gaël Richard pour leur présence en qualité d'examinateurs.

Je souhaiterais remercier tous les membres de l'équipe CVA, et tout particulièrement, Jérôme Daniel pour les discussions captivantes sur la spatialisation sonore ainsi que pour le partage de belles découvertes musicales, mais également, Marc Émerit et Grégory Pallone pour les discussions sur le rendu binaural, l'appui technique et le temps consacré pour m'aider à la mise en place des différents tests subjectifs. Je souhaiterais remercier chaleureusement Théo Ladune pour m'avoir transmis son enthousiasme et son engouement pour les méthodes de compression par réseaux de neurones ainsi que pour l'aide et les conseils éclairés qu'il m'a partagé.

Je souhaiterais également remercier : Léa Krief, Rémi Rigale, David Espinel, Tangui Le Gléau, Sylvain Barthomeuf, Bruno Thiao-Layel, ainsi que toute la communauté des doctorants d'Orange Labs Lannion pour le soutien de tous les jours durant les moments de doute et de victoire, les nombreuses discussions enrichissantes et avec qui j'ai partagé l'aventure de la thèse.

Pour finir, je souhaiterais remercier toutes les personnes que j'ai croisées pendant ces trois années de thèse et qui ont contribué, même indirectement, à son bon déroulement. Ces trois années m'ont beaucoup appris et m'ont permis de mettre un pied dans le monde de la recherche. J'ai beaucoup aimé travailler à Orange Labs Lannion, notamment pour les personnes captivantes, et singulières que j'ai pu rencontrer. Je suis aussi heureux d'avoir pu découvrir les richesses de cette région et j'espère avoir su profiter des opportunités qu'elle proposait.

Table des matières

Table des figures

Liste des tableaux

Liste d'abréviations

Introduction

Contexte et enjeux

Actuellement les codecs téléphoniques sont essentiellement mono. Avec la démocratisation de la réalité virtuelle et des contenus immersifs, de plus en plus de systèmes d'enregistrement et de restitution pour le contenu audio spatialisé ont vu le jour. Les nouvelles générations de téléphones étant équipées de plusieurs microphones, il serait donc possible de traiter l'audio 3D et de proposer des communications plus immersives. Pourtant, mis à part dans le domaine du cinéma, avec des contenu comme le 5.1, 22.2, *Dolby Atmos*... l'audio immersif reste encore peu utilisé.

Cette faible utilisation peut être en partie expliquée par la quantité de données nécessaires pour transmettre et stocker ce type de signaux audio. Ce frein est d'autant plus grand dans le domaine de la téléphonie où les débits sont relativement restreints. De plus, les codecs téléphoniques actuel ne sont pas adaptés pour compresser des contenus multicanaux. Ce manque de codecs efficaces pour coder l'audio immersif a fait naitre le besoin d'élaborer de nouvelles méthodes de compression dédiées aux signaux immersifs répondant aux contraintes du domaine téléphonique.

Le sujet de cette thèse porte sur la problématique de la compression de signaux audio immersifs pour les communications immersives (la téléphonie, réalité virtuelle...). Pour représenter un contenu audio spatial, il existe plusieurs types de représentations. L'une d'elles, l'Ambisonie permet une représentation d'une scène audio captée en un point. Pour cela, le champ sonore est décomposé selon une base d'harmoniques sphériques, appelées composantes ambisoniques. En captant directement le champ sonore, le format ambisonique permet d'avoir une représentation indépendante du contenu de la scène ainsi que du matériel de captation et de restitution sonore. La captation de la scène sonore peut être faite à l'aide d'une antenne microphonique, aussi appelée microphone ambisonique, de manière similaire à une prise de son dite *live*. Cette flexibilité fait de l'Ambisonie une représentation prometteuse pour les nouveaux codecs audio. Par ailleurs, ce format est de plus en plus plébiscité par un certain nombre d'acteurs de la communication et du divertissement en ligne comme Orange, YouTube, Facebook. De plus, les équipes d'Orange ont fortement contribué à son développement et son utilisation. Le format ambisonique a également été intégré dans la norme *MPEG-H 3D* audio [MPEG, 2013].

Dans le domaine de la téléphonie, le 3GPP, l'un des organismes principaux en charge de standardisation pour les nouveaux codecs a lancé la normalisation d'un nouveau codec téléphonique immersif appelé *IVAS* [3GPP TR 26.997, 2019]. Cette normalisation a débuté en 2017 et devrait se terminer à l'horizon 2023. Tout au long de cette thèse, nous nous sommes appuyés sur les contraintes et recommandations liées à la normalisation du codec *IVAS* pour fixer le cadre de nos investigations.

Dans le domaine du codage audio, l'élaboration d'un nouveau codec se fait de manière incrémentale, chaque codec se basant sur les codecs précédents. Pour la compression de signaux audio multicanaux, cela ce traduit par l'utilisation de codecs mono ou stéréo existants comme briques de base du nouveau système. Les nouvelles méthodes consistent à ajouter des traitements en amont et aval du codec cœur pour permettre de traiter des signaux multicanaux. Cette approche permet

de bénéficier des avancées faites dans le domaine du codage mono ou stéréo.

Dans ces travaux de thèse, nous n'avons pas voulu proposer des améliorations liées à un codec particulier, mais proposer des méthodes qui pourraient être utilisables pour un ensemble de codecs cœur. Nous nous sommes intéressés principalement à deux codecs conversationnels : *EVS* et *Opus*. Pour l'évaluation de la qualité subjective de nos méthodes de compression, nous nous sommes concentrés sur la compression de l'Ambisonie de premier ordre. Cependant, les méthodes développées dans nos recherches pourraient être étendues à des signaux ambisoniques d'ordres supérieurs.

Contributions et plan de la thèse

Le travail de cette thèse a tout d'abord consisté à analyser l'état de l'art dans le domaine de la perception sonore, de la captation et de la restitution d'une scène sonore, présenté dans les chapitres 1 et 2. Dans le domaine de la compression audio, l'évaluation subjective de la qualité audio est un point essentiel pour comparer les différentes méthodes de codage. Une revue des méthodes existantes est présentée au chapitre 2. Une analyse des différentes méthodes de codage audio ambisonique, a été réalisée. Le fonctionnement des différentes méthodes de codage est décrit dans le chapitre 3.

La première partie nos travaux de recherche, décrits au chapitre 4, s'est tournée vers l'amélioration de l'approche multimono. Dans cette approche chaque canal ambisonique est codé de manière indépendante par un codec cœur mono. Une évaluation de l'approche a été menée pour en déterminer les capacités et les limites. Puis, nous nous sommes intéressés à l'élaboration d'une extension de l'approche multimono pour essayer d'améliorer la qualité spatiale. Notre système se base sur la mise en œuvre d'un rematriçage des composantes. Ce rematriçage a pour but de décorréler les composantes avant le codage multimono. Pour déterminer ce rematriçage, une analyse en composante principale, ou Principal Component Analysis (PCA), est réalisée sur chaque trame du signal ambisonique. Pour minimiser la latence ajoutée par l'extension, aucun recouvrement temporel n'est fait entre deux trames consécutives. Pour garantir la continuité du signal entre les trames, une interpolation des matrices PCA est réalisée dans le domaine des quaternions.

Ces premières réalisations nous ont conduit à nous intéresser à la visualisation du champ sonore ainsi que sur l'extraction de l'information spatiale à partir d'un signal ambisonique. Dans une seconde partie de nos travaux, nous avons cherché à créer des méthodes de codage exploitant cette information spatiale. Une méthode de post-traitement, décrite dans le chapitre 5, a été élaborée. Ce post-traitement vise à corriger les déformations spatiales du signal après un codage multimono. Le principe de la méthode est de calculer un ensemble de cartographies de l'énergie du signal d'origine et du signal décodé. À partir de ces cartographies, une matrice de correction est déterminée grâce à l'utilisation de la factorisation de Cholesky. En appliquant cette matrice sur le signal ambisonique décodé, les déformations spatiales de ce dernier sont grandement atténuées.

Par la suite, des travaux ont été réalisés pour adapter le principe utilisé dans la méthode de post-traitement en créant une méthode de codage paramétrique. Pour cette méthode, décrite au chapitre 6, seule la composante omnidirectionnelle du signal ambisonique est transmise. Au décodage, à l'aide de l'information spatiale, transmise en tant que méta-donnée, la spatialisation du signal

d'origine est reproduite. Un des points clé de cette méthode a été la décorrélation du signal pour recréer la largeur des sources ainsi que l'enveloppement de la scène. Une étude comparative a été réalisée entre notre méthode et la méthode paramétrique DirAC [Pulkki *et al.*, 2018].

Dans une dernière partie, plus exploratoire, nous avons étudié la possibilité de remplacer le codec cœur mono par une méthode de compression par réseau de neurones. Pour nos recherches, présentées au chapitre 7, nous nous sommes basés sur le modèle de compression d'image par auto-encodeur variationnel, ou Variational Autoencoder (VAE) présenté dans l'article [Ballé *et al.*, 2017]. Le modèle a été adapté pour lui permettre de traiter du signal audio mono. Dans un second temps, une extension de ce modèle a été proposée pour permettre d'améliorer la qualité audio. Cette extension est basée sur le rafinement proposé dans l'article [Minnen *et al.*, 2018]. Une analyse des performances des modèles développés a été conduite selon plusieurs métriques objectives, ces résultats ont été comparés avec des performances de deux codecs traditionnels *MP3* et *Opus*.

CHAPITRE 1

Représentation d'une scène sonore

Sommaire du chapitre

Il existe plusieurs manières de représenter une scène sonore spatialisée. Les représentations peuvent être regroupées sous 3 grandes catégories : la représentation basée canaux, la représentation basée objets et la représentation basée scène. Selon les besoins et applications, l'une ou l'autre de ces représentations est utilisée.

Dans les représentations basées canaux, comme la stéréo ou le 5.1, le système de restitution est fixé lors de la création du contenu. Cela permet d'avoir un contrôle très précis sur le rendu de la scène sonore restituée, cependant toute variation entre le système défini et le système réellement utilisé par l'utilisateur, en terme de nombre de haut-parleurs ou de leurs positions impactera grandement la scène sonore.

Dans les représentations basées objets, l'espace sonore est constitué d'un ensemble de sources sonores. Chaque source sonore utilisée dans le mixage doit être enregistrée séparément. La localisation et la trajectoire de chacune d'elles dans l'espace doivent être définies individuellement. Lors de la restitution, un moteur de rendu permet de spatialiser les sources en fonction du système de restitution. Cette représentation est adaptée pour le contenu synthétique ou issu d'un mixage, pour les enregistrements studio ou les jeux vidéo, mais elle demande une phase de post-production pour définir les positions de chaque source.

Dans les représentations basées scène, dont l'Ambisonie est le formalisme le plus populaire, l'ensemble de la scène sonore est capturée en un point par une antenne microphonique. Une conversion de la scène, appelée encodage spatial, est faite vers une représentation interne. Lors de la restitution, cette représentation interne est convertie vers le système de restitution, cette étape est appelée décodage spatial. Cette représentation interne est indépendante du système de captation et de restitution utilisé. Les représentations basées scène sont adaptées pour des captations

dites *live*, où la fidélité de la scène sonore est importante et où le contenu de la scène n'a pas pour vocation à être modifié. Dans le domaine de la téléphonie et des communications immersives, les systèmes de captation et restitution peuvent être différents pour chaque utilisateur. De plus, lors d'appel, la scène sonore doit pouvoir être partagée, sans retouche ou opération de mixage de la part des utilisateurs. Ces contraintes font de la représentation ambisonique, la représentation intéressante pour les codecs de communications immersives.

1.1 La représentation ambisonique

Initialement formalisée dans [Gerzon, 1985] pour décrire une scène sonore à l'ordre 1 ($N = 1$) dans les années 80, la représentation est appelée Ambisonie d'ordre 1, ou First-Order Ambisonics (FOA). Le formalisme de l'Ambisonie a été étendu [Daniel, 2000] aux ordres supérieurs, ou Higher-Order Ambisonics (HOA). Le but est de reproduire le champ sonore en un point théorique, correspondant à la position de l'auditeur. Pour cela, l'Ambisonie décompose le champ sonore en un point.

Les coordonnées sphériques s'avèrent être la représentation la plus adaptée pour manipuler l'Ambisonie. Tout point de l'espace peut être localisé par 3 coordonnées : la distance à l'origine r, l'azimut θ et l'élévation ϕ. Plusieurs conventions existent, la figure 1.1 indique la convention utilisée pour décrire l'espace dans ce manuscrit. Dans la convention utilisée, l'élévation ϕ est définie à partir du plan horizontal et les angles sont mesurés selon le sens trigonométrique. La représentation ambisonique fait l'hypothèse que toutes les ondes incidentes sont des ondes planes, les sources sont situées à l'infini, la coordonnée de distance r est donc généralement omise, seules les coordonnées de directions (θ, ϕ) sont utilisées pour repérer une source dans l'espace. Cependant, la coordonnée de distance peut s'avérer nécessaire dans le cas de la simulation de sources en champ proche [Daniel, 2003].

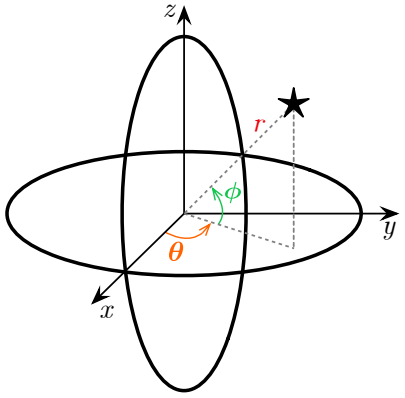


FIGURE 1.1 – Système de coordonnées sphériques utilisé.

Dans la représentation ambisonique, le champ sonore est décomposé selon un ensemble de fonctions propres. Ces fonctions propres combinent un ensemble de fonctions pour capturer à la fois la dépendance radiale et angulaire de l'onde acoustique. La dépendance radiale est décrite par une série de Fourier-Bessel du premier ordre $j_m(kr)$, où k est le nombre d'onde et r le diamètre

de la sphère. La dépendance angulaire est, elle, décrite par les harmoniques sphériques $Y_{nm}^{\sigma}(\theta, \phi)$. De manière générale, le champ de pression acoustique $p(r,\omega)$ peut être exprimé comme :

$$\underbrace{p(r,\ \omega,\ \theta,\ \phi)}_{\substack{\textbf{pression}\\ \textbf{acoustique}}} = \sum_{n=0}^{+\infty} i^n \underbrace{j_n(kr)}_{\substack{\textbf{dépendance}\\ \textbf{radiale}}} \sum_{m=-n}^{n} \underbrace{B_{nm}^{\sigma}(\omega)}_{\substack{\textbf{coefficients}\\ \textbf{HOA}}} \underbrace{Y_{n,m}^{\sigma}(\theta,\phi)}_{\substack{\textbf{dépendance}\\ \textbf{angulaire}}} \tag{1.1}$$

avec ω correspond à la vitesse angulaire et k est le nombre d'onde. Ces deux grandeurs peuvent être reliées par la formule $k = \frac{\omega}{c}$ où c correspond à la célérité du son. Grâce à l'équation 1.1, le champ de pression acoustique peut être entièrement décrit par les coefficients $B_{nm}^{\sigma}(\omega)$ associés à chaque harmonique sphérique indicé par $0 \leq m \leq n$, $n \geq 0$ et $\sigma = \pm 1$. Ces coefficients sont appelés composantes ambisoniques et constituent la représentation ambisonique du champ sonore. L'ensemble de ses composantes est également appelé format-B. Il est théoriquement possible de représenter parfaitement le champ sonore avec nombre de composantes $n \to +\infty$. Cependant en pratique, la représentation ne peut pas être représentée par un nombre de termes infini. Une troncature de la représentation est donc faite à un ordre donné N. Pour une représentation d'une scène 3D à l'ordre N, le nombre de composantes B_{nm} sera de $(N + 1^2)$. Pour une représentation d'un scène planaire, le nombre de composantes est de $2N + 1$.

La figure 1.2 représente les fonctions harmoniques sphériques pour les 4 premiers ordres. La première ligne correspondant à l'harmonique d'ordre 0. La seconde ligne correspond aux harmoniques à ajouter à l'ordre 0 pour obtenir l'ordre 1. L'ajout des harmoniques d'une nouvelle ligne aux harmoniques de l'ordre précédent permet d'obtenir l'ordre suivant. Pour la représentation de l'Ambisonie planaire, seuls les harmoniques selon le plan horizontal, $m = \pm n$, sont à considérer. Ces harmoniques correspondent aux harmoniques situés sur l'extérieur du triangle de la figure 1.2.

Dès l'ordre 1, la sensation d'enveloppement de la scène est présente, mais la localisation de la direction des sources sonores sera imprécise. Pour améliorer la localisation des sources, l'ordre ambisonique devra être augmenté. Cependant, le nombre de composantes augmentant au carré de l'ordre ambisonique, plus l'ordre sera élevé plus le nombre de composantes nécessaires pour représenter la scène sera importante. Ce compromis entre précision et quantité de données est un élément à prendre en compte lors de l'élaboration des méthodes de compression.

Dans la suite de cette thèse, le terme général "Ambisonie" sera utilisé pour désigner toute représentation d'une scène 3D, quelque soit l'ordre utilisé. Les termes FOA et HOA seront utilisés uniquement quand le besoin de distinction entre les différents ordres sera nécessaire. Le terme "planaire" pourra être précisé pour certains cas spécifiques.

Il existe plusieurs conventions dans la manière de numéroter les composantes : Single Index Designation (SID) [Daniel, 2000], Ambisonic Channel Number (ACN) [Chapman *et al.*, 2009], Furse-Malham (FuMa) [Malham, 1999]. La convention acACN tend à s'imposer comme standard, c'est cette convention qui a été utilisée pour les travaux de cette thèse. Il est à noter que le choix du système de numérotation des composantes n'a pas d'impact sur la faisabilité et les performances des méthodes développées par la suite dans ce manuscrit.

Il existe également plusieurs conventions de normalisation : maxN, N3D, SN3D ... Dans nos travaux, nous avons utilisé la normalisation SN3D. Il est possible de convertir un signal ambiso-

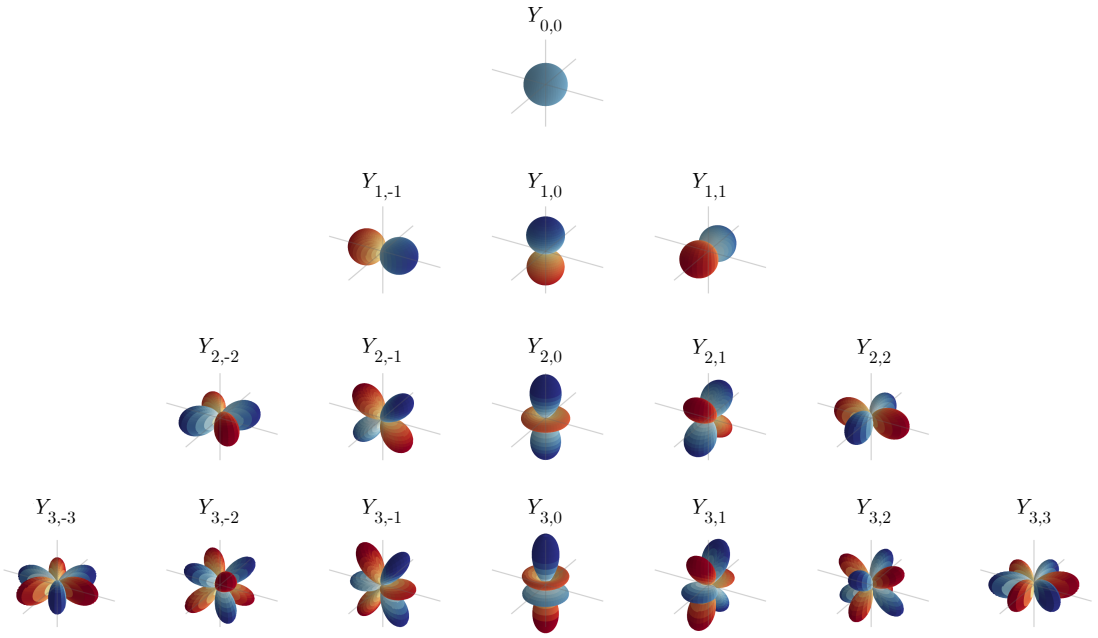


FIGURE 1.2 – Représentation des premiers harmoniques sphériques. La partie en bleu correspond aux valeurs positives, la partie en rouge aux valeurs négatives des harmoniques.

nique d'une normalisation vers une autre en multipliant les composantes par un facteur de normalisation. La conversion des composantes d'une normalisation vers le SN3D avant l'entrée des méthodes serait suffisante pour pouvoir utiliser les méthodes développées durant ces travaux de thèse avec n'importe quel signal, quel que soit la convention de normalisation de ce dernier.

1.1.1 Spatialisation d'une source à l'ordre 1

Il est possible de projeter artificiellement une source mono dans le domaine ambisonique. En FOA ($N = 1$), le nombre de composantes est égal à $n = 4$. Ces composantes sont nommées par convention w, x, y, z. Pour une source $s(t)$ à la position (θ, ϕ), la source sera encodée selon la formule :

$$\mathbf{b}(t) = \mathbf{e} \cdot s(t) = \begin{bmatrix} 1 \\ \cos\,\theta\,\cos\,\phi \\ \sin\,\theta\,\cos\,\phi \\ \sin\,\phi \end{bmatrix} s(t) = \begin{bmatrix} w(t) \\ x(t) \\ y(t) \\ z(t) \end{bmatrix} \tag{1.2}$$

où $\mathbf{e}$ est le vecteur d'encodage pour la source dans la direction donnée. Ces 4 composantes FOA peuvent être interprétées comme la captation du champ acoustique en un point de l'espace par quatre microphones coïncidents : un omnidirectionnel w, et trois microphones bidirectionnels x, y, z, chacun selon un axe de l'espace différent.

Pour spatialiser plusieurs sources en même temps, la même opération d'encodage devra être répétée pour chaque source. Puis l'ensemble des composantes w (resp. x, y et z) du format-B devra

être ajouté. Avec cette méthode, la source sonore est encodée comme une onde plane, sa position peut être considérée comme à l'infini. Pour reproduire une source avec un effet de champs proche, où l'onde ne peut plus être assimilée à une onde plane, nous renvoyons le lecteur aux travaux présentés dans [Daniel, 2003].

1.1.2 Encodage ambisonique d'une prise de son réelle

Une scène sonore ne peut pas être capturée directement dans le domaine ambisonique. L'utilisation d'une antenne microphonique est nécessaire pour capturer le champ sonore dans une représentation intermédiaire, le format-A. Cette représentation intermédiaire correspond au format brut venant des capsules de l'antenne, c'est pour cela qu'il peut être parfois appelé format capsule. Cette représentation doit ensuite être convertie vers la représentation ambisonique, ou le format-B, à l'aide d'un rematriçage des composantes. Des filtres d'égalisation peuvent être appliqués aux composantes pour compenser les éventuelles déformations générées par la forme de l'antenne microphonique. Par abus de langage, les antennes microphoniques sont simplement désignées par la dénomination : microphone ambisonique. Le calcul d'une représentation ambisonique à l'ordre N demande, au minimum, autant de capsules que de coefficients ambisoniques $N + 1$. La figure 1.3 montre un exemple d'une antenne microphonique : l'*Eigenmike* [Mh-Acoustics, 2013]. Une antenne microphonique est constituée de 32 capsules réparties de manière quasi-uniforme sur l'ensemble d'une sphère de $8{,}4$ cm de diamètre. Le nombre de capsules permet un encodage HOA allant jusqu'à l'ordre 4.

FIGURE 1.3 – Antenne de microphones Eigenmike de Mh Acoustics [Mh-Acoustics, 2013].

Comme pour tout système de mesure, l'utilisation d'un nombre fini de capsules produira un effet d'échantillonnage spatial. Quand le nombre de points de mesures deviendra trop faible par rapport au phénomène observé, un repliement spatial va apparaître. Lors de la conception d'une antenne microphonique, deux contraintes sont à prendre en compte :

Dans les hautes fréquences, un repliement spatial a lieu quand la longueur d'onde λ est inférieure à la distance inter-capsule d. Ce qui se traduit par un besoin de maintenir la distance inter-capsule d la plus faible possible : soit par une sphère de diamètre le plus petit possible, soit par un très grand nombre de capsules, pour avoir une représentation spatiale la plus cohérente possible en hautes fréquences.

À l'inverse, quand les longueurs d'onde $\lambda \gg d$, la représentation des basses fréquences seront approximatives. Pour bien représenter les basses fréquences, les capsules ont besoin de mesurer une différence de pression entre elles. Si la distance entre les capsules n'est plus significative, l'erreur d'estimation du gradient de pression sera plus importante que le gradient de pression lui-même. Pour l'encodage correct des basses fréquences, la sphère a besoin d'avoir un diamètre important. La sensibilité et qualité intrinsèque de chaque capsule utilisée pour l'antenne peut également apporter des approximations dans l'encodage spatial. Les questions du repliement spatial et de la conception d'antenne microphonique ont été étudiées de façon détaillée dans la thèse de Moreau [Moreau, 2006].

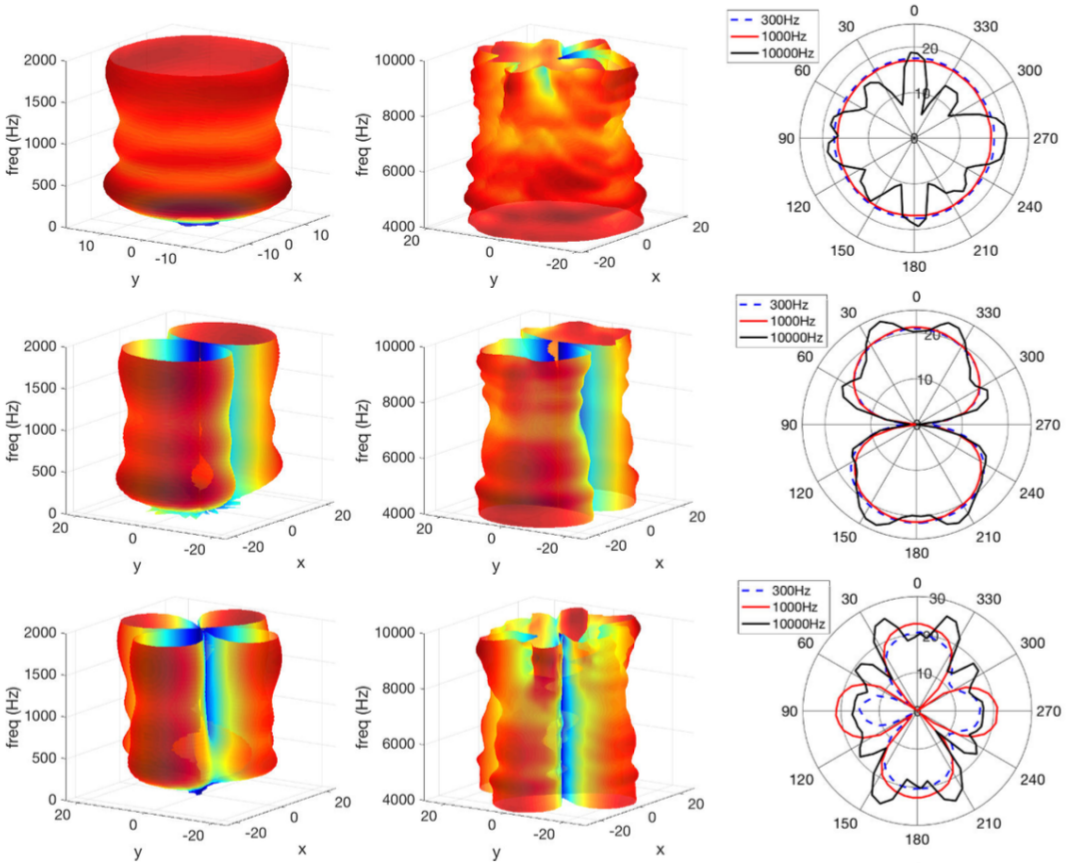


FIGURE 1.4 – Directivité mesurée en fonction de la fréquence des composantes ambisoniques pour le microphone Eigenmike dans le plan horizontal. De haut en bas : composante d'ordre 0 à 2, de gauche à droite : représentation pour les basses et les hautes fréquences ainsi que le diagramme de directivité. Cette figure est issue de [Baque, 2017].

Dans sa thèse [Baque, 2017], Baqué présente les résultats détaillés sur les mesures de directivités réelles de l'Eigenmike. La figure 1.4 montre la directivité mesurée pour cette antenne. Pour améliorer la lisibilité des figures polaires, la valeur unitaire utilisée, pour chaque graphique, est différente. Seules les courbes d'un même graphique peuvent être comparées. En fonction de la fréquence, il est possible de voir que la directivité obtenue comporte des déformations par rapport à la directivité théorique. Pour la composante omnidirectionnelle W, la limite de l'encodage en

basses fréquences (à gauche) se situe autour de 200 Hz. Pour les hautes fréquences, un repliement spatial est visible dès 8 kHz. La plage de validité spatiale pour l'ordre 0 pour l'Eigenmike est donc comprise entre 200 à 8000 Hz. Cette plage de validité est fixée en observant 2 critères : la corrélation de la directivité ambisonique mesurée avec la directivité théorique (*Corr.*) et l'énergie moyenne de l'ordre donné par rapport à une composante omnidirectionnelle théorique (*Gain*). Le seuil pour chaque critère est défini arbitrairement, dans [Baque, 2017], ils ont été fixés comme *Corr.* $> 0{,}95$ et *Gain* > -3 dB. Pour les composantes d'ordre supérieur, le même type d'étude de validité peut être fait. Plus l'ordre est important plus cette plage de validité sera étroite. Pour les composantes d'ordre 4, la plage de validité a une largeur de moins de 3 kHz, allant de $3{,}4$ kHz à $6{,}2$ kHz.

1.1.3 Restitution sonore

L'Ambisonie permet une représentation de la scène sonore indépendante du système de diffusion utilisé pour la restitution. Le contenu enregistré n'est pas lié à une configuration de haut-parleurs, contrairement au contenu basé canal, comme le *surround* (5.1, 7.1, 22.2...). Avant de pouvoir écouter le contenu ambisonique, il a besoin d'être converti du domaine ambisonique au domaine des haut-parleurs.

Ce décodage sur haut-parleurs d'un contenu ambisonique peut être fait de plusieurs manières (*basique*, max-r_E, *in-phase*...) [Daniel, 2000]. Le décodage présenté ici sera le décodage spatial dit *basique*. Ce décodage est uniquement basé sur une transformation linéaire sans procédure d'optimisation perceptive ou de contrainte de parcimonies. L'onde sonore émise par les haut-parleurs est considérée comme plane dans la zone d'écoute.

Le principe du décodage basique est de reproduire le champ sonore à partir d'un ensemble de haut-parleurs dans la zone de l'espace où se situe l'auditeur. Les haut-parleurs peuvent être situés sur une sphère dans le cas de l'ambisonique 3D ou un cercle pour l'ambisonique planaire. Le décodage basique utilise le principe dit du "réencodage" [Gerzon, 1985], ce principe implique la relation :

$$\mathbf{ES} = \mathbf{B} \tag{1.3}$$

$$\text{avec } \mathbf{E} = \begin{bmatrix} \mathbf{Y}_1(\theta_1, \phi_1) & \dots & \mathbf{Y}_1(\theta_L, \phi_L) \\ \vdots & \ddots & \vdots \\ \mathbf{Y}_N(\theta_1, \phi_1) & \dots & \mathbf{Y}_N(\theta_L, \phi_L) \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_L \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_N \end{bmatrix} \tag{1.4}$$

où L est le nombre de haut-parleurs utilisés, la matrice $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_L]$ contient le vecteur des harmoniques sphériques associés à la direction de chaque haut-parleur. Le vecteur $\mathbf{B}$ contient les composantes ambisoniques et $\mathbf{S}$ les signaux joués par les haut-parleurs. Le décodage basique peut se formuler comme un système de $(N+1)^2$ équations linéaires à L inconnues qu'il faut résoudre. Pour trouver une solution exacte au système, le nombre L de haut-parleurs doit être au moins de $(N+1)^2$ en 3D et $2N+1$ en 2D. En dessous de ce nombre, seule une solution approximative pourra être trouvée produisant un champ sonore détérioré par rapport au champ sonore représenté par les composantes $\mathbf{B}$.

Une matrice de décodage $\mathbf{D}$ est déterminée pour permettre de convertir les composantes ambisoniques vers l'espace des haut-parleurs. Chaque terme D_{ij} correspond au gain à appliquer à la $i^{\text{ième}}$ composante ambisonique pour le haut-parleur j, tel que :

$$\mathbf{S} = \mathbf{DB} \tag{1.5}$$

La matrice d'encodage $\mathbf{E}$ dépend de la position des haut-parleurs. Comme le nombre L de haut-parleurs peut être supérieur au nombre de composantes de $\mathbf{B}$, la matrice d'encodage n'est pas obligatoirement inversible. Une solution générale peut être trouvée à l'équation 1.5 en calculant le pseudo-inverse de Moore-Penrose [Golub et Van Loan, 1983] :

$$\mathbf{D} = \mathbf{E}(\mathbf{E}\mathbf{E}^t)^{-1} \tag{1.6}$$

Cas idéal

Dans le cas, où le nombre de haut-parleurs L est égal au nombre de composantes de $\mathbf{B}$ et si ces haut-parleurs sont répartis de manière régulière sur la surface d'une sphère pour le cas 3D, ou le contour d'un cercle pour le cas 2D, le système peut être résolu de manière exacte. Cela revient à trouver un ensemble de directions (θ_l, ϕ_l) avec $l = \{1, \ldots, L\}$ préservant la propriété d'orthonormalité des harmoniques sphériques, ce qui peut s'exprimer mathématiquement comme :

$$\mathbf{E} \cdot \mathbf{E}^t = \frac{1}{L} \cdot \mathbf{I}_L \tag{1.7}$$

où $\mathbf{I}_L$ est la matrice identité de dimension $L \times L$. Dans ce cas idéal, la matrice de décodage s'exprime alors simplement comme :

$$\mathbf{D} = \frac{1}{L}\mathbf{E}^t \tag{1.8}$$

Le problème est de trouver un ensemble de directions (θ_L, ϕ_L) de L haut-parleurs répartis de manière régulière sur une sphère. Ce problème revient à devoir trouver une discrétisation uniforme d'une sphère. Pour un nombre limité de haut-parleurs, une telle discrétisation est possible, notamment grâce aux solides de Platon. Malheureusement quand le nombre de haut-parleurs devient trop important, il n'existe pas de solution exacte. Seule une discrétisation quasi-uniforme est possible, ce qui oblige à utiliser la solution générale donnée par l'équation (1.5). Un ensemble de méthodes de discrétisation pseudo-uniforme sont présentées dans [Rafaely, 2019]. Dans le cadre de cette thèse, nous avons utilisé une discrétisation de Lebedev [Beentjes, 2015].

En plus du décodage basique, il existe d'autres décodages qui proposent d'améliorer l'effet subjectif de localisation. Ces méthodes prennent en compte des critères comme le vecteur énergie ou le vecteur vélocité pour optimiser la localisation des sources [Moreau, 2006, Zotter et Frank, 2012]. Dans les moteurs de rendu actuels, les différents types de décodage sont souvent utilisés de manière conjointe. Le signal ambisonique est divisé en plusieurs bandes de fréquences. Un décodage différent est utilisé selon les bandes. En règle générale, 2 bandes de fréquences sont utilisées, un décodage basique est appliqué pour la bande basse alors que pour la bande haute un décodage max-r_E est appliqué. À titre d'exemple, dans l'analyse de différents

décodeurs [Heller *et al.*, 2008], tous les décodeurs découpent le signal en 2 bandes avec une fréquence de coupure autour de 400 Hz.

1.1.4 Binauralisation

Pour restituer une scène sonore ambisonique sur casque, une étape supplémentaire doit être ajoutée au processus de décodage ambisonique. Cette étape est appelée la synthèse binaurale ou binauralisation. Le traitement de binauralisation consiste à appliquer une fonction de transfert entre chacun des haut-parleurs et chaque tympan de l'auditeur. Cette fonction de transfert, aussi appelée filtre binaural, dépend de deux éléments : la position (θ, ϕ) du haut-parleur et la morphologie de l'auditeur [Blauert, 1997].

Ces filtres binauraux, ou Head-Related Transfer Function (HRTF) doivent être mesurés pour chaque auditeur. La mesure des HRTF consiste à placer un micro dans chaque conduit auditif de l'individu puis à l'aide de haut-parleurs, de mesurer la réponse impulsionnelle pour un ensemble de positions de l'espace. Cet ensemble de positions correspond à un maillage discrétisant une sphère. Plus le nombre de mesures est important, plus le nombre de positions possibles pour les haut-parleurs sera important. La distance entre deux points est de l'ordre de $10°$ (en azimut et en élévation) pour les systèmes actuels de mesures [Armstrong *et al.*, 2018]. Les HRTF sont exprimées dans le domaine fréquentiel, elles peuvent également être exprimées dans le domaine temporel, où elles sont appelées Head-Related Impulse Response (HRIR). À défaut de connaître les HRTF de l'auditeur, il est possible d'utiliser des HRTF de référence mesurée sur un mannequin, comme le *Neumann KU-100*, censé représenter la morphologie moyenne.

Le traitement de binauralisation est réalisé en filtrant l'ensemble des L signaux $\mathbf{S}$ issus des équations de décodage (1.5) par les HRTF correspondantes aux directions (θ_l, ϕ_l) de chaque haut-parleur. Les signaux binauraux résultants sont ensuite reproduits sur le casque d'écoute. Puis de sommer les contributions associées à chacune des oreilles. Les signaux binauraux $\mathbf{s}_G$ et $\mathbf{s}_D$, respectivement gauche et droit, peuvent être exprimés, tels que :

$$\begin{cases} \mathbf{s}_G = \sum_{l=1}^{L} h_G(\theta_l, \phi_l) \cdot S_l = \mathbf{h}_G^t \cdot \mathbf{S} \\ \\ \mathbf{s}_D = \sum_{l=1}^{L} h_D(\theta_l, \phi_l) \cdot S_l = \mathbf{h}_D^t \cdot \mathbf{S} \end{cases} \tag{1.9}$$

où $h_G(\theta_l, \phi_l)$ et $h_D(\theta_l, \phi_l)$ correspond à la paire de filtres HRTF, associés respectivement à l'oreille gauche et l'oreille droite, pour le $l^{\text{ième}}$ haut-parleur.

En recombinant les équations de décodage sur haut-parleur et le filtrage, il est possible de faire directement une projection du domaine ambisonique vers le domaine binaural sans calculer les signaux intermédiaires [Moreau, 2006].

Pour limiter encore le coût de calcul lié à la binauralisation, d'autres optimisations sont possibles, dans l'article [Daniel, 2000] l'auteur propose une simplification des HRTF en se basant sur l'hypothèse de la symétrie des filtres HRTF selon le plan sagittal, ce qui permet de diminuer le nombre total de convolutions à réaliser. Comme pour le décodage sur haut-parleurs, des méthodes proposent d'ajouter des contraintes lors du filtrage binaural pour améliorer la localisation des sources sonores. Dans [Mccormack et Delikaris-Manias, 2019], les auteurs proposent une méthode de binauralisation paramétrique basée sur estimation de la direction d'arrivée, ou Direction

of Arrival (DOA) de la source prédominante pour chaque carreau temps-fréquence du signal ambisonique. La source prédominante est ensuite binauralisée avec le filtre HRTF le plus proche de la direction d'arrivée extraite. Un signal mono représentant le champ diffus est estimé. Ce signal est décorrélé pour minimiser la cohérence inter-canal puis le signal est ajouté au canal de l'oreille gauche et de l'oreille droite. Ces méthodes de binauralisation paramétriques semblent montrer de bonnes performances . Cependant, elles posent un certain nombre d'hypothèses sur le contenu de la scène sonore (nombre de sources, niveau du champ diffus...), ce qui les rend difficilement utilisables comme rendu binaural lors de l'évaluation de qualité audio de codecs.

1.2 Manipulation de l'ambisonique

La représentation ambisonique est une représentation du champ sonore indépendante du système de captation et du système de restitution sonore. Cette représentation très mathématique du champ sonore rend les manipulations de l'espace sonore relativement simples et élégantes.

1.2.1 Rotation

La représentation ambisonique discrétisant l'espace de manière homogène, une rotation de la scène sonore correspond à une multiplication des composantes $\mathbf{B}$ avec une matrice de rotation. Cette matrice de rotation effectue une simple recombinaison linéaire des composantes. Pour un signal ambisonique d'ordre 1 avec les composantes dans l'ordre : W, X, Y, Z (convention SID). Les composantes transformées $\mathbf{B}'$ du champ sonore d'origine $\mathbf{B}$, peuvent être exprimées comme :

$$\mathbf{A}(t) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & & & \\ 0 & & \mathbf{R} & \\ 0 & & & \end{bmatrix} \mathbf{B}(t) \tag{1.10}$$

où $\mathbf{R}$ est une matrice de rotation élémentaire autour d'un axe (x, y ou z) selon les angles de Cardan :

$$\begin{aligned} \mathbf{R}(\alpha_x) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha_x) & -\sin(\alpha_x) \\ 0 & \sin(\alpha_x) & \cos(\alpha_x) \end{bmatrix} \\ \mathbf{R}(\alpha_y) &= \begin{bmatrix} \cos(\alpha_y) & 0 & -\sin(\alpha_y) \\ 0 & 1 & 0 \\ \sin(\alpha_y) & 0 & \cos(\alpha_y) \end{bmatrix} \\ \mathbf{R}(\alpha_z) &= \begin{bmatrix} \cos(\alpha_z) & -\sin(\alpha_z) & 0 \\ \sin(\alpha_z) & \cos(\alpha_z) & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{1.11}$$

avec α l'angle de rotation appliqué à la scène sonore.

1.2.2 Projection d'un espace à l'autre

À la section 1.1.3, nous avons vu que pour écouter un contenu ambisonique, une conversion de l'espace ambisonique vers l'espace des haut-parleurs est nécessaire. Dans le cas d'une configuration de haut-parleurs permettant de trouver une solution à l'équation 1.5, ce changement de domaine est complètement équivalent. De manière analogue, il est possible de transformer une représentation ambisonique $\mathbf{B}$ vers une autre représentation de l'espace sonore $\mathbf{A}$ par une matrice de transformation $\mathbf{D}$ de taille $N \times M$:

$$\mathbf{A}(t) = \mathbf{D}\mathbf{B}(t) \tag{1.12}$$

où $\mathbf{D} = [\mathbf{d}(\theta_1, \phi_1), \ldots, \mathbf{d}(\theta_M, \phi_M)]$ et $\mathbf{d}(\theta_m, \phi_m)$ le vecteur des harmoniques sphériques associés à la direction (θ_m, ϕ_m). La matrice $\mathbf{D}$ est de taille $N \times M$, où N est la dimension de la base ambisonique de B et M est la dimension de l'espace de A. Pour garantir l'équivalence des deux domaines, la dimension de l'espace d'arrivée doit être au moins équivalente à la dimension du domaine d'origine $M \geq N$ et la matrice $\mathbf{D}$ doit être unitaire, telle que :

$$\mathbf{D}^t\mathbf{D} = \mathbf{I} \tag{1.13}$$

Cette transformation d'un espace à un autre est fréquemment utilisée comme lors d'opérations de rematriçage des composantes ambisonique [Skoglund, 2018], le passage du domaine ambisonique vers le domaine spatial équivalent, ou Equivalent Spatial Domain (ESD) [3GPP TS 26.918, 2018] ou lors du passage du format-B à un format-A théorique, le format d'une antenne microphonique idéale.

1.2.3 Formation de voie

À partir d'un réseau de microphones quelconque, il est possible de recombiner les signaux des capsules pour réaliser une formation de voie aussi appelée filtrage spatial, ou *beamforming*. Pour effectuer cette recombinaison, deux étapes sont nécessaires : le calcul d'un réalignement temporel entre les différents microphones puis un matriçage des canaux. La figure 1.5 montre un diagramme du fonctionnement du mécanisme de formation de voie pour un réseau de microphones quelconque. Le format ambisonique est une captation du champ sonore en un point de l'espace, les composantes sont donc déjà des signaux coïncidant, le module d'alignement τ_n peut donc être éliminé. La complexité dans l'élaboration de la formation de voie réside donc uniquement dans le calcul de la pondération w_n de chaque composante.

La méthode basique pour faire de la formation de voie est la méthode appelée *Delay-Sum* [Van Trees, 2002]. Dans le cas de l'ambisonique, la méthode se simplifie en une somme pondérée des composantes :

$$\mathbf{y}(t) = \mathbf{w}^t(b)(t) \tag{1.14}$$

où les valeurs du vecteur $\mathbf{w} = \{\mathbf{w}[1], \ldots, \mathbf{w}[N]\}$ sont déterminées dans le cas de l'ambisonique par son vecteur d'encodage $\mathbf{d}(\theta_i, \phi_i)$ où (θ_i, ϕ_i) correspond à la direction que pointe la formation

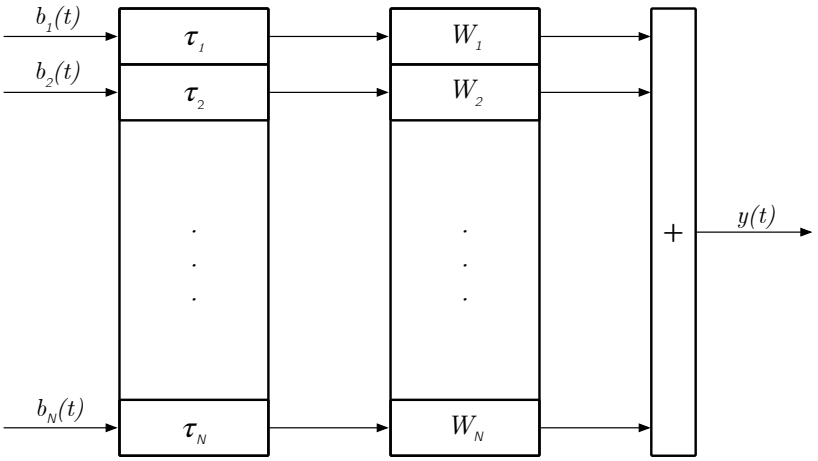


FIGURE 1.5 – Diagramme du fonctionnement de la formation de voie.

de voie. La figure 1.6 montre la directivité d'un faisceau. Des méthodes plus élaborées, avec l'ajout de contraintes, peuvent être trouvées dans [Rafaely, 2019].

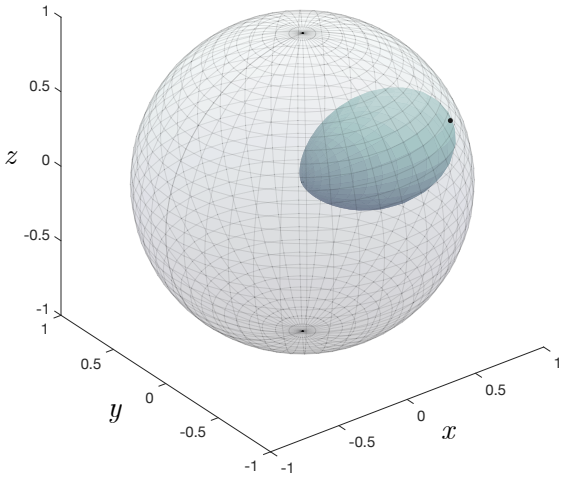


FIGURE 1.6 – Formation de voie pour la direction $(-45°, 30°)$.

Dans le domaine de l'Ambisonie, la formation de voie peut être utilisée pour différentes applications. Dans des méthodes de la séparation de source [Perotin, 2019], la formation de voie peut permettre d'extraire le signal provenant de la direction de la source. Une autre utilisation du filtrage spatial peut être dans les méthodes de rehaussement de la parole ou d'un élément du mix musical [Gorlow et Marchand, 2012]. Dans [Kronlachner, 2014b] les auteurs proposent également d'utiliser des mécanismes de formation de voie pour effectuer des modifications de l'espace sonore, comme la dilatation ou la contraction de certaines zones de l'espace.

1.2.4 Cartographie de l'espace sonore

En répétant l'opération de formation de voie pour un ensemble de points sur la sphère, il est possible de projeter ces points sur une représentation équirectangulaire pour visualiser une cartographie de la puissance sonore de la scène selon la direction (θ, ϕ), aussi appelée image spatiale. La cartographie générée par des formations de voie *Delay-Sum* est appelée réponse de puissance dirigée, ou Steered Response Power (SRP), [Jarrett *et al.*, 2010].

Il existe d'autres méthodes de formation de voie, comme Minimum Variance Distortionless Response (MVDR), Linearly Constrained Minimum Variance (LCMV) [Rafaely, 2019], ou des méthodes paramétriques comme CroPaC [Delikaris-Manias et Pulkki, 2013]. Selon la méthode de formation de voie utilisée, le résultat de la cartographie ne sera pas le même.

Avec l'approche SRP, pour une trame donnée $t = [1, \ldots, T]$, la puissance du signal provenant de la direction i peut être calculée par :

$$E_i = \|s_i\|^2 = \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{d}(\theta_i, \phi_i) \cdot \mathbf{b}[t]\|^2 \tag{1.15}$$

En répétant l'opération pour l'ensemble des points, il est possible de tracer la cartographie de la puissance de la trame. La figure 1.7 montre un exemple de cartographie calculé pour une trame de 50 ms de l'échantillon FOA avec un nombre de maillages de 648 points uniformément répartis sur la sphère. Sur la figure, les valeurs ont été normalisées par la puissance maximale du maillage. L'image spatiale étant calculée sur un certain intervalle de temps, selon la taille de trame utilisée l'image spatiale peut varier. Pour une source sonore en mouvement , un effet de flou de mouvement sera visible, la position de la source correspondra à la moyenne des positions sur l'intervalle de temps.

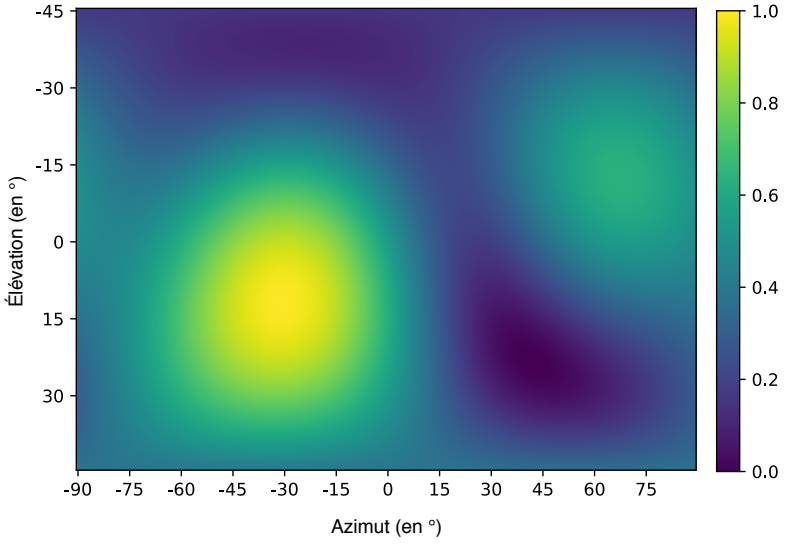


FIGURE 1.7 – Exemple de cartographie de la puissance d'une trame du signal FOA "Nature" décrit à l'annexe A.

1.2.5 Vecteur intensité acoustique

Le vecteur intensité acoustique permet de caractériser le champ sonore. Le vecteur représente le flux d'énergie en chaque point de l'espace, il est défini comme le produit de la pression sonore par le conjugué de la vitesse particulaire :

$$\mathbf{I}(t,f) = p(t,f)\mathbf{v}^*(t,f) \tag{1.16}$$

où $\mathbf{v}(t,f)$ est la vélocité particulaire et $p(t,f)$ le champ de pression généré par l'onde pour chaque échantillon t et chaque fréquence f.

Le vecteur intensité peut également être calculé dans le domaine fréquentiel. Pour chaque carreau temps-fréquence, un vecteur intensité peut être estimé. Ce vecteur est à valeur complexe, il peut être décomposé en deux parties : la partie réelle, appelée partie active, et la partie imaginaire, appelé partie réactive. Pour un signal ambisonique, le vecteur d'intensité acoustique active est donné par :

$$\mathbf{I}_a[t,f] = \Re \begin{bmatrix} W[t,f]\ X^*[t,f] \\ W[t,f]\ Y^*[t,f] \\ W[t,f]\ Z^*[t,f] \end{bmatrix} \tag{1.17}$$

avec $\Re$ l'opérateur partie réelle et $W[t,f]$, $X[t,f]$, $Y[t,f]$, $Z[t,f]$ respectivement la transformée de Fourier à court terme des composantes ambisoniques $w(t)$, $x(t)$, $y(t)$, $z(t)$. La partie réactive du vecteur intensité utilise la parte imaginaire du vecteur, définie comme :

$$\mathbf{I}_r[t,f] = \Im \begin{bmatrix} W[t,f]\ X^*[t,f] \\ W[t,f]\ Y^*[t,f] \\ W[t,f]\ Z^*[t,f] \end{bmatrix} \tag{1.18}$$

avec $\Im$ l'opérateur partie imaginaire. Pour un signal ambisonique composé d'une seule onde plane, la partie active est opposée à la direction de propagation de l'onde. Quand plusieurs ondes sont présentes, le vecteur intensité $\mathbf{I}_a[t,f]$ pourra avoir des triplets de valeurs différentes selon les carreaux temps-fréquence, en fonction de la contribution énergétique de chaque onde présente pour le carreau temps-fréquence. À partir d'un triplet de valeurs, il est possible de déterminer la DOA d'une fréquence donnée :

$$\text{DOA}[t,f] = \angle\left[-\mathbf{I}_a[t,f]\right] \tag{1.19}$$

avec $\angle$ l'opérateur qui donne l'angle (θ, ϕ) associé au vecteur. La figure 1.8 illustre pour deux trames de signal, les directions pointées par le vecteur intensité pour chacune des fréquences de la transformée de Fourier. Selon le contenu de la scène dans la trame analysée, les directions du vecteur intensité peuvent être plus ou moins reparties dans l'espace. Une forte dispersion des valeurs dans l'espace traduit un fort champ diffus. Pour une scène synthétique avec une seule source, les valeurs du vecteur intensité seront identiques pour tous les carreaux temps-fréquence. Dans le cas de signaux ambisoniques réels, le vecteur intensité peut être bruité, un lissage selon l'axe fréquentiel et temporel est souvent appliqué avant l'estimation de la DOA [Pulkki *et al.*, 2018, Chap. 5].

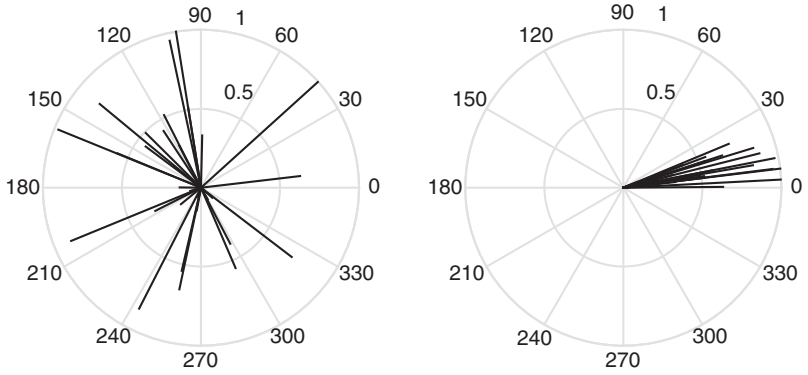


FIGURE 1.8 – Direction d'arrivée déterminée par le vecteur intensité pour une trame d'un signal ambisonique planaire exprimé en degré. À gauche, le signal est fortement diffus, à droite, le signal est non-diffus. La figure est issue de [Pulkki *et al.*, 2018].

Malgré des travaux essayant d'exploiter à la fois la partie active et réactive [Perotin *et al.*, 2019, Daniel et Kitić, 2020], de manière générale, seule la partie active est utilisée. Cela peut être expliqué par le fait que la partie active permet une interprétation directe de ces valeurs, là où l'interprétation de la partie réactive est beaucoup moins aisée. L'information contenue dans la partie réactive semble plus difficile à exploiter.

CHAPITRE 2

Perception du son et évaluation de la qualité audio

Sommaire du chapitre

2.1 Perception du son

2.1.1 Sensibilité de l'oreille

Le système auditif humain a un fonctionnement particulier, constitué de mécanismes complexes et d'un grand nombre de non-linéarités. Ce système a été très étudié pour comprendre et approximer son fonctionnement. Dans l'histoire de la compression audio, il a été essentiel de comprendre les mécanismes de base de l'audition pour pouvoir appliquer certaines approximations et simplifications aux signaux lors du codage sans dégrader la qualité perçue. La figure 2.1 représente la sensibilité de l'oreille selon la fréquence et le niveau de pression sonore audible pour une oreille moyenne. Les fréquences sont perceptibles de 20 Hz à 20 kHz. En terme de niveau sonore,

il existe 2 seuils : le seuil de silence qui représente le seuil minimum pour qu'un son soit audible et le seuil de douleur. Ces deux seuils sont dépendants de la fréquence audio. Par ailleurs, il est également intéressant d'analyser la courbe isosonique. Cette courbe représente l'intensité sonore, en dB, nécessaire pour provoquer la même sensation d'intensité sonore pour l'oreille humaine, en fonction de la fréquence. La figure 2.2 montre la courbe isosonique pour un ensemble de niveaux sonores. Ces niveaux sont représentés en phones, correspondant au dB SPL pour une fréquence de 1000 Hz.

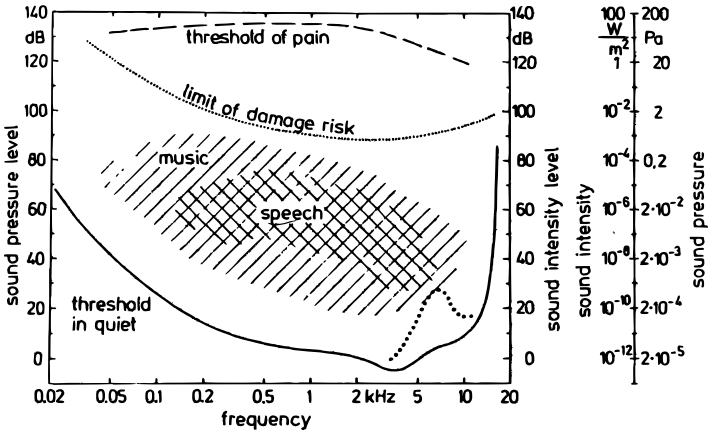


FIGURE 2.1 – Zone audible de l'oreille avec les seuils d'audition et de douleur ainsi que les zones de fréquence pour la musique et la voix. La figure provient de [Zwicker et Fastl, 2013].

2.1.2 Phénomène de masquage temporel et fréquentiel

En plus d'une sensibilité variable du système auditif en fonction des fréquences, un second phénomène rentre en jeu dans la perception sonore. Quand deux sons sont produits simultanément, un phénomène de masquage peut se produire. La source sonore la plus forte, la source masquante va cacher le son le moins fort, la source masquée. Il existe 2 formes de masquage différentes.

La première forme de masquage est appelée masquage fréquentiel. Il apparaît quand un son d'une certaine fréquence est émis au même moment qu'un son plus faible à une fréquence voisine. Le son le plus fort, le son masquant, va rendre inaudible le son plus faible. La fréquence et l'intensité jouent un rôle important sur l'étalement fréquentiel de ce masquage. Plus le son masquant sera fort plus la zone de fréquence voisine masquée va être importante. La figure 2.3(a) montre un schéma du phénomène avec un son masquant, ainsi que la courbe de masquage que doivent dépasser les autres sons pour être distingués par l'auditeur. Une étude détaillée sur les seuils de masquage selon l'intensité et la durée peut être trouvée dans [Zwicker et Fastl, 2013].

La seconde forme est appelée le masquage temporel. L'idée est la même que pour le masquage fréquentiel, mais cette fois selon l'axe temporel. Il apparaît quand un son fort est émis juste avant ou juste après un son plus faible. En plus de masquer les autres sons au moment où il est joué, le son masquant va également rendre inaudible les sons joués légèrement après. Ce masquage est appelé *post-masking*. De manière moins intuitive, ce son va également affecter la perception des sons

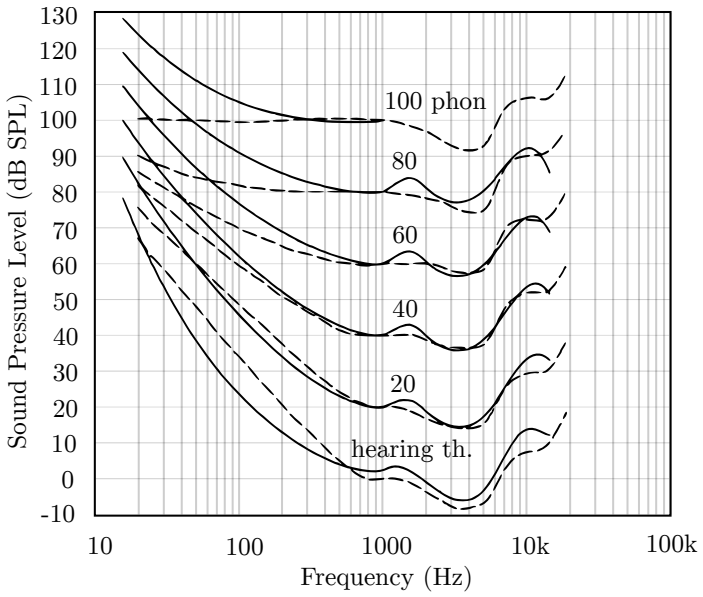


FIGURE 2.2 – En trait continu : courbes isosoniques issues de [ISO 226, 2003]. Trait discontinu, courbes de Fletcher-Munson pour comparaison. La figure provient de [Suzuki *et al.*, 2003].

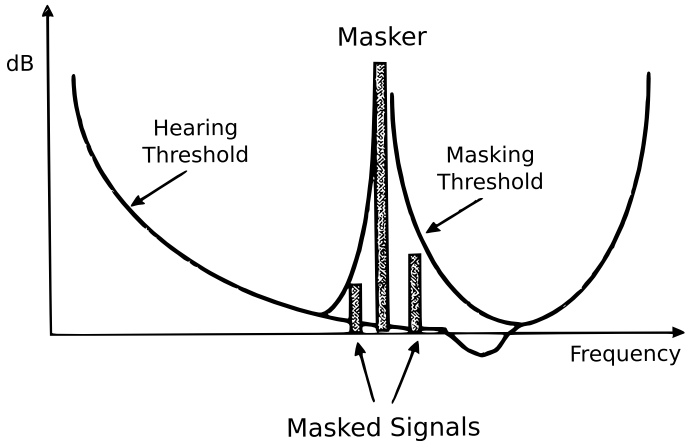
émis avant lui. Ce masquage, beaucoup plus court, est appelé *pre-masking*. La figure 2.3(b) donne une vision schématique du phénomène, ainsi que l'ordre de grandeur de la durée du phénomène de masquage. Les deux phénomènes de masquage ne sont pas exclusifs, le seuil de masquage doit être visualisé comme une surface 2D qui varie en fonction du temps et de la fréquence.

2.1.3 Les bandes critiques

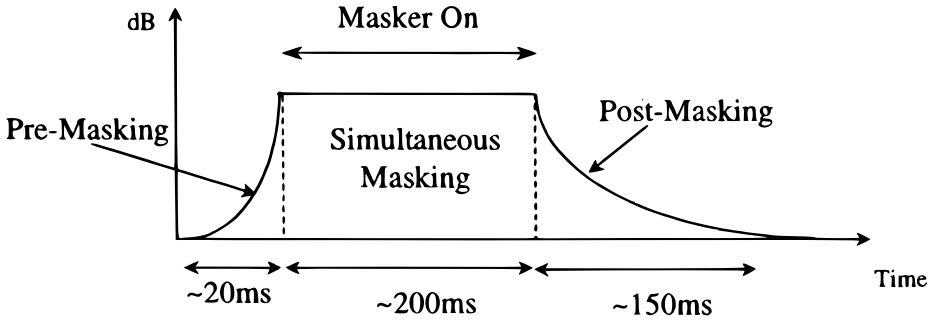
Le concept de bandes critiques a été introduit par [Fletcher, 1940], les bandes critiques servent à modéliser la manière dont le système auditif traite les sons. Les bandes critiques découpent le spectre sonore en plusieurs bandes de fréquences. La taille des bandes varie en fonction de la fréquence, plus la fréquence est élevée plus la bande de fréquence est large. Les bandes critiques sont liées au phénomène de masquage. Un son masquant dans une certaine bande critique impactera plus fortement les fréquences de la même bande critique que les fréquences des autres bandes. De plus, le découpage en bandes critiques permet d'avoir une découpe homogène du spectre d'un point de vue perceptif. Dans cette échelle, une distance perceptive entre deux bandes est perçue de manière identique que la fréquence soit basse ou haute, là où pour l'échelle des fréquences la distance perceptive entre 100 et 200 Hz n'est pas la même qu'entre 1000 Hz et 1100 Hz.

L'échelle de Bark

Plusieurs découpages en bande critique existent. L'un des plus utilisé en codage est l'échelle de Bark [Zwicker et Fastl, 2013]. L'échelle se divise en 24 sous-bandes allant de 20 Hz à 15,5



(a) Masquage fréquentiel.



(b) Masquage temporel.

FIGURE 2.3 – Les deux types de masquages fréquentiels. La figure provient de [Bosi et Goldberg, 2012].

kHz. Les sous-bandes sont découpées selon une échelle pseudologarithmique. Les fréquences de coupure $z(f)$ de chaque sous-bande sont indiquées dans des tables. Il existe également des formules permettant la conversion approchée entre l'échelle des fréquences et l'échelle de Bark. La première approximation a été faite par la formule présentée dans [Zwicker et Fastl, 2013] :

$$Bark(f) = 13 \arctan\left(\frac{0{,}76f}{0{,}001}\right) + 3{,}5 \arctan\left[\left(\frac{f}{0{,}0075}\right)^2\right] \tag{2.1}$$

où f correspond à la fréquence en Hz et $Bark(f)$ à la valeur de l'échelle de Bark. La valeur retournée par la formule permet de convertir les fréquences de l'échelle des fréquences vers l'échelle de Bark. Cette valeur arrondie à l'entier supérieur permet de savoir à quelle bande de Bark cette fréquence appartient.

Autres échelles perceptives

En parallèle à l'échelle Bark, d'autres échelles existent, comme l'échelle Equivalent Rectangular Bandwidth (ERB) [Moore et Glasberg, 1996] ou l'échelle Mel. L'échelle ERB propose une autre approximation du système auditif, notamment avec un découpage différent par rapport aux bandes de Bark en dessous de 500 Hz. L'échelle de Mel [Stevens *et al.*, 1937] a été initialement créée pour le traitement de la musique avec une distance égale entre les notes de musique, son utilisation s'est peu à peu étendue au reste du domaine audio. Cette échelle permet un découpage perceptif des fréquences selon une échelle logarithmique. Elle a l'avantage de ne pas avoir de contrainte sur les bornes de fréquences ainsi que sur le nombre de bandes, ce qui rend cette échelle plus souple que l'échelle de Bark.

2.1.4 Localisation du son

2.1.4.1 Indices de localisation

Le système auditif est capable de localiser la position de sources dans l'espace. La localisation est obtenue par la captation du son par nos deux oreilles et les traitements effectués par le cerveau pour en déduire un certain nombre d'informations sur la source sonore.

La localisation de la source, en terme d'azimut, a été principalement attribuée à deux indices binauraux [Blauert, 1997] : la différence d'intensité interaurale, ou Interaural Level Difference (ILD), c'est-à-dire la différence d'intensité du signal entre l'oreille droite et l'oreille gauche et la différence de temps d'arrivée, ou Interaural Time Difference (ITD), c'est-à-dire le retard du signal entre les deux oreilles. La figure 2.4 montre une source sonore ponctuelle arrivant d'une direction donnée. Le son produit par cette source va arriver l'oreille droite avant l'oreille gauche, ce qui va produire une différence de temps d'arrivée ΔT entre les deux oreilles. Pour arriver à l'oreille gauche, le son a parcouru une plus grande distance, le son en plus d'arriver avec un retard aura également une différence d'intensité ΔI, entre les deux oreilles. De manière moins importante, la tête engendrera un phénomène de filtrages et de diffractions selon les fréquences, ce qui donnera des informations supplémentaires au cerveau pour déterminer la position de la source [Blauert, 1997].

La location d'une source sonore, en terme d'élévation, est quant à elle déterminée principalement par le filtrage du son engendré par la forme de la tête, de l'oreille ainsi que dans une moindre mesure le haut du corps. L'onde sonore qui provient de la source va être réfléchie et diffusée différemment par le corps de l'auditeur selon la direction d'arrivée de l'onde. Pour chaque être humain, les filtres associés à chaque direction sont appris à la naissance. Ces filtrages évoluent au cours de la vie, le cerveau adapte en permanence son estimation de filtrage [Honda *et al.*, 2007]. Ces filtrages sont dépendants de la morphologie de chaque personne [Blauert, 1997]. La mesure des HRTF de l'individu permet de mesurer les fonctions de transfert liées à sa morphologie.

2.1.4.2 Erreur de localisation

Malgré les mécanismes mis en œuvre pour la localisation de source, le système auditif humain peut faire des erreurs sur la position des sources dans l'espace. De très nombreuses études se

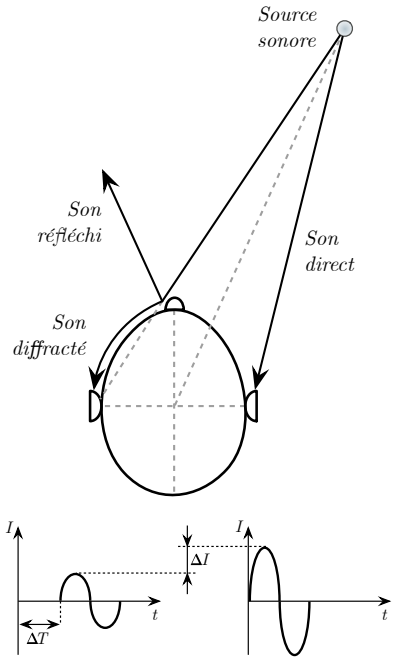


FIGURE 2.4 – Différence d'intensité et de temps d'arrivée d'une source ponctuelle.

sont penchées sur les performances de localisation du système auditif [Rayleigh, 1907, Perrott et Saberi, 1990, Damaske et Wagener, 1969]. Ces études ont demandé à un panel de sujets de localiser une source dans l'espace, en azimut ou en élévation, puis de comparer leurs réponses avec la localisation réelle. Selon les études différents stimuli ont été utilisés : bruit blanc, son à large bande, son à bande étroite ou un sinus pur à différentes fréquences. En fonction des stimuli utilisés, les performances de localisation en azimut et en élévation varient de manière significative. Ce qui suggèrerait que le système auditif ne perçoit et ne traite pas tous les sons de la même manière. Dans sa thèse [Daniel, 2011, Chap. 1], Daniel fait une synthèse détaillée de l'ensemble de ces méthodes. Même si les valeurs varient d'une étude à une autre, il est possible de dégager des tendances et des ordres de grandeur.

Sur le plan azimutal, la localisation pour les sources venant de l'avant a une erreur moyenne de l'ordre de 3-4°, alors qu'elle est de 5-6° pour les sources venant de l'arrière. Pour les sources latérales, l'erreur est plus imprécise avec une erreur moyenne de localisation de l'ordre de 8-10°. La moindre performance pour les sources latérales venant du fait que le filtrage de la tête est maximal pour cette position. Le son arrive avec un niveau plus faible à la seconde oreille, ce qui ajoute de l'imprécision au niveau des indices binauraux.

Pour l'élévation, la localisation est moins précise que pour l'azimut avec une erreur moyenne de l'ordre de 10-15°. Il est à noter que le type de stimulus semble jouer un rôle plus important dans l'estimation de l'élévation, ce qui semble renforcer l'hypothèse que l'estimation de l'élévation est déterminée par un filtrage fréquentiel [Daniel, 2011, Chap. 1].

2.1.4.3 Démasquage spatial

À la section 2.1.2, les phénomènes de masquage temps-fréquence ont été présentés dans le cadre d'un signal mono. Pour les contenus où les deux oreilles ne reçoivent pas le même signal audio (stéréo ou multicanaux) de nouveaux mécanismes de masquage semblent entrer en jeu. Par exemple, des fréquences masquées dans un signal lors d'une écoute en mono peuvent être audibles lors de l'écoute des deux canaux stéréo. Traiter chaque canal séparément peut donc produire des altérations audibles du signal. Dans les travaux de [Blauert, 1997], des tests ont montré que la direction d'arrivée des sources jouait un rôle important pour le masquage. L'étude met en évidence que deux sources venant de la même direction pouvaient se masquer alors que les mêmes sources venant de deux directions différentes pouvaient être dissociées. Le même phénomène a aussi été mis en évidence pour la distance des sources avec l'auditeur [Shinn-Cunningham *et al.*, 2001]. Dans des études plus récentes [Shinn-Cunningham, 2005], des liens ont été établis entre orientation de la tête et démasquage spatial. Le système auditif n'ayant pas la même précision spatiale suivant la direction de la tête, deux sources non distinguables lorsqu'elles sont localisées toutes les deux à gauche, deviennent distinguables si elles sont placées de manière frontale à l'auditeur. Un dernier phénomène à prendre en compte dans le phénomène de démasquage est l'attention spatiale. Dans [Arbogast *et al.*, 2005], les auteurs ont étudié l'influence de l'attention de l'auditeur sur la scène sonore. Selon l'endroit où l'auditeur focalise son attention dans la scène sonore, les sources à cet endroit deviennent plus distinguables.

L'ajout de la dimension spatiale au contenu audio a amené un nombre important de nouveaux facteurs perceptifs à prendre en compte pour le codage. Malgré des études qui tentent de mieux caractériser ces phénomènes, il n'existe malheureusement pas encore de modèle perceptif permettant de prendre en compte l'ensemble des facteurs pour fournir une modélisation complète du masquage spatial.

2.2 Évaluation de la qualité audio

L'évaluation de la qualité audio est un point essentiel dans le domaine de la compression audio avec perte où le but des méthodes de codage est de minimiser la quantité de données à transmettre tout en limitant le plus possible les dégradations perçues du signal codé par rapport à un signal d'origine. La fiabilité des méthodes d'évaluation est donc cruciale, car elles sont la seule manière de comparer la qualité audio produite par les différentes méthodes. Pour comparer les dégradations, il existe 2 grandes familles de méthodes d'évaluation. La première est la famille des méthodes subjectives, où le but va être de faire évaluer par un panel de sujets la qualité audio. Chaque sujet va procéder à un test d'écoute pour comparer et noter différentes conditions de codage du même signal d'origine. Les résultats sont ensuite étudiés avec des analyses statistiques permettant de dégager des tendances et des différences significatives. Ces méthodes nécessitant un nombre important d'auditeurs, leurs mises en œuvre sont donc souvent assez lourdes et contraignantes.

La seconde famille est la famille des méthodes dites objectives. Leur but est d'estimer la qualité audio subjective en se basant sur la mesure de certains indicateurs et éléments du signal. Dans le domaine de la compression audio, le signal d'origine est généralement accessible. Il est donc possible de faire une comparaison des valeurs des indicateurs du signal d'origine et du signal

codé pour estimer la dégradation de ce dernier. Pour prendre en compte la perception humaine, certaines métriques intègrent une modélisation de la perception. Cela permet d'avoir des résultats plus proches des résultats qu'il est possible d'obtenir avec des tests subjectifs.

L'avantage des méthodes objectives est qu'il est possible d'estimer la qualité perçue par un sujet. De plus, ces méthodes permettent de réaliser un grand nombre de tests. Cela permet de facilement comparer des variantes d'un même codec ou de faire de l'exploration de paramètres dans le but de trouver les meilleurs réglages d'un codec. Cependant, les modèles perceptifs ne sont qu'une approximation de la perception humaine, des divergences peuvent être présentes entre les résultats des mesures objectives et subjectives.

Pour déterminer la fiabilité d'une métrique objective, plusieurs indicateurs sont à analyser [ITU-T P.1401, 2020], dont la corrélation entre les résultats produits par la métrique et les résultats subjectifs. Plus la corrélation entre les deux scores est importante, plus la métrique est jugée fiable. L'évaluation de la fiabilité d'une métrique objective est cruciale, car de nombreux biais (signal audio utilisé lors de l'élaboration de la méthode, représentativité et culture du panel) peuvent intervenir lors de l'élaboration de la méthode et de sa validation.

2.2.1 Test subjectif

Il existe différentes méthodologies de test pour évaluer la qualité, en fonction du type de dégradation du signal, et du critère qui est évalué, certaines méthodes sont plus adaptées que d'autres. Pour permettre une reproductibilité plus importante des résultats ainsi qu'une compréhension des résultats plus aisée, les protocoles d'évaluation sont, généralement, définis par des standards. Ces standards définissent, le protocole d'évaluation ainsi que les conditions de tests (traitement acoustique de la salle, dispositif d'écoute...).

Dans le monde de l'industrie et plus particulièrement dans le monde des télécommunications, les recommandations de différents standards sont définies par l'Union Internationale des Télécommunications (UIT). Selon le type de contenu sonore, ainsi que la quantité de dégradation apportée par le système de codage, plusieurs méthodologies sont recommandées. Les méthodes subjectives peuvent être utilisées dans différentes configurations d'écoute : écoute au casque, sur haut-parleurs... En fonction de la configuration d'écoute, les recommandations donnent également des consignes en terme de volume, de temps de réverbération et de traitement acoustique que doit respecter la salle de test pour pouvoir être considérée comme valide.

Pour les écoutes binaurales, de nouvelles questions se posent. Quel moteur de rendu utilisé ? C'est-à-dire quels filtres HRTF et optimisations utilisés pour faire le rendu binaural : HRTF de référence, HRTF personnel, décodage simple, décodage max-r_E. Mais aussi la possibilité d'avoir un rendu interactif de la scène, où la scène sonore bouge avec les mouvements de la tête du sujet, configuration appelée "tête mobile". À l'inverse le test peut être conduit avec un rendu fixe, quelque soient les mouvements de la tête du participant, appelé "tête fixe". Ces nouvelles possibilités amènent de nouvelles questions méthodologiques comme l'influence du retard entre le mouvement et le son généré *Motion-to-Sound* [Estrella et Plogsties, 2017], ou la dépendance des résultats au moteur de rendu binaural utilisé. Les méthodologies actuelles n'apportent pas encore de recommandations pour répondre au besoin de la compression de contenu immersif.

2.2.1.1 Méthodologie MUSHRA

La méthodologie MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) est définie par la recommandation [ITU-R BS.1534, 2015]. Cette méthode est dédiée à l'évaluation subjective de signaux dits de qualité intermédiaire, où la dégradation des signaux évalués est importante par rapport aux signaux de références.

Pour être valides, les recommandations demandent que les sujets au test soient des personnes expérimentées dans le domaine audio. De plus, les échantillons audio, aussi appelés stimuli, utilisés dans le test doivent être des stimuli critiques vis-à-vis de la méthode de codage évaluée. L'utilisation d'échantillons critiques permet de réduire le nombre d'échantillons dans le test, ce qui permet de limiter sa durée et ainsi de limiter la fatigue auditive des participants.

Chaque sujet du panel de participants, un par un, va noter les différents stimuli. Pour chacun des stimuli, le sujet va avoir à sa disposition le signal original qui va lui servir de référence. Le sujet va devoir noter la qualité de différentes versions du signal codé avec différents paramètres par rapport au signal de référence. Cette qualité est la qualité globale du signal englobant tout les aspects du signal comme le timbre, les artéfacts... Les notes vont de 0 à 100, l'échelle est découpée en 5 portions allant de mauvais (0-20) à excellent (80-100).

En plus des conditions évaluées, trois conditions de calibration sont ajoutées pour chaque stimulus. Une référence cachée pour permettre de vérifier que le sujet est capable de reconnaître le signal original et deux conditions d'ancrage sont ajoutées. Une ancre basse et une ancre moyenne qui doivent refléter le type de dégradation que peut apporter la méthode de codage. Les ancres doivent obtenir respectivement une note "mauvaise" et "moyenne". Ces conditions permettent au sujet d'entendre ce qui considéré comme des conditions fortement dégradées, de plus, cela le force à utiliser l'ensemble de l'échelle de notation.

La figure 2.5 montre l'interface de notation utilisée par les participants du test. Une fois l'ensemble des tests réalisés, la méthodologie MUSHRA prévoit une phase de post-sélection des sujets. Cette post-sélection permet de retirer les résultats des sujets n'ayant pas réussi à retrouver un nombre suffisant de références cachées pour les différents extraits. Pour les codecs mono, les ancres correspondent généralement au signal original filtré par un filtre passe-bas à 3,5 kHz et 7 kHz. L'ordre d'apparition des conditions et celui des stimuli est aléatoire pour chaque participant.

2.2.1.2 Recommandation ITU-R BS.1116

La recommandation [ITU-R BS.1116, 2015] est une méthodologie de test utilisée pour l'évaluation de la qualité de contenus audio présentant de faibles dégradations. La méthodologie partage les mêmes contraintes de la méthodologie MUSHRA en terme de conditions d'écoute, de dispositifs de reproduction sonore et de post-sélection des sujets. Toutefois, le déroulement du test et l'échelle de notation diffèrent. Pour chaque stimulus, l'auditeur doit évaluer trois conditions nommées A, B, C. La condition A correspond à la référence explicite. B et C sont la référence cachée et la référence codée par la méthode à évaluer. L'ordre de B et C est aléatoire. La notation est une échelle continue de dégradation à 5 échelons de : *dégradation très gênante (1)* à *dégradation imperceptible (5)*. Ce type d'échelle de notation de la dégradation est appelé Degradation Category Rating (DCR). La condition correspondant à la référence cachée doit être égale à 5 *dégradation*

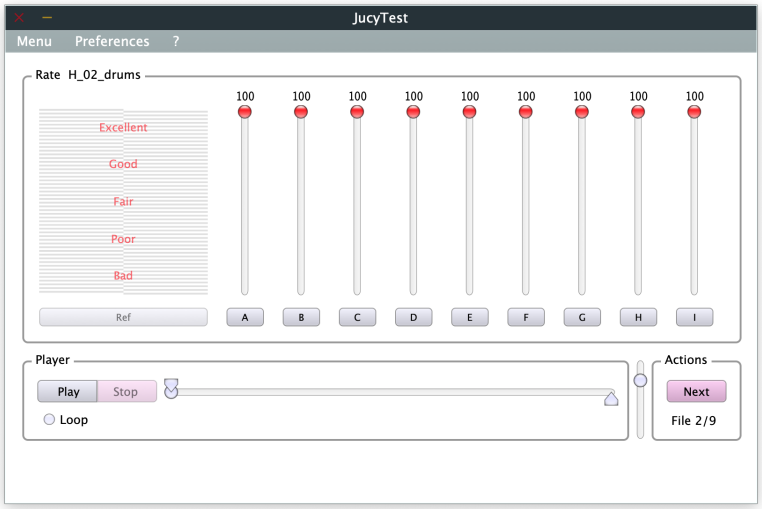


FIGURE 2.5 – Interface de test MUSHRA.

imperceptible. Une fois l'ensemble des tests réalisés, une post-sélection des sujets doit être faite pour écarter les résultats des sujets n'ayant pas réussi à trouver suffisamment de références cachées dans les conditions de test.

2.2.1.3 Méthodologie AB et Ref AB

Pour réaliser une comparaison de la qualité de deux échantillons, il est possible d'utiliser un test AB ou Ref AB. Il est demandé au participant de comparer la qualité d'une paire de stimuli, nommés A et B, ce test utilise une échelle du type Comparison Category Rating (CCR). Le score est exprimé selon le note d'opinion comparative moyenne, ou Comparative Mean Opinion Scores (CMOS). Il est généralement utilisé pour comparer la performance de deux méthodologies ou variantes de codage. Ce test est utilisé depuis de nombreuses années même si aucune recommandation propre ne lui est associée. Il est cependant possible de la rattacher à la recommandation [ITU-R BS. 1284, 2019], qui propose des briques de bases pour l'élaboration de tests subjectifs selon les besoins de chaque étude. La figure 2.6 montre l'interface utilisée pour les tests. Chaque stimulus correspond au même signal d'origine codé selon différentes conditions. À la conception du test, selon que les auteurs veulent réaliser une évaluation de fidélité (Ref AB) ou de préférence (AB), une référence explicite peut être incluse à côté des deux stimuli à évaluer.

Pour chaque échantillon audio, le participant doit noter la paire de stimuli selon une échelle discrète allant de -3 à $+3$. Un score positif signifie que la condition A est perçue comme de meilleure qualité que la condition B. Un score négatif signifie que la condition B est de meilleure qualité que la condition A. L'échelle de notation correspond à l'échelle de comparaison du tableau 2.1. Pour limiter tout biais lors du passage du test, l'ordre d'apparition des échantillons et des stimuli est aléatoire.

Qualité	Dégradation	Comparaison
5 Excellent	5 Imperceptible	3 Bien meilleure
4 Bon	5 Perceptible, mais non gênant	2 Meilleure
3 Moyen	3 Légèrement gênant	1 Légèrement meilleure
2 Médiocre	2 Gênant	0 Identique
1 Mauvais	1 Très gênant	-1 Légèrement plus mauvaise
		-2 Plus mauvaise
		-3 Bien plus mauvaise

Tableau 2.1 – Échelles proposées par la recommandation ITU-R BS.1284

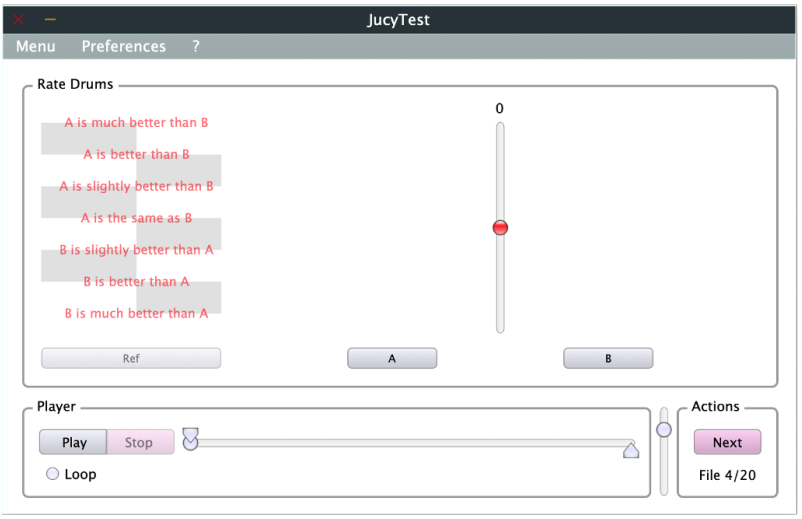


FIGURE 2.6 – Interface de test Ref AB

2.2.2 Test objectif

Le but des premiers codecs mono était de reproduire au plus proche la forme d'onde du signal origine, aussi appelé *waveform-matching*. Pour ces codecs, une simple mesure objective suffisait à comparer les dégradations engendrées par les codecs. Avec le raffinement des procédés de codage, notamment avec l'intégration de modèle perceptif, les métriques objectives ont dû elles aussi devenir plus sophistiquées.

2.2.2.1 SNR et SegSNR

L'une des métriques les plus connues est la métrique appelée rapport signal à bruit, ou Signal-to-Noise Ratio (SNR). Elle consiste en un calcul du rapport entre l'énergie du signal d'origine et l'énergie du bruit apporté par le codage. Le SNR peut être calculé dans le domaine temporel ou dans le domaine fréquentiel.

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_{i=1}^{N} x(i)^2}{\sum_{i=1}^{N} (x(i) - y(i))^2} \tag{2.2}$$

où x et y correspondent au signal d'origine et au signal codé et N au nombre d'échantillons du signal.

Le SNR donne un ratio moyen sur l'ensemble du signal, ce qui ne permet pas de déterminer si la dégradation du signal codé y est localisée dans le temps. Un bruit constant aura la même influence sur la qualité qu'une dégradation ponctuelle. Pour pouvoir prendre en compte ces variations temporelles, il est possible d'introduire une nouvelle métrique appelée rapport signal à bruit par segment, ou Segmental Signal-To-Noise Ratio (SegSNR). Dans cette métrique, le SNR n'est pas calculé sur l'ensemble du signal, mais sur des segments courts, généralement de 5 à 20 ms. En observant l'évolution du SNR d'un segment à l'autre, il est possible de détecter des dégradations ponctuelles du signal.

2.2.2.2 Métriques orientées parole

Les deux métriques présentées sont très sensibles à la présence de retard temporel, de distorsion ou d'une fluctuation de volume. Ces altérations peuvent entrainer une forte dégradation des scores de qualité, alors que d'un point de vue perceptif, ces altérations peuvent avoir un impact mineur. Avec l'amélioration des codecs et l'arrivée des modèles perceptifs à l'intérieur de ces derniers, de nouvelles métriques plus élaborées ont dû être proposées. Ces méthodes se sont divisées en deux groupes : les métriques orientée parole et les métriques orientée musique.

La méthode Perceptual Evaluation of Speech Quality (PESQ) est l'une des premières méthodes d'évaluation du signal de parole avec une prise en compte de la perception humaine. Présentée dans [Rix *et al.*, 2001], la méthode a été normalisée en 2001, elle a fait partie des méthodes d'évaluation de référence pour la compression de signal de parole bande étroite, ou Narrowband (NB), c'est-à-dire pour des signaux avec une bande de fréquence de 300 Hz à 3,4 kHz. Un mécanisme d'alignement du signal codé par rapport au signal d'origine est ajouté pour éviter la chute des scores lors de décalage entre les deux signaux. Un découpage du signal est fait selon l'échelle de Bark, et une égalisation des gains entre les trames et les gains sont réalisés. Un modèle psychoacoustique, assez rudimentaire, permet d'ajouter des masques temporel et fréquentiel pour pondérer l'importance de certaines altérations du signal. Une fois ces traitements réalisés, une *p-distance* est calculé entre les deux signaux. Initialisement conçu pour des signaux NB, la méthode PESQ a été par la suite étendu pour permettre de traiter des signaux de parole en large bande, ou Wideband (WB) sous le nom de Perceptual Evaluation of Speech Quality Wideband (PESQ-WB) [ITU-T P.862, 2005]. Cette extension permet de traiter des signaux allant de 80 Hz à 7 kHz.

À partir des méthodes déjà développées, la méthode Perceptual Objective Listening Quality Analysis (POLQA) [ITU-T P.863, 2018] a été normalisée en 2011. Elle a un fonctionnement assez proche de la méthode PESQ. La méthode POLQA se distingue par son plus grand nombre d'indicateurs pris en compte pour prédire le score de qualité. Dans ces indicateurs, certains sont ajoutés pour prendre un compte les dégradations apportées par la transmission par internet, comme la Voice Over Internet Protocol (VoIP) et le Web Real-Time Communication (WebRTC). POLQA propose également un mode de fonctionnement permettant d'estimer la qualité pour des signaux pleine bande, ou Fullband (FB), allant de 20 Hz à 20 kHz.

2.2.2.3 Métriques orientées musique

Parallèlement aux méthodes développées pour le codage de parole mono, des méthodes pour les signaux musique et son ambiant ont été développées. La méthode Perceptual Evaluation of Audio Quality (PEAQ) a été normalisée par l'ITU-R en 2001 [Thiede *et al.*, 2000]. Comme pour les méthodes d'évaluation du codage de la parole, un modèle perceptif est utilisé. Le modèle est optimisé pour mieux prendre en compte les sons musicaux. Pour être adaptée aux signaux musicaux, la méthode permet de traiter des signaux pleine-bande, FB échantillonnés à 48 kHz. Contrairement aux autres méthodes orientées parole, PEAQ permet la prise en compte des signaux stéréo et multicanaux. Un score de qualité est estimé pour chaque canal indépendamment. Puis la moyenne des scores est calculée pour obtenir le score final. Ce traitement indépendant de chaque canal induit le fait que les modèles perceptifs utilisés sont des modèles perceptifs mono notamment le masquage temps-fréquence.

2.2.2.4 ViSQOL et ViSQOLAudio

Plus récemment une nouvelle métrique objective nommée Virtual Speech Quality Objective Listener (ViSQOL) a été présentée [Hines *et al.*, 2015]. De manière similaire à la méthode POLQA [Beerends *et al.*, 2013], cette méthode a été conçue pour produire une meilleure estimation de la qualité dans les conditions de codage avec perte de paquets et déformation temporelle des signaux caractéristiques des communications par internet. D'abord déclinées en deux versions ViSQOL et *ViSQOLAudio*, respectivement pour la parole et la musique, les deux méthodes seront de plus en plus proches jusqu'à se fondre dans une appellation commune ViSQOL v3 speech et audio [Chinen *et al.*, 2020]. Cette dernière version de la métrique est *open-source*. Dans une étude [Sloan *et al.*, 2017] comparative des méthodes d'évaluation : ViSQOL, POLQA et PEAQ, les auteurs ont montré une corrélation plus importante entre les notes subjectives et ViSQOL que celle obtenue par les autres méthodes dans le cas de perte de paquets et lors d'activation de mécanisme de correction d'erreur. Pour les autres conditions, aussi bien parole que musique, les performances de ViSQOL sont similaires aux autres scores de corrélation obtenus par les autres méthodes.

Pour traiter les signaux stéréo, la méthode ViSQOL propose deux modes de fonctionnement. Dans le premier, le score est calculé pour le signal gauche et le signal droit de manière indépendante, puis le score maximum entre les deux est conservé. Dans sa version v3, le choix a été fait de faire un rematriçage des canaux stéréo gauche/droite vers une stéréo *Mid/Side*. Seul le canal *Mid* est conservé pour évaluer le score de qualité.

2.2.2.5 AMBIQUAL

Pour combler le manque de métrique objective pour le contenu audio spatialisé et notamment pour les contenus ambisoniques, une méthode nommée AMBIQUAL [Narbutt *et al.*, 2018] a été présentée récemment. Le format-B ambisonique est pris comme donnée d'entrée. Pour chaque composante ambisonique, un spectre de phase est extrait, appelé phasogramme. Les composantes sont divisées en deux groupes. La composante omnidirectionnelle W est utilisée pour déterminer la qualité audio *LQ*. Cette qualité d'écoute est mesurée par la méthode ViSQOL appliquée sur le phasogramme de la composante W du signal original et codé. Le second groupe de composantes,

constitué de toutes les composantes directionnelles, est utilisé pour mesurer la qualité de localisation *LA*. Cette précision finale est calculée à partir d'une somme pondérée de la similarité entre les paires de composantes directionnelles (originale et codée). Les deux indices *LQ* et *LA* sont étudiés séparément, l'un pour la qualité du timbre et l'autre la qualité spatiale.

Dans l'étude présentant la méthode [Narbutt *et al.*, 2018], des tests ont été menés pour étudier la fiabilité des scores de qualité produits par la méthode *AMBIQUAL* et la qualité subjective MUSHRA donnée par un panel de sujets. L'analyse statistique a montré une forte corrélation, selon la corrélation de Pearson et de Spearman, entre les résultats obtenus par les deux méthodes. Cependant, les résultats obtenus ne sont pas comparés avec la corrélation obtenue avec d'autres méthodes. De plus, l'étude ne se penche pas sur la corrélation entre les scores LQ et LA, ce qui ne permet pas de conclure sur l'intérêt d'utiliser deux scores distincts. Des travaux supplémentaires pour étudier la question de l'indépendance et la pertinence de conserver ces deux indices auraient besoin d'être menés. La relative jeunesse de cette méthode ainsi que le manque de travaux vérifiant les résultats obtenus par les auteurs de la méthode rend difficile l'utilisation d'*AMBIQUAL* dans l'élaboration de nouveaux codecs.

CHAPITRE 3

Codage audio

Sommaire du chapitre

3.1 Codage monocanal

3.1.1 Codage audio sans perte

La manière la plus simple de représenter un signal audio est la représentation Pulse Code Modulation (PCM). Cette représentation modélise le signal par un simple échantillonnage de la pression acoustique suivi par une quantification uniforme de ces échantillons. La qualité de cette représentation dépend de 2 facteurs, la fréquence d'échantillonnage et du nombre de bits utilisés pour quantifier le coefficient associé à chaque échantillon. Cette représentation du signal demande cependant une grande quantité de données. Par exemple, la qualité CD est obtenue avec 16 bits par échantillon et une fréquence d'échantillonnage de $44{,}1$ kHz, le débit de codage PCM d'un signal stéréo nécessaire est de $2 \times 16 \times 44100 \approx 1{,}4$ Mbit/s, ce qui est débit très important pour le codage audio.

Pour réduire le débit nécessaire pour transmettre le signal sans perte, les codeurs vont essayer de réduire la redondance du signal. Cette réduction de la redondance va être en partie réalisée par l'exploitation de la structure du signal. Cette structure est transmise en tant que métadonnée et la partie non prédictible du signal va être envoyée sous la forme d'un signal résidu. Lors du décodage, les métadonnées et la partie non prédictible du signal vont être réassemblées pour reproduire le signal d'origine.

3.1.2 Codage audio avec perte

Les méthodes de codage de parole pour les applications de téléphonie se sont d'abord appuyées sur des techniques de codage de forme d'onde de type PCM [Jayant et Noll, 1984] pour la qualité téléphonique dite *toll quality*. Pour les bas débits, les approches ont consisté à utiliser des modèles paramétriques de type Linear Prediction Coding (LPC) ou codage sinusoïdal [Kleijn et Paliwal, 1995]. Pour l'approche LPC, la structure du signal est modélisée à l'aide d'une prédiction linéaire [Marple, 1980]. Pour chaque trame, les coefficients d'un filtre IIR vont être déterminés pour approximer l'enveloppe spectrale du signal, ce filtre correspond à la prédiction à court terme du signal. Les coefficients vont être envoyés sous forme de métadonnées à chaque trame. L'erreur de prédiction du signal, aussi appelée résidu de prédiction, peut être codé par un codeur entropique type codeur de Huffman ou codeur arithmétique ou bien avec un codeur avec perte. Dans ce cas, la reconstruction du signal d'origine ne sera que partielle.

Par la suite, l'approche qui s'est imposée dans l'état de l'art est constituée d'approche hybride avec un modèle paramétrique type LPC estimé grâce à une analyse par synthèse avec principalement des variantes du codage Code-Excited Linear Prediction (CELP) [Schroeder *et al.*, 1979]. Le codage CELP a donné la majorité des standards de codage de parole jusqu'aux débuts des années 2000 avec le codec *G.729* pour la téléphonie fixe et *AMR*, *AMR-WB* pour la téléphonie mobile. Une prise en compte sommaire de la perception humaine est faite dans ces approches par l'application d'un filtre de pondération perceptuelle des coefficients LPC.

En parallèle, le codage musique et audio haute qualité s'est développées autour des techniques de codage en sous-bandes et par transformée Quadrature Mirror Filters (QMF) et Modified Discrete Cosine Transform (MDCT) avec des outils complémentaires (commutation de bancs de filtres ou de fenêtrages, mise en forme du bruit ...). Des exemples de codecs sont donnés par les standards *MPEG*, comme le codec *MPEG 1* et *2 Layer III* ou *MPEG 4 AAC*. Dans ces codecs, un des éléments clés pour réduire le débit va être la prise en compte de la perception humaine à l'aide de modèle perceptif. Les modèles perceptifs [Painter et Spanias, 2000] déterminent les parties du spectre sonore qui vont être inaudibles, soit parce qu'elles sont masquées par d'autres fréquences, soit parce que l'énergie de ces parties est en dessous des limites du seuil d'audition. Ces parties du spectre peuvent être supprimées sans altérer la qualité subjective perçue, ce qui permet d'économiser du débit. Une seconde stratégie de ces codecs est la mise en forme du bruit de codage, c'est-à-dire travailler dans des représentations où les approximations ajoutées par le codeur, aussi appelé bruit de codage, restent inférieures au seuil de masquage et donc sont inaudibles pour l'auditeur.

Les techniques d'extension de bande ont ensuite permis d'atteindre les très bas débits. Dans ces techniques, dont la plus connue est Spectral Band Replication (SBR) [Dietz *et al.*, 2002], le

principe consiste à transmettre uniquement la partie basse du signal, typiquement de 0 à 6 kHz. Puis, à partir de l'information de l'enveloppe et des paramètres du signal d'origine, reconstruire la bande haute du signal. Ces informations sont calculées dans le codeur et transmises sous forme de métadonnées. Pour effectuer la reconstruction des hautes fréquences, le décodeur va répliquer la bande basse au niveau de la bande haute. Puis à l'aide de l'information de l'enveloppe transmise par le codeur pour la bande haute, le signal répliqué est ajusté pour que la bande haute du signal décodé ait la même répartition d'énergie que le signal d'origine en hautes fréquences. La figure 3.1 montre les étapes de fonctionnement de la méthode. Le débit alloué pour coder le signal ne permet pas de maintenir le bruit de codage sous le seuil de masquage. Le signal est ensuite tronqué, seule la partie basse est conservée, le débit ainsi libéré permet de limiter le bruit de codage. Au décodage, le signal en bande basse et répliqué en bande haute. Grâce à l'enveloppe du signal d'origine, la forme du signal générale est reproduite. Le SBR est une brique de base qui peut-être intégrée dans n'importe quel codec perceptuel. Des versions plus avancées de la méthode ont été intégrées dans des codecs comme dans le codec High-Efficiency Advanced Audio Coding (HE-AAC) [Wolters *et al.*, 2003] ou plus récemment le codec Enhanced Voice Services (EVS) [Dietz *et al.*, 2015].

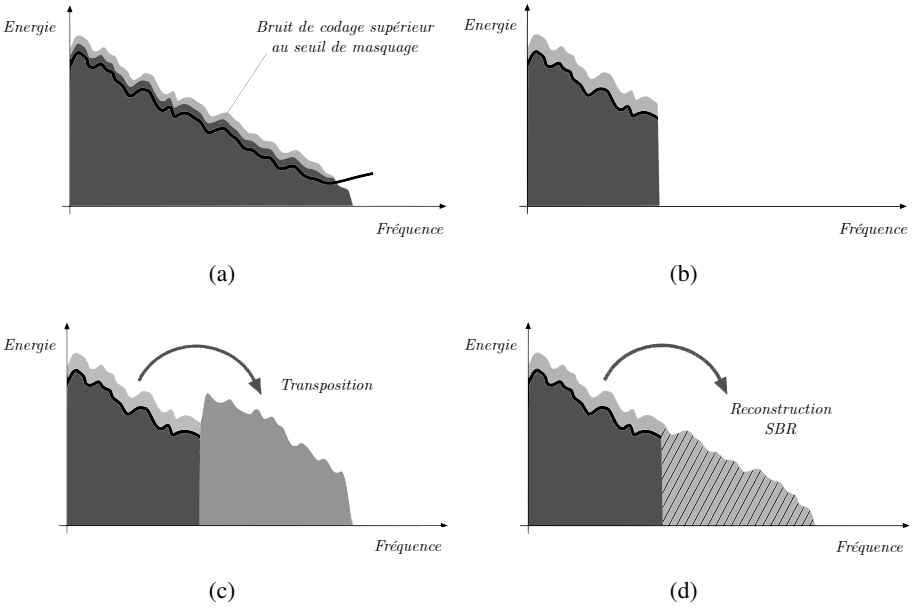


FIGURE 3.1 – Fonctionnement de la réplication de bande spectrale. Figure traduite de [Dietz *et al.*, 2002]

À partir des années 2000, des approches de commutation entre codage temporel par prédiction type CELP et codage fréquentiel par transformée type MDCT sont apparues [Tancerel *et al.*, 2000] comme *AMR-WB+* et *ATCELP*. Cette approche a été reprise dans des standards *MPEG-D USAC* et EVS. Les codecs actuels sont constitués d'un ensemble de modes de codage et de modules de pré-traitement et post-traitement. Selon le débit et le contenu audio à coder, une sélection du ou des modules les plus adaptés est réalisé. La figure 3.2 montre un schéma bloc haut niveau du codec EVS.

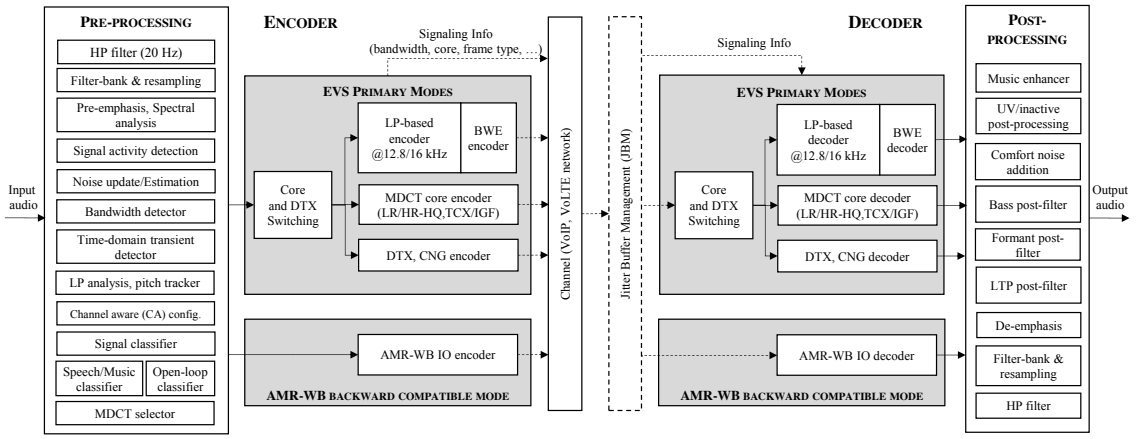


FIGURE 3.2 – Schéma fonctionnel du codec EVS. La figure est issue de [Dietz *et al.*, 2015]

3.2 Codage stéréo

Les méthodes de compression mono présentées précédemment permettent de réduire la quantité d'informations à transmettre en diminuant la redondance intra-canal. Pour la compression stéréo, une approche simple pour coder un signal stéréo est de considérer comme deux canaux mono séparés et de coder chaque canal avec une instance de codec mono indépendant. Cette approche est appelée dual-mono. Une telle approche permet de profiter des dernières avancées et optimisations faites dans le domaine de compression mono. Cependant, cette approche ne prend pas en compte les redondances d'informations qu'il peut y avoir entre les deux canaux. Pour les signaux stéréo, plutôt que de coder chaque canal indépendamment, des méthodes de compression stéréo proposent de réduire la quantité de débit nécessaire en exploitant la redondance inter-canal.

3.2.1 Codage stéréo par matriçage fixe

Pour les signaux stéréophoniques, les deux canaux sont généralement corrélés entre eux. Ce qui signifie qu'une partie de l'information contenue dans le canal droit et aussi présent dans le canal gauche. L'approche dual-mono, en traitant chaque canal de manière indépendante, va transmettre la même information deux fois, ce qui correspond à un surcoût en terme de débit.

Pour le signal stéréophonique, la méthode *Mid/Side Stereo Coding* [Johnston et Ferreira, 1992] propose un rematriçage des canaux *gauche-droite* vers une représentation *somme-différence*. L'idée de cette approche est de convertir les canaux stéréo vers une représentation où la corrélation entre les deux canaux est moindre. L'information commune entre les deux canaux stéréo se retrouve alors uniquement dans le canal *Mid*. Le rematriçage proposé est défini tel que :

$$\begin{bmatrix} c_m(t) \\ c_s(t) \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} s_l(t) \\ s_d(t) \end{bmatrix} \tag{3.1}$$

où s_l et s_d sont respectivement le signal gauche et droit et c_m et c_s le signal *Mid* et *Side*. L'entropie de la paire de canaux c_m/c_s est plus bas que la paire s_l et s_d, le débit nécessaire pour transmettre

les signaux rematricés est donc moins important que les signaux d'origine. Les signaux c_m et c_s sont ensuite codés par deux codeurs mono indépendants, appelés codecs cœur. Selon la quantité d'information contenue dans le canal c_m et c_s, le débit alloué à chaque canal peut être différent. De manière générale, un débit plus important est donné au canal *Mid* c_m par rapport au canal *Side* c_s. Au décodage, un matriçage inverse est appliqué aux signaux codés pour retrouver les signaux gauche et droit d'origine.

3.2.2 Codage stéréo paramétrique

3.2.2.1 Codage stéréo d'intensité

Le codage d'intensité, présenté dans [Herre *et al.*, 1994], se base sur le fait que la perception des composantes hautes fréquences est principalement liée à L'ILD. Dans leurs méthodes, les auteurs proposent pour la partie haute fréquence de transmettre uniquement un seul signal, la somme du canal gauche et droit c_m. Lors du décodage, la partie haute du signal c_m est spatialisée entre les canaux gauche et droit. Cette partie hautes fréquences est découpée en plusieurs sous-bandes. Pour chaque sous-bande, un gain est calculé en fonction de l'ILD entre le canal gauche et le canal droit.

Le gain permet au décodeur de réaliser un panoramique d'intensité, ou *Intensity Panning*, du signal haute fréquence. Le codage d'intensité peut être considéré comme une méthode paramétrique, car seule une partie des canaux est transmise. Le procédé qui permet de passer d'un nombre canaux à un nombre moins important est appelé *downmix*. La reconstruction des canaux manquants est faite en s'appuyant sur les métadonnées transmises, ce procédé est appelé *upmix*.

Dans le cas théorique où d'une source mono spatialisée artificiellement avec une panoramique d'intensité, la méthode permet une reconstruction parfaite du signal. Cependant, la technique ne permet qu'une reconstruction approximative de la spatialisation lorsque les canaux présentent une décorrélation. La plupart du temps, cette décorrélation est liée à l'utilisation d'un dispositif de prise de son stéréo non-coïncident (couple AB, arbre Decca...). Pour améliorer le codage de ces signaux stéréo, la méthode [Lindblom *et al.*, 2005] propose de faire un alignement temporel des canaux avant le codage d'intensité.

3.2.2.2 Codage BCC et parametric stereo

Les technologies Binaural Cue Coding (BCC) [Faller et Baumgarte, 2003] et *Pamateric Stereo* [Breebaart *et al.*, 2005] sont des méthodes de codage stéréo paramétrique. L'idée de ces méthodes est de faire un codage paramétrique où les indices binauraux seraient conservés entre le signal d'origine et le signal décodé. Plutôt que de transmettre les deux canaux, ces méthodes vont effectuer un *downmix* des signaux d'entrée pour transmettre uniquement un signal. Dans le codeur une extraction des indices binauraux est effectuée. Basé sur ces indices binauraux, le décodeur effectue un *upmix* pour recréer un signal stéréo. Nous présentons ici la structure de l'implémentation de *pamateric stereo* du codec *Enhanced aacPlus* [3GPP TS 26.401, 2007] car l'encodeur est également décrit contrairement au codec standardisé par MPEG, où seul le décodeur est décrit. La structure générale du codeur et du décodeur est donnée dans la figure 3.3.

Dans le codeur, figure 3.3(a), les canaux stéréo x_1 et x_2 vont passer dans un banc de filtres qui transforme en signaux temps-fréquences X_1 et X_2. À partir de ces canaux, les indices binauraux sont extraits. Les indices utilisés par le codec le codec *Enhanced aacPlus* sont : la cohérence inter-canal (ICC) et la différence d'intensité inter-canal (ICLD), la différence de phase inter-canal (ICPD) ou la différence de phase globale (OPD). Selon le codec et le mode de fonctionnement les indices binauraux utilisés peuvent être plus ou moins nombreux. Par exemple, le codec *Enhanced aacPlus* dans son profil basse complexité n'utilise que l'ICC et l'ICLD. Une fois les indices stéréo extraits, une réduction de canaux est effectuée pour créer le canal mono S. De manière générale, cette opération peut s'exprimer sous la forme :

$$S[f] = w_1 X_1[f] + w_2 X_2[f] \tag{3.2}$$

avec w_1 et w_2 les coefficients réels ou complexes qui permettent de choisir la projection de X_1 et X_2 pour former S. Pour un $w_1 = 0{,}5$ et $w_2 = 0{,}5$, le *downmix* va consister en une moyenne des signaux X_1 et X_2. Cependant, un *downmix* avec des poids fixes peut conduire à une variation forte de puissance du signal S en fonction de l'intercorrélation des deux canaux stéréo. Dans l'implémentation *Enhanced aacPlus* [3GPP TS 26.401, 2007] dans le mode *parametric stereo*, le codeur propose un raffinement du *downmix* par demi-somme des canaux en ajoutant un gain dynamique γ pour compenser les faibles variations d'énergie induite par l'opposition de phase :

$$S[f,b] = \frac{X_1[f,b] + X_2[f,b]}{2} \gamma[f,b] \tag{3.3}$$

avec b la sous-bande $b \in \{1, \ldots, B\}$. Une fois la réduction de canal réalisé, le signal obtenu est transformé dans un domaine temps-fréquence QMF à 64 bandes. Puis une extraction des paramètres SBR est effectuée. Un sous-échantillonnage est appliqué pour ne conserver que les 32 bandes fréquentielles inférieures. Une transformation inverse transforme le signal S dans le domaine temporel. Le signal s obtenu est ensuite codé par le codeur cœur mono AAC.

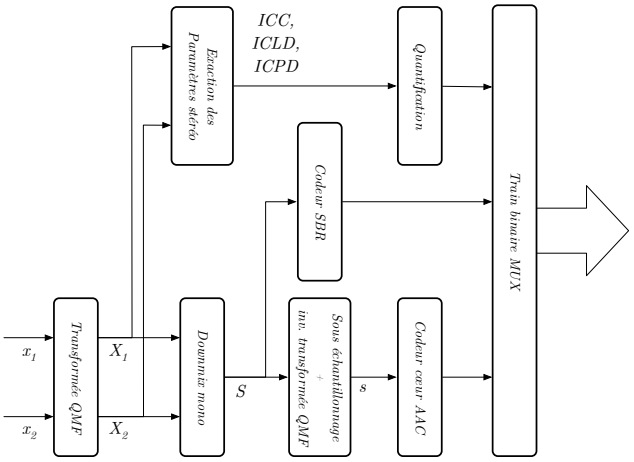
Dans le décodeur, figure 3.3(b), les traitements inverses sont appliqués au signal. Le signal mono est décodé par le décodeur cœur AAC. Puis à l'aide des paramètres SBR, une extension de bande est effectuée pour reconstruire la partie haute fréquence du signal. Un filtre passe-tout est appliqué au signal S pour créer une version décorrélée du signal D. Les signaux S et D sont ensuite fournis au module de synthèse spatiale *upmix*. À partir des paramètres spatiaux, le module recrée deux canaux stéréo qui sont ensuite transformés dans le domaine temporel.

3.3 Codage ambisonique

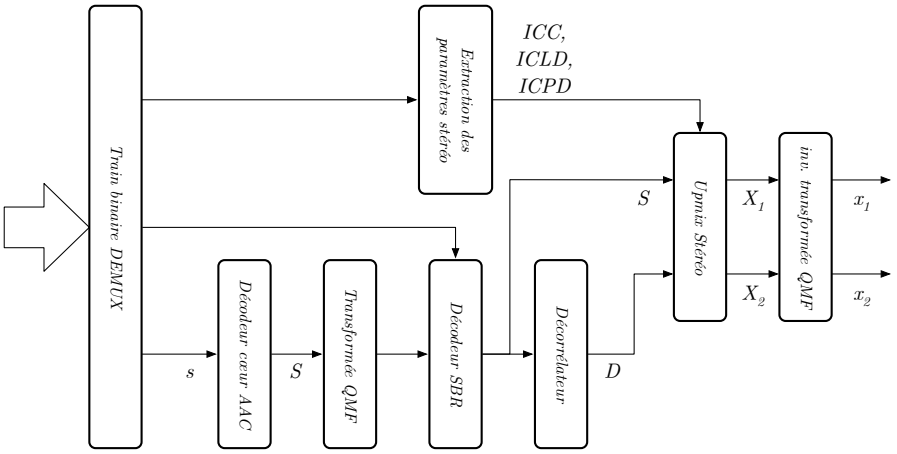
3.3.1 Approches basiques par matriçage fixe

3.3.1.1 Codage multimono

Nous considérons maintenant, non plus des signaux mono ou stéréo, mais des signaux ambisoniques. De manière analogue au codage stéréo, l'approche la plus basique consiste à considérer chaque composante comme un signal audio distinct, et d'appliquer un codage mono à chaque composante de manière séparée. Cette approche est appelée codage multimono. La figure 3.4 montre



(a) Codeur *parametric stereo*.



(b) Décodeur *parametric stereo*.

FIGURE 3.3 – Schéma du fonctionnement du codeur et du décodeur *parametric stereo*.

un schéma de fonctionnement de l'approche. Chaque composante ambisonique, au format-B, est transmise à une instance d'un codec cœur mono. Selon les besoins (téléphonie, haute fidélité, conversationnel ...), différents codecs cœur peuvent être utilisés. Le débit total est réparti de manière égale entre les différentes instances de codec cœur.

Un des avantages de cette méthode est qu'il est possible de profiter de toutes les avancées faites dans le domaine des codecs mono. De plus, le codage multimono a comme atout de ne pas apporter de retard de codage supplémentaire. Le retard global de la méthode correspond uniquement au retard apporté par le codec cœur. D'un point de vue de la complexité, l'approche multimono à une complexité 4 fois plus importante que le codec cœur mono utilisé, ce qui est un surcoût non négligeable. De plus, le codage indépendant provoque des déformations de la base ambisonique ce qui peut provoquer l'apparition d'artefacts spatiaux [Mahé *et al.*, 2019a].

Pour limiter l'apparition de ces artefacts et améliorer la qualité globale, la méthode [Brettle et Skoglund, 2016] propose de changer la répartition du débit entre les composantes, en transformant

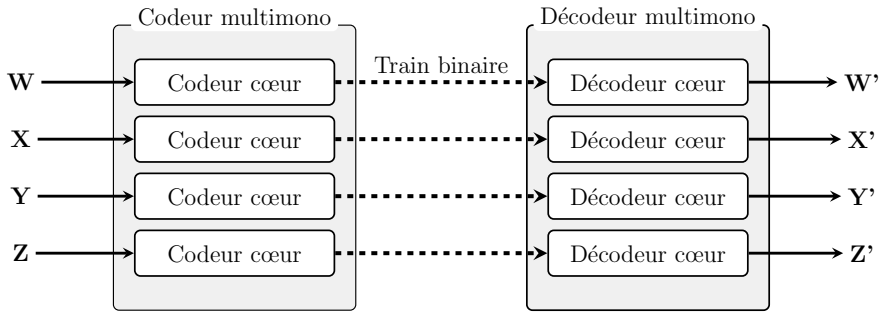


FIGURE 3.4 – Fonctionnement de l'approche multimono.

la répartition égale entre composantes vers une répartition où le débit est plus important pour coder la composante W. La méthode fait l'hypothèse que la dégradation de la composante W est plus préjudiciable pour la qualité globale que la dégradation des autres composantes. Comme pour la métrique *AMBIQUAL*, une distinction est faite entre la composante omnidirectionnelle et les composantes directionnelles X, Y, Z pour l'estimation de la qualité. Pour un signal FOA, la répartition du débit est faite selon le rapport : $\{1,\ 0{,}75,\ 0{,}75,\ 0{,}75\}$.

Inspirée des travaux de rematriçage *somme-différence* pour les signaux stéréo, la méthode [3GPP TS 26.918, 2018] propose de faire un rematriçage fixe des composantes FOA vers une représentation dans un format-A théorique, correspondant à une représentation ESD du format-B [Fliege et Maier, 1996]. Les 4 composantes sont rematricées, telles que :

$$\mathbf{A} = \mathbf{MB}$$
$$\begin{bmatrix} \text{FL} \\ \text{FR} \\ \text{BU} \\ \text{BD} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{12}} \\ \frac{1}{2} & \frac{-1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{12}} \\ \frac{1}{2} & 0 & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{12}} \\ \frac{1}{2} & 0 & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{12}} \end{bmatrix} \times \begin{bmatrix} \text{W} \\ \text{Y} \\ \text{Z} \\ \text{X} \end{bmatrix} \tag{3.4}$$

où $\mathbf{A}$ et $\mathbf{B}$ correspondent aux composantes ambisoniques respectivement dans le format-A et le format-B et $\mathbf{M}$ correspond à la matrice de transformation. La matrice $\mathbf{M}$ recombine les composantes pour former 4 faisceaux qui pointent dans les directions (θ, ϕ) : $(+54{,}7,\ 0)$, $(-54{,}7,\ 0)$, $(+180,\ +54{,}7)$,$(+180,\ -54{,}7)$

Le format-A est initialement un format qui correspond au signal enregistré par les 4 capsules de l'antenne microphonique. Les capsules sont placées sur les faces d'un tétraèdre. Pour désigner les 4 signaux issus des capsules, un nom est donné à chacune selon la position de la capsule : *Front Left (FL)*, *Front Right (FR)*, *Back Upward (BU)* et *Back Downward (BD)*. Dans la méthode, la même convention de nommage a été utilisée pour les composantes après rematriçage. Une fois converties dans le format-A, les composantes ambisoniques sont codées avec une approche multimono avec un débit équivalent pour chaque codec cœur. Après décodage, la transformation vers le format-A est inversée pour retrouver le format-B d'origine. Pour cela, la matrice $\mathbf{M}$ est inversée, puis elle est appliquée aux composantes au format-A, telle que : $\mathbf{B} = \mathbf{M}^{-1}\mathbf{A}$.

3.3.1.2 Codage multistéréo

Pour réduire la quantité d'informations à transmettre, la méthode [3GPP TS 26.918, 2018, Section 6] propose un codage FOA en utilisant le même type de rematriçage des composantes, mais en remplaçant le codage multimono par un codage multistéréo, aussi appelé *dual-stereo*. Le principe est de remplacer les 4 codecs coeurs mono par 2 codecs cœur stéréo. Une fois le signal converti dans le format-A, les composantes avants (*FL* et *FR*) sont codées par une instance du codeur cœur stéréo. Les composantes arrières (*BU* et *BD*) sont codées par une seconde instance du codeur cœur. Dans l'étude, le test de qualité subjectif a été réalisé avec le codec cœur *HE-AAC* v2. Cependant, la méthode peut être applicable à n'importe quel codec stéréo.

Le codec Opus dans son mode *Channel Mapping Family 3* [Skoglund, 2018] étend le codage multistéréo aux ordres supérieurs HOA. Comme pour la méthode précédente de multistéréo, un matriçage est appliqué aux composantes ambisoniques pour obtenir un format-A FOA ou HOA selon le signal d'entrée. La matrice de transformation n'est pas standardisée. Elle peut donc être choisie selon l'utilisation et les besoins. Dans l'implémentation de référence de *Opus* v1.3, une matrice est proposée pour chaque ordre ambisonique. Ces matrices correspondent à des discrétisations de la sphère. Elle transforme le signal ambisonique d'entrée vers un domaine spatial équivalent. Le nombre de canaux pour les deux représentations est identique. Quelle que soit la matrice utilisée, elle doit être présente dans la table contenant les matrices de transformation du codeur et du décodeur. L'indice de la table, qui correspond à la matrice utilisée, est transmis dans l'en tête de chaque trame de signal. La matrice de transformation détermine également l'appairage des composantes pour le codeur cœur stéréo. Chaque paire est codée avec le codec cœur stéréo d'Opus.

Pour son codage stéréo, Opus adopte une stratégie différente pour le codage de la bande basse et de la bande haute. Pour la bande basse, le codec code la forme d'onde du signal. Selon la similarité des deux canaux, la bande basse du signal est codée par un codage *somme-différence* ou un codage indépendant des deux canaux, appelé dans Opus : codage *Dual Stereo*. Pour la bande haute, un codage d'intensité est réalisé. Un seul signal est codé, cet unique signal codé correspond à la somme des deux canaux.

Lors du décodage, le signal est divisé en un ensemble de sous-bandes. Pour chaque sous-bande, le signal est respatialisé en utilisant un panoramique d'intensité pour répartir le signal de la bande entre les deux canaux. Dans le cas du codage stéréo traditionnel, ces deux canaux correspondent aux signaux des haut-parleurs gauche et droit. Dans le cas du FOA, l'approche multistéréo doit respatialiser le signal de chaque sous-bande dans une direction (θ, ϕ), chaque codec cœur stéréo définissant la direction de la source selon un axe. Le premier codec cœur va respatialiser le signal selon l'axe gauche/droite. Le second codec va quant à lui respatialiser le signal selon un axe haut/bas.

Dans le cas où une seule source est présente dans une sous-bande dans le contenu ambisonique d'origine, les deux panoramiques d'intensité permettront de respatialiser la source dans la bonne direction. Si plusieurs sources sont actives sur la même bande, chaque codec stéréo va faire un panoramique d'intensité correspondant à la direction moyenne sur l'axe gauche/droite ou haut/bas. Ces directions moyennes sont pondérées par l'énergie de chaque source.

Pour des scènes simples, avec au plus une source dominante dans chaque bande de fréquence,

l'hypothèse d'assimiler l'ensemble du signal à la direction moyenne par sous-bande semble être valide. Pour les scènes plus complexes, l'hypothèse est plus discutable, car la direction moyenne peut correspondre à une zone de l'espace, où il n'y avait pas d'énergie dans le signal d'origine.

En plus du codage multistéréo, le codec Opus propose également un codage multimono directement sur le format-B, appelé *Channel Mapping Family 2*. Selon le type de signal et de la scène sonore, il semble préférable d'utiliser l'un ou l'autre mode de codage.

3.3.2 Codage ambisonique paramétrique

3.3.2.1 Codage par prédiction

La méthode [3GPP TR 26.118, 2018] propose de réaliser une prédiction inter-canal pour le codage des composantes FOA. Puis de coder les résidus de prédiction avec un codage multimono. La prédiction s'adapte très bien au format ambisonique, car les composantes sont toutes des signaux coïncidents. Dans le codeur, une prédiction est faite entre la composante w et chacune des composantes directionnelles telle que :

$$\mathbf{B}' = \begin{bmatrix} w(t) \\ x'(t) \\ y'(t) \\ z'(t) \end{bmatrix} = \begin{bmatrix} w(t) \\ w(t) - \text{pred}_x(t) \\ w(t) - \text{pred}_y(t) \\ w(t) - \text{pred}_z(t) \end{bmatrix} \tag{3.5}$$

où $\text{pred}(t)$ est la prédiction faite pour chacune des composantes directionnelles à partir de la composante w. Pour chaque composante la prédiction est contrôlée par un gain de prédiction $g(t)$. La prédiction est formulée comme : $\text{pred}_x(t) = g_x(t) \times w(t)$. Pour une trame donnée, le signal va être découpé en K sous-bandes. Pour chacune des sous-bandes k, un gain $g(t,k)$ est déterminé. Ce gain $g_x(t,k)$ correspond au rapport d'énergie entre la composante w et la composante directionnelle x, tel que : $g_x(t,k) = \frac{x(t,k)}{w(t,k)}$. Le gain est quantifié puis transmis au décodeur en tant que métadonnée.

Une fois la prédiction réalisée, la composante w et les résidus de prédictions des composantes directionnelles x', y', z' sont codés par l'approche multimono. Pour la compression du FOA le codec alloue un débit deux fois plus important à la composante w. Le signal de la composante w apporte plus d'information que les résidus de prédictions, de plus une dégradation sur cette composante impacterait également la qualité de reconstruction des autres composantes x, y, z.

3.3.2.2 La méthode DirAC

La méthode DirAC [Pulkki *et al.*, 2018] propose une approche de codage paramétrique pour les signaux FOA. Il existe plusieurs variantes et modes de fonctionnement pour cette méthode, selon les cas, le nombre de canaux transmis, le découpage en sous-bandes ou le système de rendu peuvent varier. Nous nous intéressons dans cette section à la version décrite dans [Pulkki, 2007] sous l'appellation de *Dirac for telecommunication*. Dans cette variante, seule la composante W est transmise entre le codeur et le décodeur. À partir de cette unique composante et de l'information sur la scène sonore, le décodeur spatialise la composante décodée W' pour recréer un signal FOA

complet. Pour effectuer cette respatialisation, le décodeur s'appuie sur deux caractéristiques de la scène sonore : la direction de la source principale et le caractère diffus, ou *diffuseness*, de la scène. Ces caractéristiques sont estimées par le codeur sur le signal FOA d'origine, puis cette information est codée et transmise en tant que métadonnée au décodeur. La méthode est dédiée au codage ambisonique bas débit.

Pour estimer ces caractéristiques, le codeur se base sur le vecteur intensité acoustique, présenté à la section 1.2.5. Le signal ambisonique est découpé en trames avec un recouvrement de 50 %. Pour chaque trame, une transformée de Fourier est calculée pour chacune des composantes ambisoniques. Le vecteur intensité acoustique $\mathbf{I}_t(f)$ est calculé pour chaque carreau fréquentiel f. Le vecteur intensité est à valeurs complexes, cependant seule la partie active est prise en compte par la méthode First-Order Directional Audio Coding (DirAC). Le vecteur intensité active est calculé par :

$$\mathbf{I}_t(f) = \Re \begin{bmatrix} W_t(f)X_t^*(f) \\ W_t(f)Y_t^*(f) \\ W_t(f)Z_t^*(f) \end{bmatrix} \tag{3.6}$$

où $W_t(f)$, $X_t(f)$, $Y_t(f)$, $Z_t(f)$ sont les transformées de Fourier des composantes ambisoniques W, X, Y, Z pour la trame t. L'opérateur $(.)^*$ désigne le conjugué complexe et $\Re$ la partie réelle.

Les valeurs du vecteur intensité sont regroupées en sous-bandes de fréquence. Pour chacune des sous-bandes b, une direction d'arrivée est calculée. Il est possible d'estimer la direction prédominante de la trame à partir du vecteur intensité par la formule :

$$\text{DOA} = \angle\mathbb{E}[-\mathbf{I}] \tag{3.7}$$

avec $\angle$ l'opérateur qui donne l'angle (θ, ϕ) associé au vecteur. L'opérateur $\mathbb{E}[.]$ désigne l'espérance mathématique. Cet opérateur permet de calculer la moyenne des valeurs du vecteur intensité sur l'ensemble de la sous-bande. Pour rendre l'estimation plus stable à travers le temps, les valeurs (θ, ϕ) de la DOA peuvent être lissées d'une trame sur l'autre par un filtre type exponentiel. La DOA permet donc d'obtenir la direction (θ, ϕ) de la source principale de la scène pour chaque sous-bande.

Dans la méthode DirAC, l'hypothèse est faite que le champ diffus est supposé homogène, l'énergie provenant de toutes les directions est identique. Pour chaque sous-bande, une seule valeur est donc suffisante pour caractériser toute la scène sonore. Le caractère diffus Ψ peut être estimé de plusieurs manières, la méthode [Ahonen et Pulkki, 2009, Del Galdo *et al.*, 2012] semble selon les auteurs offrir le meilleur compromis entre fiabilité et complexité de calcul. Cette estimation de la valeur Ψ consiste à calculer la norme de la moyenne du vecteur intensité divisée par la moyenne de la norme du vecteur :

$$\Psi = \sqrt{1 - \frac{\|\mathbb{E}[\mathbf{i}]\|}{\mathbb{E}\,\|\mathbf{i}\|}} \tag{3.8}$$

où $\|.\|$ désigne l'opérateur distance euclidienne. Comme pour l'information de direction, un lissage temporel avec la trame précédente peut être réalisé. La figure 3.5 montre le fonctionnement de la méthode DirAC pour le codeur et le décodeur. Dans le codeur, seule la composante W, désignée

sur la figure par b_w, est transmise au décodeur. Pour reconstruire un signal FOA pour chacune des sous-bandes, la direction de la source et la valeur de caractère diffus sont transmis.

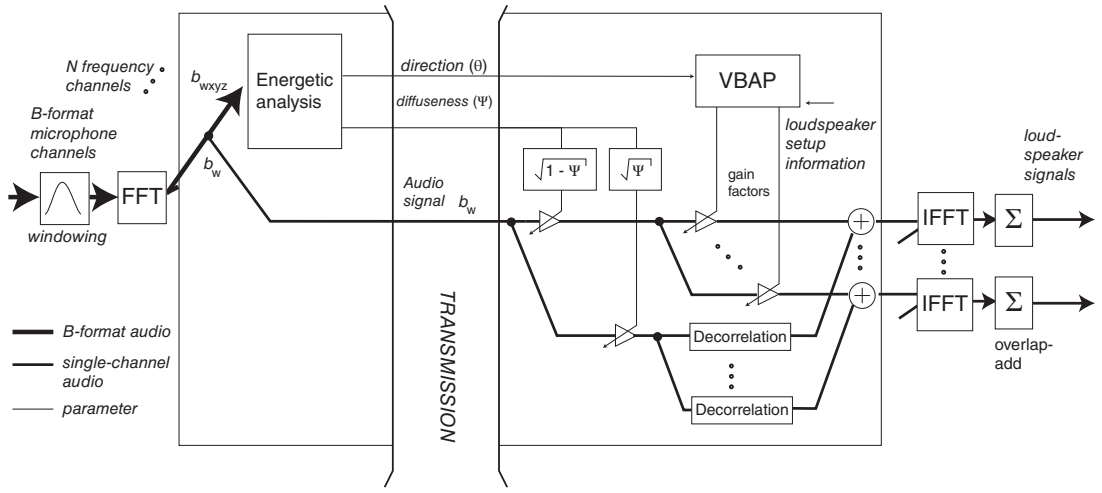


FIGURE 3.5 – Fonctionnement de la méthode DirAC. La figure est issue de [Pulkki *et al.*, 2018].

Dans la partie décodeur, les informations de direction (θ, ϕ) la valeur du caractère diffus de chaque sous-bande sont extraites. Puis ces informations de directions sont utilisées pour spatialiser la composante décodée W'.

La spatialisation est réalisée par la méthode Vector Base Amplitude Panning (VBAP) [Pulkki, 1997]. Cette méthode consiste à spatialiser une source dans une certaine direction par un panoramique d'intensité sur les haut-parleurs adjacents à la direction de la source. Selon le contenu qui est reproduit (FOA, ambisonique planaire ...), la position et le nombre de haut-parleurs peuvent varier. Pour une restitution ambisonique du signal d'entrée, les haut-parleurs sont généralement placés sur une sphère tout autour de l'auditeur.

La largeur de la source est simulée par l'ajout d'une version décorrélée de la composante W' sur l'ensemble des haut-parleurs. Cette largeur de la source est contrôlée par le caractère diffus ainsi que l'énergie de la composante W' sur la sous-bande. Plus le caractère diffus est important plus la quantité de signaux décorrélés est ajoutée aux haut-parleurs.

3.3.2.3 La méthode HO-DirAC

La méthode DirAC, dédiée uniquement au codage FOA, a été par la suite étendue pour les signaux HOA sous le nom de Higher-Order Directional Audio Coding (HO-DirAC) [Politis *et al.*, 2015]. La méthode consiste à découper le signal HOA en différents secteurs angulaires. La figure 3.6 montre le découpage fait pour les différents ordres ambisoniques. Pour un signal HOA d'ordre 2, le nombre de secteurs angulaires est de 4, pour un signal HOA d'ordre 3, le nombre de secteurs angulaires est de 9.

Pour chaque secteur angulaire, la méthode détermine un triplet de composantes, puis applique

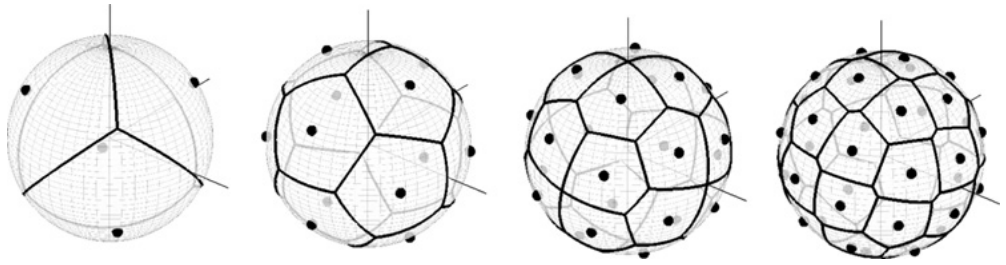


FIGURE 3.6 – Découpage de l'espace sonore selon l'ordre ambisonique. De gauche à droite, découpage fait pour l'ordre 2 à 5. Figure issue de [Politis *et al.*, 2015].

la méthode DirAC sur ces composantes. Pour un secteur, le triplet de 3 composantes (X_{s_i}, Y_{s_i}, Z_{s_i}) est créé par un rematriçage des composantes du signal HOA pour former une base de décomposition de l'espace du point de vue du secteur. Ces composantes jouent le rôle des composantes directionnelles dans la méthode DirAC. Une $4^{\text{ième}}$ composante W_{s_i}, correspondant de la somme des 3 composantes, joue le rôle de la composante omnidirectionnelle. Pour un signal ambisonique d'ordre 2, un exemple des composantes calculées pour un secteur est montré sur la figure 3.7.

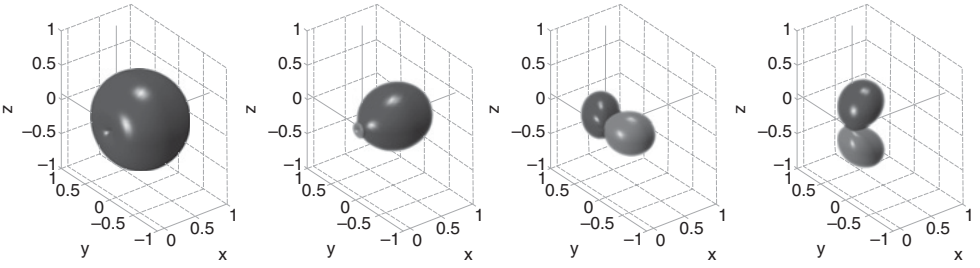


FIGURE 3.7 – Composantes constituant un secteur pour un signal ambisonique d'ordre 2. De gauche à droite, la directivité de la composante W_{s_i}, ainsi que les trois composantes directionnelles X_{s_i}, Y_{s_i}, Z_{s_i}. Figure issue de [Pulkki *et al.*, 2018].

Seule la composante W_{s_i} est codée par un codec cœur et transmise au décodeur. Une analyse de la scène sonore est effectuée pour chaque secteur angulaire pour déterminer la direction de la source principale $(\theta_{s_i}, \phi_{s_i})$ et du caractère diffus Ψ_{s_i} de la scène. Au décodage, ces informations sont utilisées pour permettre de recréer l'image spatiale de chaque secteur angulaire.

3.3.3 Approche avancée

3.3.3.1 Le codec MPEG-H

Le codec MPEG-H [Herre *et al.*, 2015, Bleidt *et al.*, 2017] a été développé pour compresser les signaux immersifs. Ce codec permet de traiter des signaux immersifs avec une représentation

basée : canaux, objets et scène. Pour coder un signal HOA, le codec MPEG-H réalise un *downmix* du signal. La méthode va faire une analyse du signal pour séparer le contenu prédominant et le contenu d'arrière-plan. Les deux types de contenu vont avoir un traitement différent. Pour le contenu prédominant, les sources vont être extraites et transmises comme des canaux séparés. Le nombre de canaux transmis peut varier en fonction de la scène sonore et du débit disponible. Pour le contenu d'arrière-plan, une représentation approximative de la scène sonore va être faite par un signal FOA. Les 4 canaux de cette représentation vont être transmis. Pour déterminer et extraire le contenu prédominant et le contenu d'arrière-plan du signal, deux modes sont proposés : un mode de décomposition du champ sonore en ondes planes, proche du fonctionnement de la méthode HARPEX [Berge et Barrett, 2010], et un mode par décomposition en valeurs singulières, ou Singular Value Decomposition (SVD), des composantes ambisoniques. Pour le mode SVD, la décomposition est réalisée dans le domaine temporel. La figure 3.8 montre un schéma du fonctionnement du codage pour la compression d'un signal HOA par SVD.

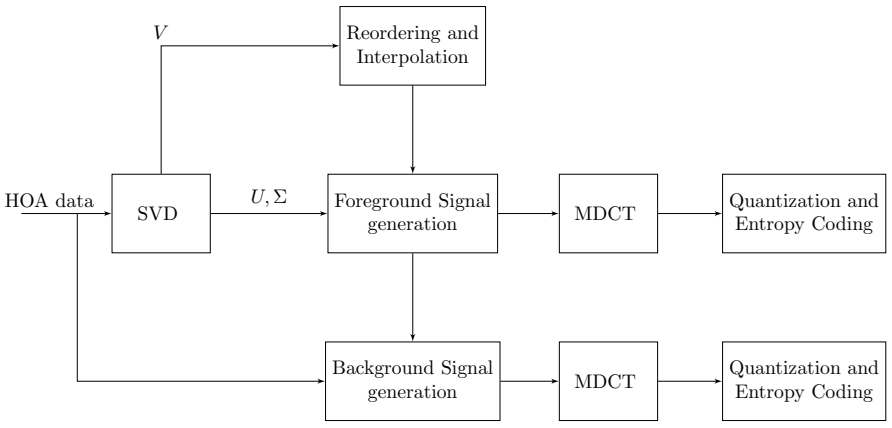


FIGURE 3.8 – Fonctionnement du codec MPEG-H pour la compression du HOA. Figure issue de [Zamani *et al.*, 2017].

Pour un signal HOA d'ordre N, le nombre de composantes est $M = (N+1)^2$. Le signal ambisonique est découpé en trame de longueur $2L$ avec un recoupement de 50 % pour former une matrice de taille $2L \times M$. Une SVD est réalisée sur cette matrice, telle que :

$$\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{3.9}$$

où $\mathbf{B}$ est le signal HOA de taille $2L \times M$, $\mathbf{U}$ et $\mathbf{V}$ les vecteurs singuliers de taille respectivement $2L \times 2L$ et $M \times M$. La matrice diagonale Σ de taille $2L \times M$ correspond aux valeurs singulières de la matrice. Le contenu prédominant correspond aux r plus grandes valeurs singulières Σ. Pour isoler le contenu prédominant, les r premiers vecteurs de la matrice $\mathbf{V}$ sont extraits pour former la matrice $\mathbf{V}_r$. Chacun des r vecteurs de la matrice permet de réaliser un matriçage des composantes, ce qui correspond à une formation de voie dans le domaine ambisonique. Selon la complexité de la scène, chaque formation de voie peut être focalisée dans une direction précise, la direction d'une source, ou pointer dans plusieurs directions, dans le cas de plusieurs sources simultanées. La figure 3.9 montre des exemples de formations produits par un vecteur de la matrice $\mathbf{V}_r$. La figure de gauche correspond à un faisceau obtenu pour l'extraction d'une source ponctuelle. Les

figures du centre et de droite correspondent à des faisceaux qui capturent un signal plus réparti dans l'espace.

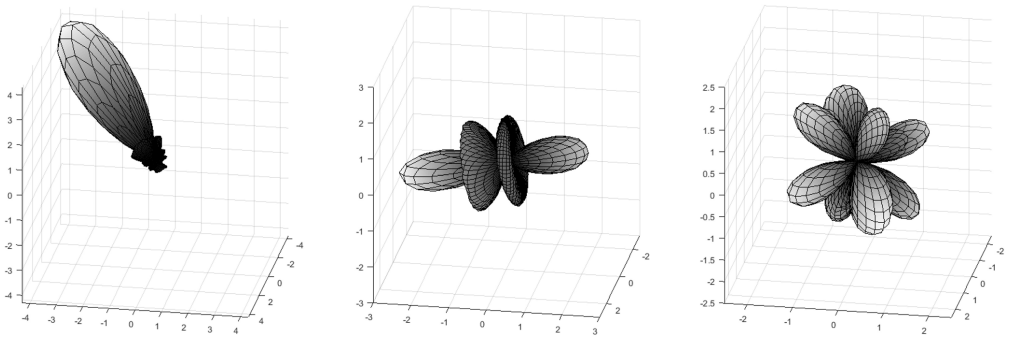


FIGURE 3.9 – Différent type de formation de voie que peut produire un vecteur de $\mathbf{V}_r$. La figure est issue de [Bleidt *et al.*, 2017].

Pour obtenir le contenu d'arrière-plan, la contribution du contenu prédominant est soustraite du signal HOA d'entrée. À partir du signal obtenu, les 4 composantes formant le FOA sont conservées. Les canaux extraits, formés de r canaux pour le contenu prédominant et du signal FOA d'arrière-plan sont ensuite codés individuellement par un codec cœur mono USAC. Pour chaque trame, la matrice $\mathbf{V}_r$ est transmise en tant que métadonnée pour permettre au décodeur de respatialiser du contenu prédominant.

3.3.3.2 Proposition d'amélioration du codeur MPEG-H

Dans le codec *MPEG-H*, le contenu prédominant extrait lors du calcul de la SVD peut varier fortement d'une trame à l'autre. Entre deux trames consécutives, un élément peut être considéré comme faisant partie du contenu prédominant dans la première trame, puis à la trame suivante être considéré comme faisant partie de l'arrière-plan. Ce phénomène peut produire des changements brutaux dans les canaux du contenu prédominant, ce qui peut rendre l'image spatiale instable à l'écoute. De plus, il peut avoir une permutation de l'ordre des vecteurs dans $\mathbf{V}_r$ d'une trame à l'autre, ce qui peut créer des discontinuités dans le signal codé et dégrader fortement la qualité audio.

Dans l'article [Zamani *et al.*, 2017], les auteurs proposent des améliorations pour le codec *MPEG-H*. Dans leur méthode, la SVD est calculée dans le domaine fréquentiel MDCT et non dans le domaine temporel. Ce changement permet de tirer avantage du recouvrement de 50 % lié à la MDCT. Lors de la reconstruction du signal dans le décodage, ce recouvrement effectuera un lissage du signal reconstruit, ce qui permet des transitions moins brutales entre les trames. Pour minimiser le nombre de permutations des vecteurs dans la matrice $\mathbf{V}$, un mécanisme de réorganisation des vecteurs d'une trame à l'autre a été mis en place. Cette réorganisation se base sur l'analyse de l'inter-corrélation des vecteurs $\mathbf{V}_t$ et $\mathbf{V}_{t-1}$.

La figure 3.10 montre le fonctionnement du codec *MPEG-H* amélioré. Le signal HOA est divisé en N bandes de fréquences. Pour chaque bande de fréquences, une SVD est calculée pour déterminer les r canaux pour le contenu prédominant et le signal FOA d'arrière-plan. Une fois

l'ensemble des signaux r extraits des N sous-bandes, les signaux des différentes bandes de fréquences sont concaténés pour former r signaux. La même opération est faite pour les signaux FOA d'arrière-plan. La matrice $\mathbf{V}_r$ associée à chaque bande de fréquences est quantifiée et transmise au décodeur. Pour retrouver le signal HOA, les N matrices $\mathbf{V}_r$ sont transmises au décodeur. L'envoi des N matrices entraine une multiplication par un facteur N du débit nécessaire pour les métadonnées par rapport au codec *MPEG-H* sans les améliorations.

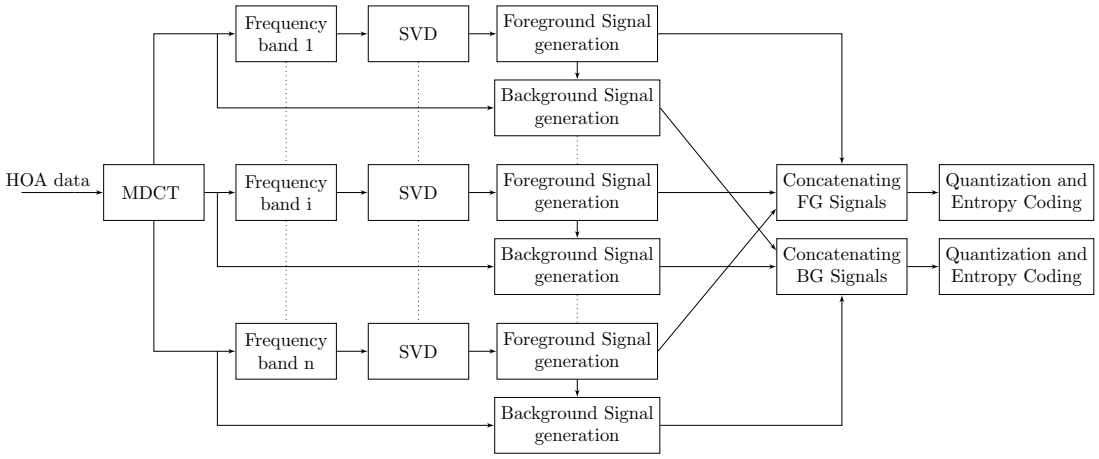


FIGURE 3.10 – Fonctionnement de la proposition d'amélioration du codec MPEG-H pour la compression du HOA. La figure est issue de [Zamani *et al.*, 2017].

Les décompositions SVD et la sélection du contenu prédominant et du contenu d'arrière-plan sont indépendantes entre les bandes de fréquences. Cette indépendance implique que d'une bande à l'autre, la formation de voie peut extraire un contenu totalement différent. Cela signifie que lors de la concaténation des N signaux pour les bandes de fréquences, des discontinuités peuvent apparaître dans les signaux. Malgré cela, l'évaluation la qualité semble montrer que la méthode proposée permet d'obtenir une qualité significativement meilleure que la méthode *MPEG-H* classique [Zamani *et al.*, 2017].

3.3.3.3 Méthode MAEC

À la section 3.3.2.1, nous avons présenté une approche de codage effectuant une prédiction inter-canal pour le codage FOA. Le codec Metadata Assisted EVS Codec (MAEC) propose d'utiliser cette approche pour effectuer un codage paramétrique pour des signaux HOA [McGrath *et al.*, 2019]. L'idée du codec est d'effectuer un *downmix* de HOA vers FOA puis de coder le signal FOA par l'approche par prédiction. La figure 3.11 montre un schéma du fonctionnement du codec. Ce codec a été conçu pour la compression audio immersive dans le cadre de la réalité virtuelle. Le codec peut prendre en entrée un signal HOA d'ordre N accompagné de M objets sonores. Le nombre total de canaux est de $N_{total} = M + (N+1)^2$. Ces sources sonores peuvent se déplacer au cours du temps. Il n'y a pas de contrainte quant à l'ordre et au nombre maximal de signaux théorique de la méthode. Cependant, pour des raisons pratiques, le nombre maximal de canaux a

été fixé à $N_{total} = 64$ soit un signal HOA d'ordre 7.

Dans le codeur, un *downmix* est réalisé pour tronquer le signal HOA d'ordre N vers un signal FOA à l'aide d'une matrice fixe de *downmix* $\mathbf{M}$:

$$\mathbf{B}(t) = \mathbf{M} \times \mathbf{X}(t) \tag{3.10}$$

avec $\mathbf{X}(t)$ le signal d'entrée HOA et $\mathbf{B}(t)$ le signal *downmixé*. Pour un signal d'entrée HOA d'ordre 2 et 2 objets sonores, soit au total $N_{total} = 11$, la matrice $\mathbf{M}$ est définie par :

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{obj}_{x_1}(t) & \text{obj}_{x_2}(t) & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{obj}_{y_1}(t) & \text{obj}_{y_2}(t) & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{obj}_{z_1}(t) & \text{obj}_{z_2}(t) & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.11}$$

Avec $\text{obj}_{x_o}(t)$, $\text{obj}_{y_o}(t)$, $\text{obj}_{z_o}(t)$ les coordonnées cartésiennes des objets sonores. La méthode de prédiction inter-canal, décrite à la section 3.3.2.1, est appliquée aux 4 composantes FOA pour obtenir le signal $\mathbf{B}'(t) = [w(t), x'(t), y'(t), z'(t)]^t$. Puis chaque composante est codée par un codec cœur EVS indépendant.

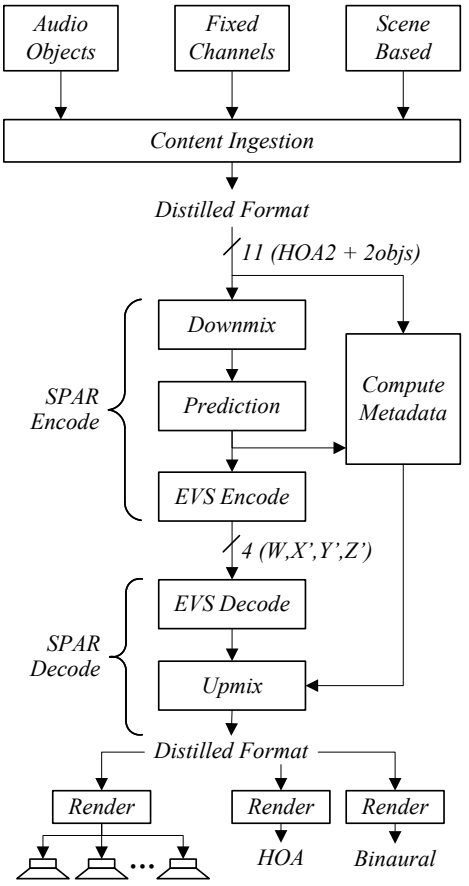


FIGURE 3.11 – Schéma bloc du fonctionnent du codec MAEC.

Pour permettre de faire recréer un signal HOA à partir du signal FOA, le signal est découpé en un ensemble de sous-bandes $b \in \{1, \ldots, B\}$. Le décodeur s'appuie alors sur 2 matrices : la matrice $\mathbf{C}_b$ et la matrice $\mathbf{P}_b$ par sous-bandes. Une matrice $mathbf{C}_b$ est estimée pour être la matrice d'*upmix* optimum au sens des moindres carrés :

$$\mathbf{C}_b = \underset{C}{\operatorname{argmin}} \int_t \left\| \mathbf{C} \times \mathbf{B}'_b(t) - \mathbf{X}_b(t) \right\|_2 \tag{3.12}$$

Si les signaux sont peu corrélés, la solution de la matrice $\mathbf{C}_b$ peut provoquer une perte d'énergie dans certaines zones de l'espace du signal généré par l'*upmix*. Pour remédier à cela, la covariance du signal d'erreur $\mathbf{C}_b \times \mathbf{B}'(t) - \mathbf{X}(t)$ est partiellement compensée par la matrice $\mathbf{P_b}$. Ces matrices sont estimées dans l'encodeur par le module *Compute Metadata* de la figure 3.11. Elles sont ensuite quantifiées et codées par un codage de Huffman puis transmises au décodeur en tant que métadonnées.

Dans le décodeur, le signal FOA transmis est décodé $\mathbf{B}''(t) = [w'(t), x''(t), y''(t), z''(t)]^t$. À partir de la première composante w'', D filtres de décorrélation sont appliqués à la composante pour créer D-canaux décorrélés $\mathbf{D}(t)$. À l'aide des matrices $\mathbf{C}$ et $\mathbf{P}$, les signaux $\mathbf{B}''(t)$ et $\mathbf{D}(t)$ sont mixés pour obtenir le signal de sortie HOA, tel que :

$$\mathbf{X}' = \mathbf{C} \times \mathbf{B}'' + \mathbf{P} \times \mathbf{D} \tag{3.13}$$

Les performances de l'approche, en terme de qualité audio, ont été évaluées par des tests MUSHRA pour diverses configurations : FOA, HOA, écoute sur haut-parleurs ou au casque. De plus les résultats ont été répliqués par plusieurs autres laboratoires [3GPP TS 26.818, 2018]. Cependant, les évaluations se sont intéressées uniquement à la qualité audio absolue, les performances de la méthode n'ont pas été comparées à d'autres méthodes de compression ambisonique.

CHAPITRE 4

Codage spatial FOA par rematriçage des composantes basé sur une PCA

Sommaire du chapitre

Dans ce chapitre, nous présentons une méthode de codage permettant d'étendre la méthode multimono pour améliorer la qualité de la compression des signaux FOA. Nous avons approfondi l'idée du rematriçage des composantes avant la compression par le codec cœur. Les composantes sont transformées de la représentation *format-B* ambisonique vers une autre représentation plus facilement codable. Des rematriçages fixes des composantes ont déjà été proposés dans des méthodes de codage comme le mode ambisonique d'Opus [Skoglund, 2018] ou dans [3GPP TS 26.918, 2018]. Contrairement aux méthodes existantes, notre méthode propose un rematriçage dynamique des composantes au cours du temps. La méthode proposée peut être vue comme une extension de [Briand *et al.*, 2006] où un rematriçage dynamique des canaux stéréo est effectué à l'aide d'une analyse statistique. Dans l'approche [Briand *et al.*, 2006], les auteurs utilisaient une analyse en composantes principales, ou PCA, permettant de créer deux signaux décorrélés à partir des deux canaux stéréo d'origine. Dans le domaine ambisonique, la méthode [Baqué *et al.*, 2016] propose de faire une décomposition de la scène sonore par analyse en composantes indépendantes, ou Independent Component Analysis (ICA), pour extraire un ensemble de contributions acoustiques indépendantes. L'utilisation d'une transformation PCA ou ICA, pour limiter la corrélation des composantes, permet de limiter la quantité de données à transmettre, l'information redondante entre les canaux ayant été retirée. Ce type de traitement permet de supprimer la redondance entre

les canaux et de supprimer la corrélation qu'ils pourrait avoir entre eux. Nos premières observations sur les codages multimono semblent indiquer que les artefacts spatiaux sont dus à la dégradation des relations entre les composantes. La suppression de la corrélation devrait donc permettre de limiter l'apparition d'artefacts spatiaux liée au codage par les codecs cœur mono.

Notre méthode a pour objectif de proposer une amélioration de l'approche multimono en ajoutant un traitement amont et aval du codec cœur sans influer sur le fonctionnement interne du codec cœur, ce qui doit permettre à l'amélioration être utilisée quel que soit le codec cœur utilisé (EVS, Opus...). Le signal d'entrée des codecs cœur est généralement dans le domaine temporel. Pour rester indépendant de la représentation interne du signal d'une part et pour ne pas ajouter un retard additionnel lié a un changement de domaine, nous avons décidé de définir ceci comme une contrainte et de proposer une méthode qui opère également dans le domaine temporel avec un signal sans recouvrement. Cette absence de recouvrement demande une attention particulière pour garantir une continuité du signal audio décodé. Dans le domaine fréquentiel, les auteurs de l'article [Zamani *et al.*, 2017] proposent un mécanisme pour réaligner les canaux audio grâce à une décomposition en valeurs singulières, ou SVD, pour chaque trame. Ce mécanisme se base sur la similarité de la base de décomposition d'une trame sur l'autre pour permuter les canaux rematricés. Cependant, cette méthode demande un recouvrement d'une trame sur l'autre, ce qui demande des traitements supplémentaires pour l'utiliser dans le domaine temporel sans recouvrement. Nous avons adressé le problème de continuité des coefficients de matriçage d'une trame à l'autre, par la mise en place d'une interpolation des coefficients de rematriçage par le biais de matrices de rotation dans le domaine des quaternions. De plus, notre méthode essaie d'être la plus agnostique possible quant au contenu et au nombre de sources de la scène sonore.

4.1 Étude préliminaire des limites de l'approche multimono

La question de la compression ambisonique est un sujet encore récent, le nombre d'études sur l'impact de la compression sur ce type de contenu est encore assez limité. Dans la section 3.3.3.1, nous avons vu que le codec MPEG-H permettait la compression ambisonique d'ordre supérieur pour une gamme de débit compris typiquement entre 512-1200 kbit/s [Herre *et al.*, 2015]. Cette gamme de débit ayant été sélectionnée pour permettre d'obtenir une bonne qualité, proche de la transparence audio, pour des contenus audio-vidéo immersifs telles que le cinéma, les vidéos 360... Dans le domaine de la téléphonie, les contraintes pour la compression audio sont beaucoup plus fortes que cela soit sur le débit de données transmises ou la latence. Cette limitation de débit ne permet pas d'utiliser des stratégies identiques à celles mises en place dans MPEG-H. Pour donner un ordre de grandeur, le dernier codec téléphonique mono standardisé EVS est déployé pour un usage commercial avec un débit allant de 9,6 kbit/s à 24,4 kbit/s pour coder un signal mono, ce qui donnerait comme pour la compression d'un signal FOA de 4 composantes un débit total allant de 38,4 kbit/s à 97,6 kbit/s.

Avant toute chose, il est important d'étudier les performances de méthodes existantes pour le codage FOA pour la téléphonie. Nous avons choisi d'étudier l'approche multimono, cette approche est considérée comme une méthode de référence, notamment car elle est relativement simple à mettre en œuvre.

Cette première analyse nous a permis de déterminer avec plus de précision les limites de la méthode multimono aussi bien en terme de qualité globale que de dégradations et d'artefacts générés. Cette analyse permet également d'affiner la gamme de débit à étudier par la suite et les mécanismes à mettre en œuvre pour améliorer le codage ambisonique.

Pour le codage FOA par l'approche multimono, l'hypothèse est faite que les composantes pouvaient être codées indépendamment les unes des autres. Chaque composante est codée par un codec cœur mono indépendant. Le débit total disponible est réparti sur les 4 composantes FOA. Cette répartition peut être égale ou inégale, avec un débit plus important attribué au canal omnidirectionnel [Brettle et Skoglund, 2016]. Pour les répartitions égales, le débit est généralement présenté avec le formalisme : $4 \times [Débit\_Mono]$, ce formalisme sera également utilisé dans le reste de ce chapitre.

4.1.1 Test subjectif de la méthode multimono

Pour étudier la qualité de l'approche multimono, un test d'écoute selon la méthodologie MUSHRA a été conduit. Ce test a eu pour but d'étudier l'impact de la méthode de codage multimono sur des composantes ambisoniques, et de mieux comprendre les dégradations produites par cette approche. Les quatre conditions évaluées ont été l'approche multimono à différent débits : 65,6 kbit/s ($4 \times 16{,}4$ kbit/s), 97,6 kbit/s ($4 \times 24{,}4$ kbit/s), 192 kbit/s (4×48 kbit/s).

En plus des conditions testées, trois conditions de calibration sont ajoutées : une référence cachée ainsi que 2 ancres. Dans la recommandation de l'ITU-R [ITU-R BS.1116, 2015] détaillant la méthodologie MUSHRA, la dégradation spatiale à apporter aux ancres est assez vague et il n'existe pas de consensus à ce jour. Dans le cadre de la normalisation de services de réalité virtuelle par le 3GPP, une large majorité des études utilisent un signal mono et un signal stéréo filtré comme ancre (basse et moyenne) [3GPP TR 26.918, 2018, Section 6]. Même si cela apporte une dégradation spatiale aux ancres, l'impact réel sur la notation des participants aux tests de telles ancres restent pour l'instant encore discutés. Du côté des recherches académiques, selon les études une grande diversité d'ancres est utilisée (filtrage fréquentiel, réductions spatiales diverses…). En l'absence de consensus, pour ce test préliminaire, nous avons décidé d'utiliser comme ancres le signal de référence filtré à 3,5 kHz et à 7 kHz sans dégradation spatiale.

Dans ce test MUSHRA, 12 extraits audio critiques ont été utilisés. Une partie des extraits sont issus des extraits proposés par Orange, lors la normalisation de MPEG-H [MPEG, 2013], l'autre partie de ces extraits ont été créés pour les besoins de cette thèse. La description détaillée des échantillons utilisés peut être trouvée à l'annexe A.

Pour cette première étude préliminaire, 8 sujets experts ont participé au test d'écoute. Un test de ce type demande généralement un nombre plus important de sujets, la recommandation de l'ITU-R [ITU-R BS.1116, 2015] conseille d'avoir au moins une quinzaine de sujets. Le nombre limité de participants à notre test ne permet pas d'avoir des résultats précis, mais permet tout de même de déterminer de premières tendances.

La figure 4.1 montre la moyenne obtenue pour les différentes conditions, ainsi que les intervalles de confiance à 95 %. Les résultats montrent qu'une augmentation de débit entraine une augmentation de la qualité. Il n'y a pas de saturation de la qualité, c'est-à-dire un débit à partir

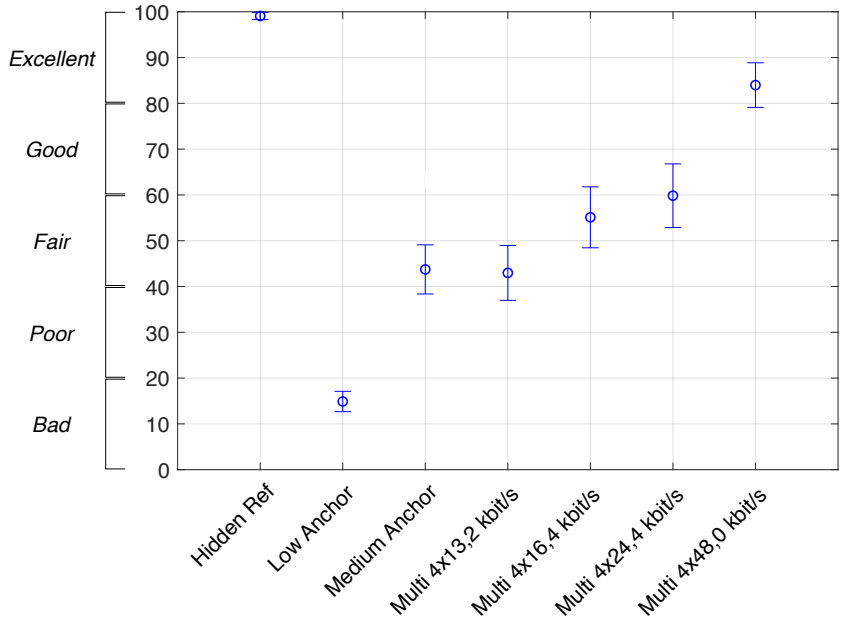


FIGURE 4.1 – Moyennes des scores MUSHRA, avec intervalles de confiance à 95% pour le codage multimono avec le codec cœur EVS.

duquel l'augmentation de débit ne permettrait pas d'améliorer la qualité globale, sur cette gamme de débits. De plus, pour les débits les moins importants, il n'y a pas de chute brutale de qualité en dessous d'un certain débit.

Dans le domaine de la normalisation audio, une condition est considérée comme satisfaisante pour un service téléphonie quand elle obtient un score de qualité supérieur à 80 sur l'échelle MUSHRA. Dans notre test, seule la condition 4×48 kbit/s arrive à atteindre cette qualité. Cela semble indiquer qu'à partir d'un certain débit, l'approche multimono serait suffisante. Pour les débits inférieurs, les résultats du test suggèrent que la méthode multimono ne permet pas d'obtenir une qualité considérée comme satisfaisante. Cela suggère que les améliorations les plus significatives de qualité peuvent être obtenues pour les débits inférieurs et que les futures investigations devraient se concentrer sur ces débits.

La figure 4.2 montre les scores moyens, et les intervalles de confiance à 95 % obtenus pour chaque échantillon en fonction des différentes conditions. Plusieurs observations ressortent de la figure. La première est que plus le débit utilisé pour le codage multimono augmente, plus la qualité subjective augmente. De plus, la grande majorité des échantillons de test obtient un score moyen > 80 avec le codage multimono 192 kbit/s. Seul l'échantillon *Noise* reste fortement dégradé même à haut débit. Pour l'ancre médium, il est possible d'observer deux groupes d'échantillons : un groupe avec des scores autour de 30 et un second groupe avec des scores plus élevés, autour de 60. Cela suggère que le seul filtrage fréquentiel n'est pas une dégradation suffisante pour remplir sa fonction d'ancrage. Une partie des échantillons est fortement impactée par le filtrage fréquentiel quand l'autre partie des échantillons n'est que peu perturbée par la dégradation.

À la suite de ce test, des entretiens ont été menés auprès des sujets. Ils ont permis de faire émerger plusieurs phénomènes. Le principal a été la perception d'artefacts de codage localisés

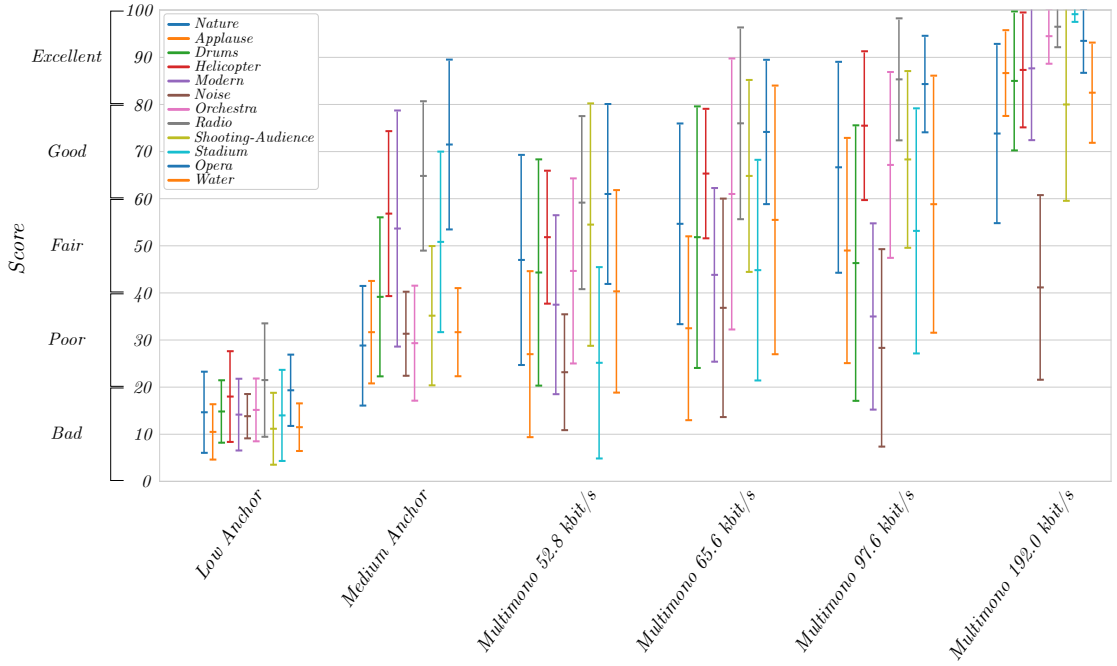


FIGURE 4.2 – Résultats détaillés du test MUSHRA, pour chaque échantillon pour les différentes conditions, avec l'intervalle de confiance à 95 %.

dans la scène spatiale. Ces artefacts spatiaux peuvent être regroupés en trois catégories :

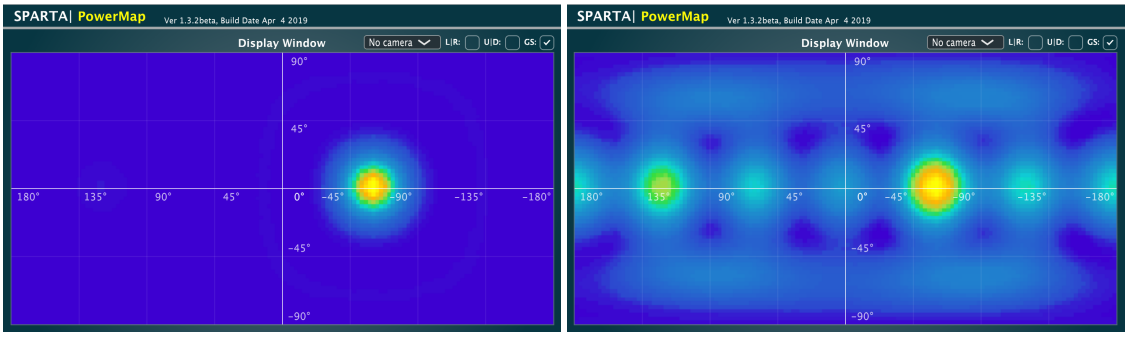
— L'ajout d'un bruit diffus à la scène sonore.

— L'apparition de sources dites "fantômes" correspondant à une copie dégradée de la source originale dans une position symétrique à la position de la source d'origine.

— Et un resserrement de l'espace sonore, la scène spatiale étant poussée vers l'avant de la scène.

Lors des entretiens, il est aussi apparu que ces phénomènes d'artefacts spatiaux sont d'autant plus perceptibles sur les échantillons où la scène sonore est dite simple, c'est-à-dire avec un petit nombre de sources jouées simultanément et avec des positions ponctuelles dans l'espace.

Il est possible de visualiser ces déformations spatiales en observant la cartographie de la puissance d'un signal synthétique contenant une source unique. La figure 4.3 représente deux cartographies de l'espace sonore pour le même signal HOA. L'une des cartographies est celle du signal original, l'autre du signal codé par une approche multimono. Pour rendre plus visible le phénomène, le signal choisi pour la visualisation est un signal ambisonique d'ordre supérieur. Le signal d'illustration HOA d'ordre 3 est un bruit rose faisant le tour de l'auditeur sur le plan horizontal spatialisé de manière synthétique. La révolution complète autour de l'auditeur est faite en 10 secondes. La figure de gauche correspond à une capture de la cartographie du signal HOA d'origine au moment où la source est positionnée à la direction $(\theta, \phi) = (-60°, 0°)$, avec θ l'azimut et ϕ l'élévation. La figure de droite représente le signal HOA d'origine codé avec une approche multimono capturé au même moment. Chaque composante est codée par le codec EVS avec un débit de 16 kbit/s. Sur cette cartographie, il est possible de voir le bruit diffus ajouté dans l'ensemble de la scène, ainsi que les multiples sources fantômes. La plus visible des sources fantômes est dans la



(a) Signal original

(b) Signal codé par l'approche multimono

FIGURE 4.3 – Cartographie d'une source à la position $(\theta, \phi) = (-60°, 0°)$.

direction $(\theta, \phi) = (130°, 0°)$. Ces cartographies ont été réalisées en utilisant le plug-in *PowerMap* du *framework SPARTA* [McCormack et Politis, 2019]. Il est à souligner que cette figure a pour vocation de mettre en évidence le phénomène. Le choix du signal ainsi que des conditions de codage, notamment l'ordre et le débit, ont été faits spécifiquement pour permettre de distinguer facilement les artefacts spatiaux. Dans les conditions de codage FOA utilisées pour le test (ordres et débits), les artefacts sont bien moins visibles.

Notre explication pour la formation de ces artefacts serait qu'ils sont liés à la manière dont chaque composante est traitée par le codec mono. L'approche multimono se base sur l'idée qu'un signal ambisonique peut être interprété comme une prise de son faite par un ensemble de microphones coïncidents. Chacune des composantes ambisoniques est codée comme un signal audio avec un codec cœur mono. Au premier abord cette approche peut sembler pertinente. Cependant, ce type de traitement ne tient pas compte du fait que les composantes sont une décomposition du champ sonore sur une base au sens mathématique. L'utilisation de codec mono indépendant pour chaque composante peut amener à une altération de la corrélation entre les composantes, ce qui correspond à une distorsion de la base ambisonique.

Cette altération de la corrélation pourrait expliquer l'apparition de tous les artefacts spatiaux. Chaque type d'artefact correspondrait à différents types d'altérations des relations entre les composantes. L'ajout de bruits diffus pourrait être lié à l'ajout de bruits de codage non corrélés sur chacune des composantes. L'apparition de sources fantômes correspondrait à la différence entre une composante directionnelle (X, Y, Z...) et la composante omnidirectionnelle (W). Pour chaque composante, une source fantôme serait créée, en fonction de l'énergie sur la composante. Ce qui expliquerait également la répétition périodique du phénomène dans l'espace sur la figure 4.3.

Il existe d'autres variantes de codage multimono. Certaines d'entre elles proposent des améliorations et paramétrages de la compression qui permettent de limiter la quantité d'artefacts spatiaux. Pour obtenir une meilleure qualité globale, la méthode [Brettle et Skoglund, 2016], recommande d'utiliser un débit plus important pour le codage la composante W, par rapport aux autres composantes. D'autres méthodes comme [3GPP TS 26.918, 2018] proposent d'utiliser l'approche multimono non pas sur le format-B mais dans un autre espace apparenté à un format-A idéal. Pour cela, un rematriçage est effectué sur les composantes pour former 4 nouveaux canaux. Puis le codage

est fait sur chaque canal avant de revenir dans le *format-B* par un rematriçage inverse. Les canaux étant une recombinaison des composantes, le bruit de codage, ajouté à chaque canal, se retrouvera réparti entre les composantes. Ce bruit devrait ajouter un bruit plus homogène entre les composantes et limiter la décorrélation des composantes. Cependant, pour ces méthodes seule la qualité globale a été étudiée. Il est donc difficile de déterminer l'impact sur les artefacts spatiaux de telles méthodes.

4.2 Codage spatial FOA par rematriçage par PCA

À la section 4.1, nous avons vu que le codage multimono génére des artefacts spatiaux dus à l'altération de la corrélation entre les composantes.

Dans notre méthode, les composantes ambisoniques sont transformées avant le codage multimono pour supprimer la corrélation des canaux lors du codage multimono. Puis, la transformation inverse est appliquée dans le décodeur pour retrouver des composantes ambisoniques d'origine. Pour cela, notre méthode propose un prétraitement des composantes ambisoniques basé sur une analyse par PCA. Pour chaque trame, une matrice de covariance des composantes est estimée. De cette matrice de covariance est dérivée une matrice de transformation obtenue en utilisant la PCA. La matrice de transformation est appliquée sur les composantes pour obtenir les canaux décorrélés qui seront codés indépendamment par le codec cœur mono. Le signal d'entrée est un signal ambisonique d'ordre 1, le nombre de composantes total est donc $n = 4$. Ces composantes sont numérotées de $i = \{1, \ldots, n\}$. Dans nos expérimentations, c'est le codec EVS qui est utilisé comme codec cœur. La méthode travaille sur des trames de 20 ms. Cette longueur de trame est une taille standard pour les codecs audio téléphoniques conversationnels.

Pour garantir la continuité du signal entre les différentes trames, plusieurs mécanismes de réalignement des valeurs de la matrice de transformation sont mis en place. Chaque trame est découpée en K sous trames, pour chacune de ces sous-trames k une interpolation des coefficients de matrice de transformation est faite entre la trame $t-1$ et la trame t. Pour permettre une interpolation à vitesse constante des coefficients entre les deux matrices de transformations V_t et V_{t-1}, cette interpolation est faite dans le domaine des quaternions. La figure 4.4 montre un schéma complet de l'architecture du codec entier. Dans la suite de cette section, nous présenterons les différents modules qui composent notre méthode. Les traits en gras représentent les flux de données quand les traits fins représentent les flux de métadonnées.

4.2.1 Calcul du rematriçage dynamique par PCA

Pour chaque trame d'indice t, la matrice de covariance $\mathbf{C}$ du signal ambisonique est estimée dans le domaine temporel.

$$\mathbf{C} = \mathbf{B}^t\mathbf{B} \tag{4.1}$$

où $\mathbf{B} = [b_1, \ldots, b_n]$ est la matrice des $n = 4$ composantes ambisoniques. Cette matrice de covariance est décomposée en éléments propres comme suit :

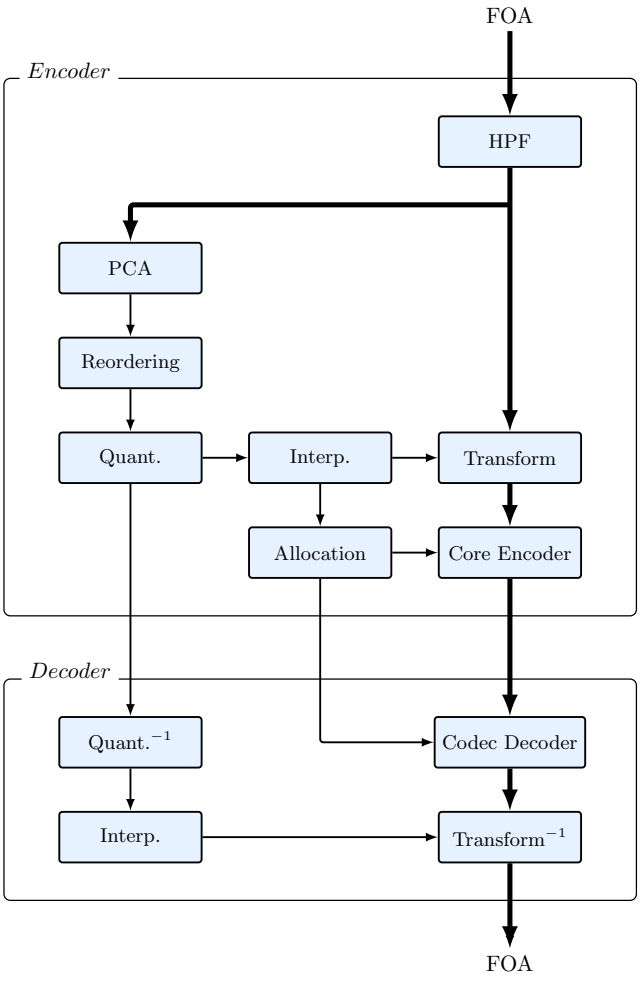
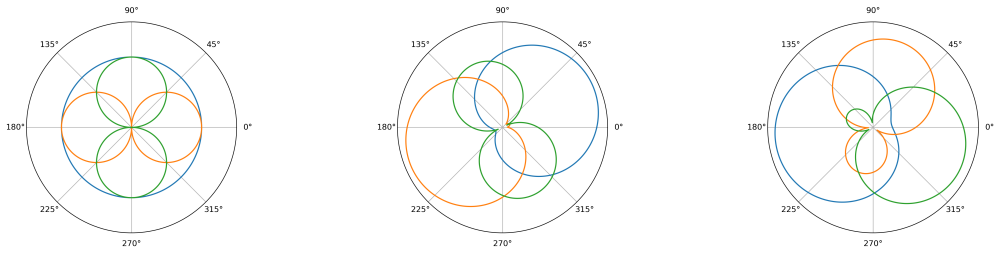


FIGURE 4.4 – Schéma global de la méthode de codage par décorrélation PCA.

$$\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^t \tag{4.2}$$

La matrice $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ correspond à la matrice de transformation qu'il faudrait appliquer aux composantes pour supprimer la corrélation pour la trame t. La matrice $\mathbf{V}$ est orthogonale, cette matrice est soit une matrice de rotation $det(\mathbf{V}) = 1$, soit une matrice de réflexion $det(\mathbf{V}) = -1$. Pour la suite, il est important que la matrice $\mathbf{V}$ soit une matrice de rotation. Pour garantir cela, le déterminant de la matrice est calculé. S'il est négatif $det(\mathbf{V}) = -1$, le signe du dernier vecteur $\mathbf{V_n}$ est inversé. Il est à noter que la matrice de transformation $\mathbf{V}$ peut également être interprétée comme une matrice permettant de créer un ensemble de faisceaux (*beams*) permet de passer du format-B classique vers un domaine spatial équivalent. La représentation sous forme de faisceaux, aussi appelés filtres spatiaux, permet la visualisation des recombinaisons des composantes et le suivi de leurs évolutions au cours du temps et du contenu de l'espace sonore. La figure 4.5 montre 3 motifs de directivité (*directivity pattern*) différents : le motif des composantes de l'Ambisonie planaire (W, X, Y), ainsi que 2 motifs de directivité obtenus par deux matrices de transformations calculées sur deux trames différentes du signal. Il est à noter que, pour chaque figure, les faisceaux sont



(a) Composantes ambisonique planaire

(b) Composantes après rematriçage 1

(c) Composantes après rematriçage 2

FIGURE 4.5 – Exemples de motifs de directivité des composantes ambisoniques planaires originales et après application de matrice de transformations.

orientés pour garder une sensibilité unitaire dans toutes les directions de l'espace. Cette contrainte limite le nombre de configurations possibles pour les faisceaux d'une même recombinaison.

Pour éviter tout biais apporté par les basses fréquences lors de l'estimation de la PCA, les composantes ambisoniques sont traitées par un filtre IIR passe-haut avec une fréquence de coupure à 20 Hz.

$$H_{pre}(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{4.3}$$

Les coefficients (b_i, a_i) du filtre sont repris des caractéristiques du filtre passe-haut IIR utilisé dans le codec EVS décrit dans [3GPP TS 26.445, 2019]. Ce filtre est placé à l'entrée de notre méthode, il est représenté à la figure 4.4 par le module HPF, pour *High-Pass Filter*.

4.2.2 Réalignement des vecteurs propres

D'une trame à l'autre, les vecteurs propres de la matrice $\mathbf{V}$ peuvent changer de manière importante. Ces changements peuvent créer des discontinuités dans le signal et ainsi dégrader la qualité audio finale. Certains changements, même très limités, de la scène sonore ou de l'intensité des sources d'une trame $t-1$ par rapport à la trame t peuvent avoir de gros impacts sur la décomposition en vecteurs propres. Un vecteur pointant dans la même direction peut changer dans l'ordre des vecteurs de la matrice des vecteurs propres entre $\mathbf{V}_{t-1}$ et $\mathbf{V}_t$. De plus, un vecteur avec une certaine direction peut garder la même orientation d'une trame à l'autre mais changer de sens. Cette inversion de sens correspond à une rotation de la phase de $180°$ et donc produit une discontinuité du signal entre les deux trames. Pour corriger cela, notre méthode met en place deux mécanismes permettant une permutation signée des vecteurs propres contenus dans la matrice $\mathbf{V}$.

Permutation des vecteurs propres

Le premier mécanisme va chercher à trouver la meilleure correspondance entre les vecteurs de la trame t et $t-1$. Le sens des vecteurs pouvant changer d'une trame sur l'autre, ce premier mécanisme prend uniquement en compte la direction des vecteurs. Ce problème peut être vu comme un problème d'affectation où le but serait de trouver l'affectation optimale entre les vecteurs de la trame t et de la trame $t-1$. De manière similaire à [Zamani *et al.*, 2017], notre méthode résout ce problème d'affectation en utilisant l'algorithme Kuhn-Munkres, aussi appelé algorithme hongrois, qui permet de trouver une solution optimale. Le critère d'optimisation de l'algorithme est la similarité entre les deux bases, définies comme :

$$J_t = tr(\left|\mathbf{V}_t \mathbf{V}_{t-1}^t\right|) \tag{4.4}$$

où $tr(.)$ correspond à la trace de matrice et $|.|$ correspond à la valeur absolue. Après avoir appliqué la permutation optimale sur la matrice $\mathbf{V_t}$, nous obtenons la nouvelle matrice $\tilde{\mathbf{V}}_\mathbf{t}$.

Inversion du sens des vecteurs propres

L'inversion du sens d'un vecteur propre d'une trame à l'autre est déterminée en utilisant la matrice réalignée $\tilde{\mathbf{V}}_\mathbf{t}$ et la matrice précédente $\mathbf{V_{t-1}}$.

$$\mathbf{\Gamma}_t = \tilde{\mathbf{V}}_t \mathbf{V}_{t-1}^t \tag{4.5}$$

En regardant le signe des valeurs de la diagonale de $\mathbf{\Gamma}_t$, il est possible de déterminer quels vecteurs propres ont changé de sens entre les deux trames. Si la valeur est négative, alors il y a eu une inversion du sens, les valeurs du vecteur propre correspondantes dans la matrice sont alors multipliées par -1. Une fois cette inversion faite, la matrice $\tilde{\mathbf{V}}_\mathbf{t}$ est mise en mémoire pour servir de trame de référence $\mathbf{V}_{t-1}$ pour la trame suivante.

D'autres mécanismes de réalignement sont possibles, notamment en s'intéressant au contenu audio extrait par les faisceaux, dans le but d'apporter une meilleure continuité notamment quand les sources changent de manière importante en position et en intensité. Des mécanismes pourraient également désactiver le réalignement quand il n'est pas nécessaire, lors de changement trop important dans la scène sonore par exemple. L'analyse du contenu et le suivi de sources sont des champs de recherche à part entière fonctionnant avec des mécanismes assez élaborés et coûteux en temps de calcul. Dans notre approche, nous avons voulu rester le plus agnostique possible sur le contenu du signal codé. C'est pour cela que notre méthode utilise des mécanismes uniquement basés sur l'énergie des canaux rematricés. De plus, les deux mécanismes mis en place dans notre méthode permettent une amélioration de la continuité tout en ayant un coût de calcul relativement faible.

4.2.3 Quantification et transmission des matrices de transformation

Une fois réalignée, la matrice ne va pas être appliquée telle quelle sur le signal. La trame t va être divisée en plusieurs sous-trames. Pour chaque sous-trame, une interpolation est faite entre la matrice $\mathbf{V}$ de la trame courante t et de la trame précédente $t-1$. Le calcul des interpolations est

fait de manière locale, l'opération est réalisée du côté codeur et du côté décodeur. Cela permet de ne pas avoir à transmettre les coefficients des matrices pour chaque sous-trame. Seule la matrice de chaque trame $\mathbf{V}_t$ doit être quantifiée et transmise au décodeur.

Pour la quantification, il est possible de transmettre les coefficients de la matrice de transformation par une simple quantification scalaire uniforme des coefficients. Cependant, cela requiert une quantité importante de données. Chaque matrice est composée de 16 coefficients pour une trame de 20 ms, ce qui représente une quantité importante de débit.

La matrice à transmettre étant une matrice de rotation $det(\mathbf{V}) = 1$, il est possible d'utiliser des méthodes moins coûteuses en terme de débit pour la transmettre. Notre méthode s'appuie sur les travaux présentés dans [Briand, 2007], proposant de convertir une matrice de rotation 2D ou 3D avec des angles d'Euler puis de les quantifier. En étendant la même méthode pour des matrices de rotations 4D, il est possible de quantifier la matrice de rotation $\mathbf{V}_t$.

Dans [Hoffman *et al.*, 1972], les auteurs ont montré qu'il était possible de convertir n'importe quelle matrice de rotation $n \times n$ par un ensemble de $n(n-1)/2$ angles d'Euler généralisés. Pour une matrice de rotation 4D, cela permet à la méthode de n'avoir à coder que 6 angles d'Euler généralisés : 3 angles allant de $[-\frac{\pi}{2}, \frac{\pi}{2}[$ et 3 angles allant de $[-\pi,\pi[$. Ces angles sont quantifiés avec une quantification scalaire avec un budget de 8 et 9 bits en fonction de la longueur de l'intervalle ($-\pi$ ou 2π). Ce qui fait un budget total pour les 6 angles de 51 bits par trame.

4.2.4 Interpolation des coefficients des matrices de transformation

Avant de présenter notre stratégie pour l'interpolation des matrices de transformation, il est important de rappeler quelques notions sur les quaternions.

Interpolation de matrices de rotation 3D par quaternion

Les quaternions sont une généralisation des nombres complexes, introduits dans [Hamilton, 1840]. Ils sont utilisés dans différents domaines comme l'animation et la robotique. Un quaternion q est défini comme $q = a + bi + cj + dk$, où a, b, c, d sont des valeurs réelles et $i^2 = j^2 = k^2 = ikj = -1$. Les quaternions sont souvent utilisés pour représenter des rotations 3D, notamment pour l'interpolation de rotations. Une rotation 3D peut être représentée par un quaternion représentant un point sur une sphère unitaire q. Cette rotation admet également une seconde représentation de signe opposé $-q$, correspondant au point antipodal sur la sphère unitaire. À partir de la représentation d'une rotation sous forme d'un quaternion, il devient possible de faire une interpolation entre deux rotations 3D à partir des 2 quaternions :

$$slerp(q_1, q_2, \gamma) = q_1(q_1^{-1}q_2)^{\gamma} \tag{4.6}$$

où q_1 et q_2 respectivement le quaternion de départ et le quaternion d'arrivée et γ le facteur d'interpolation allant de $[0, 1]$. La formule précédente peut être réécrite [Shoemake, 1985] pour obtenir :

$$slerp(q_1, q_2, \gamma) = \frac{sin((1-\gamma)\Omega)}{sin(\Omega)}q_1 + \frac{sin(\gamma\Omega)}{sin(\Omega)}q_2 \tag{4.7}$$

avec $\Omega = arccos(q_1 \cdot q_2)$ est l'angle q_1 et q_2. Cette formule permet d'obtenir une interpolation où les points intermédiaires sont tous sur une sphère unité en dimension 4. De plus, contrairement aux autres interpolations linéaires, la vitesse angulaire est constante et contrôlée par γ.

Chaque matrice de rotation peut être représentée par 2 quaternions de signe différent q et $-q$. Il est important de vérifier que l'interpolation entre q_1 et q_2 emprunte le trajet le plus court sur la sphère unité. Pour cela, la méthode [Shoemake, 1985] propose d'étudier l'angle relatif entre les quaternions interpolés calculés successivement pour déterminer à chaque interpolation s'il vaut mieux utiliser $+q_2$ ou $-q_2$ pour l'interpolation suivante.

Interpolation de matrices de rotation 4D par double quaternion

Dans notre approche, les matrices de transformation sont des matrices de rotation 4D. Pour représenter chacune des matrices dans le domaine des quaternions, il est nécessaire d'utiliser une association de 2 quaternions q et p appelée double quaternion [Hanson, 2006]. Il existe plusieurs méthodes pour obtenir cette paire de quaternions. Pour plus d'informations sur le sujet, nous renvoyons le lecteur à ces travaux [Hanson, 2006, Perez-Gracia et Thomas, 2017]. L'interpolation d'une matrice de rotation 4D est faite en interpolant de manière séparée chaque quaternion constituant la paire de quaternions. Cependant, lors de la décision du plus court chemin, il est nécessaire de prendre en compte les deux angles relatifs des deux quaternions qui forment le double quaternion. L'inversion du signe d'un des quaternions entrainera l'inversion de signe du second quaternion constituant la paire de quaternions.

Interpolation des matrices de transformation

À la section 4.2.3, nous avons vu que la méthode proposée calcule et quantifie pour chaque trame de 20 ms une matrice de transformation $\mathbf{V_t}$. Comme il n'y a pas de recouvrement entre les trames d'analyse, il n'est pas possible d'appliquer directement ces matrices de transformation sur le signal sur les trames sans créer de discontinuité d'une trame à l'autre. Pour permettre d'avoir une transition plus progressive entre les trames et ainsi ne pas avoir de discontinuité dans les canaux, la trame courante est divisée en sous-trames. Pour chaque sous-trame, une interpolation entre la trame courante et la trame précédente est calculée pour obtenir une nouvelle matrice. Cette matrice sera ensuite appliquée à la sous-trame. Les matrices de transformation $\mathbf{V}$ sont converties en paire de quaternions (q, p). Chaque trame est divisée en K sous-trames indexées de $1 \leq k \leq K$ pour la trame courante. Les paires de quaternions (q_{t-1}, p_{t-1}) et (q_t, p_t) sont interpolées en utilisant la méthode d'interpolation linéaire sphérique présentée dans la section précédente. Le facteur d'interpolation γ de l'équation 4.7 est défini par l'indice de la sous-trame $\gamma = k/K$.

Dans notre implémentation, les trames sont de 20 ms, soit $L = 640$ échantillons pour un signal échantillonné à 32 kHz. Selon la capacité de calcul disponible, le nombre sous-trame K peut être plus ou moins important. Un nombre important de sous-trames sera plus coûteux en terme de ressource de calcul mais permet une transition plus fluide entre les trames. Un nombre moins important sera moins coûteux, cependant des discontinuités pourront apparaitre.

Dans notre implémentation, nous avons utilisé $K = 128$, pour permettre d'estimer la qualité audio pouvant être obtenue sans contrainte de complexité. Chaque sous-trame a donc une durée de 10 échantillons soit 0,315 ms.

Les matrices interpolées comme les matrices de transformation sont visualisables sous forme de motifs de directivité. La figure 4.6 montre l'interpolation faite entre deux matrices de transformation selon les facteurs d'interpolation $\gamma = \{0,\ 0{,}25,\ 0{,}5,\ 0{,}75,\ 1\}$. Il est intéressant d'observer la réorganisation des faisceaux pour s'adapter au mieux au contenu. Ainsi le changement de directivité du faisceau jaune passant d'une directivité type *bidirectif* vers une directivité *cardioïde*.

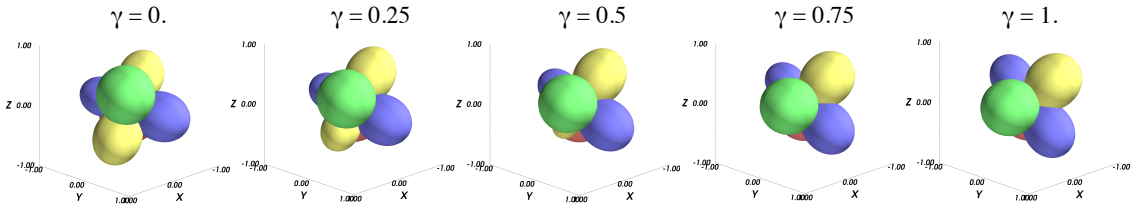


FIGURE 4.6 – Interpolation des faisceaux entre deux trames selon un facteur d'interpolation γ.

La matrice interpolée pour la sous-trame k est appliquée aux composantes FOA d'entrée pour extraire les canaux recombinés qui seront transmis au décodeur.

4.2.5 Allocation adaptative du débit

Une fois, les matrices de transformation calculées pour toutes les sous-trames, les matrices sont appliquées au signal FOA préfiltré, ce qui permet d'obtenir les quatre nouveaux canaux correspondant au rematriçage des quatre composantes ambisoniques.

Ces quatre canaux peuvent être codés avec un débit uniformément reparti entre les codecs cœurs, comme pour le codage multimono. Cependant, certains des canaux sont plus énergétiques que d'autres et véhiculent plus d'information. De manière expérimentale, nous avons constaté qu'une allocation adaptative du débit prenant en compte l'importance des canaux permettait d'obtenir une meilleure qualité. Pour permettre de faire cette répartition adaptative du débit, notre méthode utilise un algorithme glouton.

Cet algorithme cherche à maximiser un score de qualité avec pour contrainte de respecter un débit allouable maximal B_{max}. Ce score prend en compte l'énergie des canaux, ainsi que le gain de qualité obtenu par l'ajout de débit supplémentaire sur ce canal. Le score est défini comme suit :

$$S_t = \sum_{i=1}^{n} Q(b_i) \cdot E_i^{\beta} \tag{4.8}$$

où b_i et E_i correspondent respectivement au débit alloué et à l'énergie du canal i de la trame courante t. $Q(b_i)$ correspond au score de qualité obtenu par le codec cœur pour un débit donné. Le facteur β permet de pondérer l'importance de l'énergie du signal pour l'allocation. Le débit total B alloué pour le codage est défini comme $B = b_1 + \ldots + b_n$. Le débit alloué devant être inférieur au débit allouable maximal $B \leq B_{max}$.

Tableau 4.1 – Score de qualité en fonction du débit pour le codec EVS.

Débit en kbit/s	9,6	13,2	16,4	24,4	32,0	48,0	64,0	98,0	128,0
Score MOS	3,62	3,79	4,25	4,60	4,53	4,82	4,83	4,85	4,87

La qualité audio des codecs téléphoniques actuels, tels que EVS, n'augmente pas de manière linéaire avec l'augmentation du débit [3GPP TS 26.952, 2019]. Il n'est donc pas possible d'utiliser une fonction théorique de débit-distorsion pour estimer la qualité en fonction du débit alloué. Pour approximer le score de qualité $Q(b_i)$ en fonction du débit, notre méthode utilise les scores Mean Opinion Score (MOS) obtenus lors de tests subjectifs. Ces tests subjectifs mesurent la qualité subjective perçue pour chaque débit d'un codec donné. Le tableau 4.1 indique les différents scores de qualité obtenus par le codec EVS. Ces notes MOS sont issues des tests effectués lors de la caractérisation du codec EVS par le 3GPP [3GPP TS 26.952, 2019].

Ce score est dépendant du codec cœur utilisé pour coder les composantes ainsi que du contenu traité (parole, musique...). En cas de changement de codec, les valeurs $Q(b_i)$ doivent être ajustées en fonction des caractéristiques de ce nouveau codec. Pour le codec Opus, une évaluation de la qualité peut être trouvée dans [Rämö et Toukomaa, 2011].

Dans notre méthode, le codec cœur utilisé est EVS. Pour rester compatible avec le codec cœur, les seuls débits possibles pour le codage des canaux lors de l'allocation adaptative sont les débits normalisés par le 3GPP, soit : $9{,}6, 13{,}2, 16{,}4, 24{,}4, 32{,}0, 48{,}0, 64{,}0, 98{,}0, 128{,}0$ kbit/s. Le changement répété de la largeur de la bande de fréquence codée peut dégrader la qualité audio. Le débit minimum allouable par un canal est fixé à $9{,}6$ kbit/s pour garantir le codage du signal sur une bande de fréquence Super-wideband (SWB) soit d'au minimum de 16 kHz.

À l'équation 4.8, le terme de l'énergie E_i^β est élevé à la puissance β, avec $0 \leq \beta \leq 1$, pour permettre un paramétrage plus fin de l'allocation du débit. La valeur de β va jouer sur l'importance du terme d'énergie. Avec une valeur de β proche de 1, les canaux les plus énergétiques auront la majorité du débit. Une valeur proche de 0, au contraire, produira une allocation identique pour l'ensemble des canaux indépendamment de leurs contenus. La répartition de l'allocation sélectionnée pour la trame courante sera codée et transmise au décodeur.

La figure 4.7 montre un exemple d'allocation adaptative avec des trames successives de 20 ms avec $\beta = 0{,}5$. L'échantillon utilisé est l'échantillon *Nature* décrit en annexe A.1. Le débit total est de $97{,}6$ kbit/s l'équivalent de $4 \times 24{,}4$ kbit/s. Sur ce débit total, $94{,}65$ kbit/s sont utilisés pour le codage des canaux et $2{,}95$ kbit/s sont alloués à la métadonnée. Cette métadonnée fera l'objet d'une présentation détaillée dans la section 4.2.6. Sur les $94{,}65$ kbit/s allouables pour les canaux, le débit réellement alloué aux canaux (correspondant à $b_1 + b_2 + b_3 + b_4$) peut-être inférieurs. Cela est dû au nombre limité de débits possible pour le codec EVS. Dans le cas où tout le débit n'est pas alloué, l'ajout de bits de bourrage est ajouté au train binaire, ou *bitstream*.

4.2.6 Structure binaire de chaque trame

La structure binaire des trames est composée de deux parties : une partie contenant les métadonnées et une partie contenant les signaux codés. Les métadonnées sont composées des 6 angles

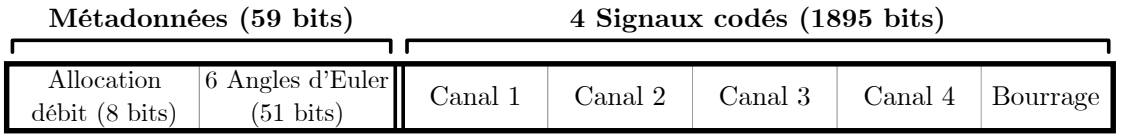


FIGURE 4.7 – Exemple de train binaire pour un signal donné au débit $4 \times 24{,}4$ kbit/s.

d'Euler généralisés, pour coder la matrice de transformation, ainsi que de l'information du débit alloué pour chacun des 4 canaux. Comme présenté dans la section 4.2.3, le codage des angles d'Euler est fait par une quantification scalaire. Les 3 angles, allant de $[-\pi, \pi[$, sont codés avec 8 bits et les 3 angles, allant de $[-2\pi, 2\pi[$ sont codés avec 9 bits. Ce qui fait un total de 51 bits pour le codage des angles d'Euler.

Pour le codage de l'allocation adaptative, le nombre de configurations possibles est restreint, du fait du peu combinaisons de débits possibles pour les codecs cœurs et la contrainte de débit global à respecter. La stratégie mise en place a été de n'envoyer que la valeur correspondant à un indice renvoyant à la table des allocations. Cette table indique l'allocation à effectuer pour les canaux pour la trame.

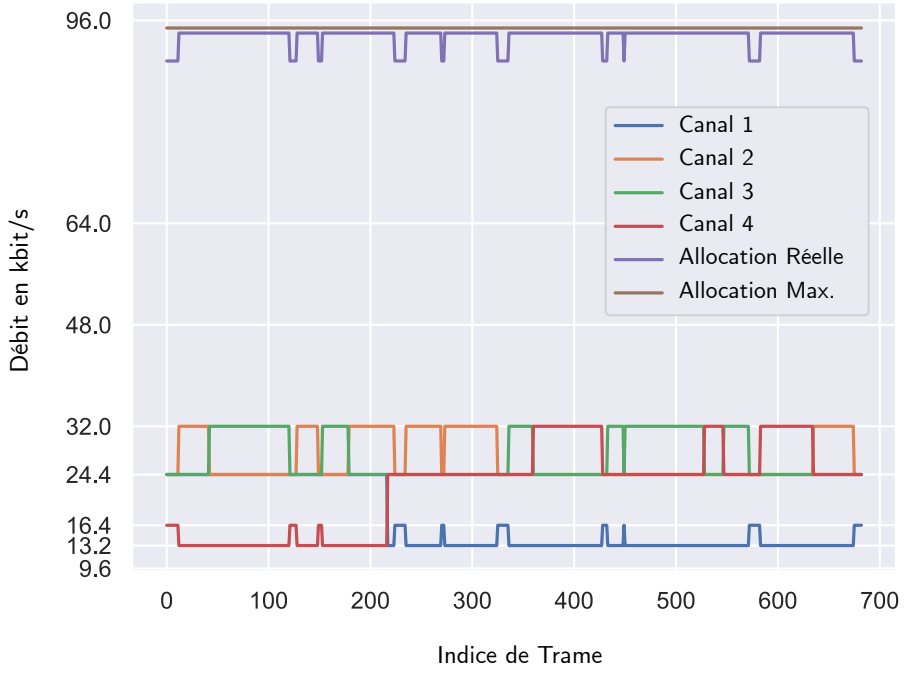


FIGURE 4.8 – Allocation binaire d'une trame pour un signal codé au débit $4 \times 24{,}4$ kbit/s.

Pour le codage d'un signal ambisonique à 97,6 kbit/s, chacun des 4 codecs cœurs peut coder le signal avec 7 débits différents. Il est possible de dresser une liste de toutes les combinaisons possibles soit 2401 possibilités. Sur ces combinaisons, certaines ne permettant pas de respecter la limite totale de débit sont écartées, ce qui réduit le nombre de combinaisons à 659. Cependant, cer-

taines de ces combinaisons sont valides par rapport à la limite de débit totale, mais sous-optimale, car une autre combinaison valide permet un meilleur score de qualité S. Par exemple, la combinaison $[48, 16{,}4, 16{,}4, 9{,}6]$ est valide, car le débit alloué aux composantes est $90{,}40 \leq 94{,}65$, mais sous-optimale par rapport à la combinaison $[48{,}0, 16{,}4, 16{,}4, 13{,}2]$ qui permet de rester sous la limite totale de débit pour les composantes $94{,}00 \leq 94{,}65$, tout en ayant un score de qualité S plus important. La première combinaison ne sera donc jamais sélectionnée par l'algorithme d'allocation adaptatif, il n'est donc pas nécessaire de conserver un indice pour cette combinaison. Les combinaisons sous-optimales sont donc éliminées de l'ensemble des combinaisons possibles. Avec la même logique de maximisation de score de qualité, il est possible, à partir du débit maximum allouable et du débit des 3 premiers canaux, de déduire le débit de la dernière composante. En reprenant l'exemple la combinaison précédente $[48, 16{,}4, 16{,}4, 13{,}2]$, si on indique le débit des 3 premiers canaux $[48, 16{,}4, 16{,}4]$ et que la limite totale de débit est connue, dans notre cas $94{,}65$ kbits/s, il est possible d'en déduire que le $4^{\text{ième}}$ débit est $13{,}2$. Cela permet de conserver uniquement 157 combinaisons, l'indexation de l'ensemble de ces combinaisons est codable sur 8 bits. La figure 4.8 montre la structure d'une trame pour un débit total de $4 \times 24{,}4$ kbit/s. La quantité de métadonnées par une trame de 20 ms est de 59 bits, ce qui représente un débit de $2{,}95$ kbit/s pour les métadonnées.

4.2.7 Décodage du signal

Du côté du décodeur, les signaux transmis sont décodés par le décodeur cœur. En parallèle, les angles d'Euler codés dans les métadonnées sont décodés et la matrice de transformation pour chaque trame est extraite. À partir de la matrice de transformation courante et de la matrice précédente, les matrices de transformation pour les sous-trames sont interpolées. Ces matrices de transformation sont appliquées aux signaux décodés pour obtenir les composantes ambisoniques reconstruites.

4.3 Évaluation de la méthode

Pour évaluer la qualité audio produite par notre méthode, un test subjectif selon la méthodologie MUSHRA [ITU-R BS.1534, 2015] a été réalisé. Ce test a pour but d'évaluer la qualité audio absolue pour notre méthode, mais également de comparer cette qualité à la qualité d'un codage multimono basique à débit équivalent.

4.3.1 Conditions du test

Le test est constitué de 10 échantillons critiques FOA représentant une diversité de situations : 4 scènes voisées, 4 scènes musicales et 2 scènes d'ambiances. La description des échantillons peut être trouvée à l'annexe A.1.

Pour chaque échantillon, les participants doivent noter 9 conditions. La méthode multimono basique pour 3 débits : $4 \times 13{,}2$ kbit/s, $4 \times 16{,}4$ kbit/s, $4 \times 24{,}4$ kbit/s, ainsi que notre méthode aux 3 débits équivalents : $52{,}8$ kbit/s, $65{,}6$ kbit/s, $97{,}6$ kbit/s. En plus de ces conditions, les 3 conditions

Tableau 4.2 – Liste des conditions utilisées lors du test MUSHRA.

Abreviation	**Description**
Hidden Ref	Référence cachée en FOA
Low Anchor	Réference FOA filtrée à 3,5 kHz avec une réduction spatiale ($\alpha = 0{,}65$)
Medium Anchor	Référence FOA filtrée à 7 kHz avec une réduction spatiale ($\alpha = 0{,}8$)
Multi $4 \times 13{,}2$	FOA codé par l'appoche multimono EVS à $4 \times 13{,}2$ kbit/s
Multi $4 \times 16{,}4$	FOA codé par l'appoche multimono EVS à $4 \times 16{,}4$ kbit/s
Multi $4 \times 24{,}4$	FOA codé par l'appoche multimono EVS à $4 \times 24{,}4$ kbit/s
PCA 52	FOA codé par la méthode proposée à $52{,}8$ kbit/s
PCA 65	FOA codé par la méthode proposée à $65{,}6$ kbit/s
PCA 97	FOA codé par la méthode proposée à $97{,}6$ kbit/s

de calibration MUSHRA sont présentes. Toutes les conditions du test ont été récapitulées dans le tableau 4.2. Le débit indiqué prend en compte le codage des canaux ainsi que de la métadonnée. Pour notre méthode, le débit est distribué dynamiquement entre les canaux, comme présenté à la section 4.2.5.

Les écoutes ont été réalisées sur casque dans une salle traitée acoustiquement. Le casque utilisé est le *Sennheiser HD 650* et la carte son *Focusrite Scarlett 6i6*. Le matériel a été le même pour tous les participants. Les échantillons ont été binauralisés par le même moteur de rendu : Resonance Audio Renderer [Gorzel *et al.*, 2019]. Le codec cœur utilisé est le codec EVS (v15.0.0) dans son implémentation à virgule fixe (*fixed-point*). Au total, 11 sujets ont participé au test. Tous les participants sont des professionnels du domaine audio ou des personnes familières avec les tests audio. Aucun participant n'avait de perte auditive connue.

Création des conditions d'ancrages du test

Lors du test sur la qualité multimono décrit à la section 4.1.1, nous nous sommes rendus compte que l'utilisation d'ancres avec uniquement des dégradations fréquentielles ne permettait pas de faire entendre aux participants ce qu'est un contenu fortement dégradé spatialement. Selon la familiarité du participant au son spatial, l'absence d'ancre dégradée spatialement a pu faire sous-estimer ou sur-estimer l'impact des dégradations spatiales dans la notation du test. La recommandation ITU-R BS.1534 [ITU-R BS.1534, 2015] propose pour l'élaboration des ancres lors de test stéréo, d'appliquer une réduction de l'image stéréo en plus du filtrage classique à 3,5 kHz et 7 kHz. Pour les autres signaux multicanaux, la recommandation ne donne ni de directives pour les altérations spatiales à appliquer aux ancres, ni la manière de les produire.

Dans ce test, nous avons décidé d'appliquer une altération inspirée de celle présentée pour les signaux stéréo en effectuant une réduction de l'image spatiale, selon la formule :

$$FOA = \begin{pmatrix} W \\ \alpha X \\ \alpha Y \\ \alpha Z \end{pmatrix}, \quad \alpha \in [0, 1] \tag{4.9}$$

La valeur de α déterminant l'importance de réduction de l'image spatiale, plus la valeur de α est basse, plus la contribution des composantes directionnelles devient faible et donc plus la spatialisation est réduite. Avec une valeur $\alpha = 0$, seule la composante omnidirectionnelle W est présente, la spatialisation du signal a entièrement disparu.

Pour nos tests, la valeur de α a été fixée à $\alpha = 0{,}65$ et $\alpha = 0{,}8$ pour l'ancre basse et l'ancre moyenne. Comme pour le filtrage fréquentiel, la réduction spatiale a un impact plus important sur certains types de contenu, notamment les contenus avec un fort champ diffus. Une série d'écoutes informelles ont été menées pour déterminer les valeurs suffisantes permettant une altération spatiale des contenus sans complètement détruire la spatialisation de la scène sonore.

4.3.2 Résultats du test

Les résultats du test sont présentés sur la figure 4.9. Cette figure inclut, pour chaque condition, la moyenne des scores obtenus, ainsi que l'intervalle de confiance à 95 %. Les résultats montrent qu'à débit équivalent, une amélioration significative de la qualité globale entre notre méthode et la méthode multimono est présente. De plus, la méthode proposée à 52,8 kbit/s permet d'obtenir une qualité équivalente à un débit de 65,6 kbit/s avec la méthode multimono.

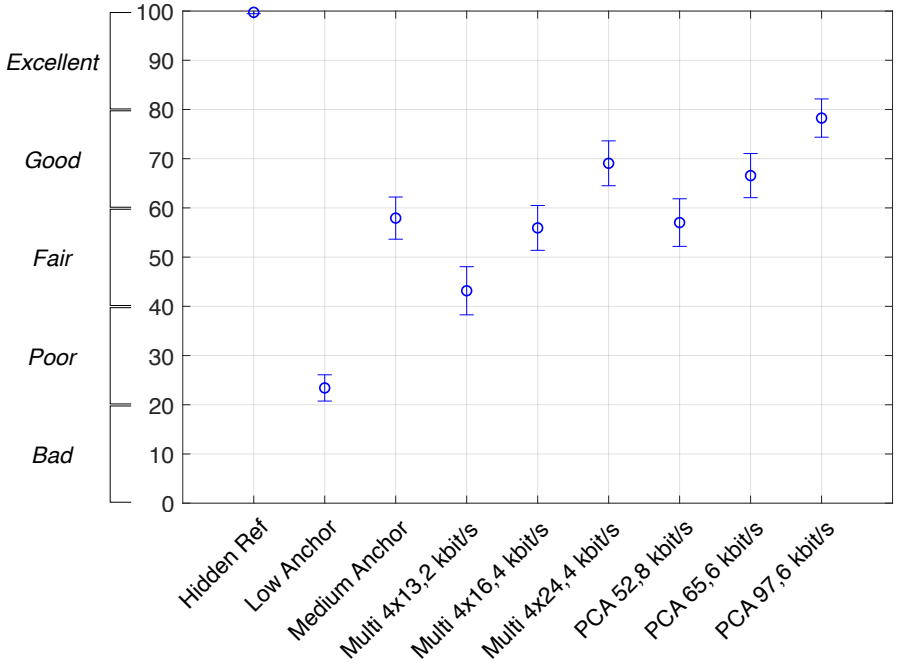


FIGURE 4.9 – Score MUSHRA pour les différentes conditions avec l'intervalle de confiance à 95%.

Deux phénomènes permettraient d'expliquer cette amélioration de la qualité. Ces deux phénomènes ne sont pas exclusifs et pourraient chacun jouer un rôle dans l'amélioration de la qualité perçue. Le premier serait la suppression des artefacts spatiaux par notre méthode. La corrélation entre les canaux codés par les codecs cœurs étant maintenant limitée, la quantité et l'intensité des artefacts spatiaux s'en trouvent grandement diminuée. Cela est d'autant plus audible sur les échantillons ayant peu de sources concurrentes. En écoutant les canaux après rematriçage des composantes, il est possible de constater que la matrice de transformation a tendance à extraire la (ou les) source(s) sonore(s) principale(s) pour la coder dans un unique canal. Ces sources étant contenues dans un seul canal, elles seront donc respatialisées sans l'ajout d'artefacts au décodage. La figure 4.10 montre l'ensemble des scores moyens pour chaque échantillon. L'échantillon *Voices-Kids*, contenant 2 personnes discutant au milieu du brouhaha d'une cour d'école obtient un score bien plus élevé avec notre méthode qu'avec une approche multimono basique. À l'écoute, des artefacts spatiaux présents dans les versions codées par le multimono ne sont plus perceptibles pour les conditions codées par notre méthode, ce qui suggère que rematricer les composantres pour coder séparément les sources principales permet de limiter la présence d'artefacts spatiaux.

Le second phénomène qui pourrait expliquer l'amélioration de la qualité serait l'allocation adaptative des canaux. En réduisant le débit des canaux contenant le moins d'éléments signifiants, et en allouant plus de débit aux éléments les plus importants de la scène sonore. Cette augmentation de la fidélité des éléments principaux de la scène pourrait augmenter la qualité perçue par les participants.

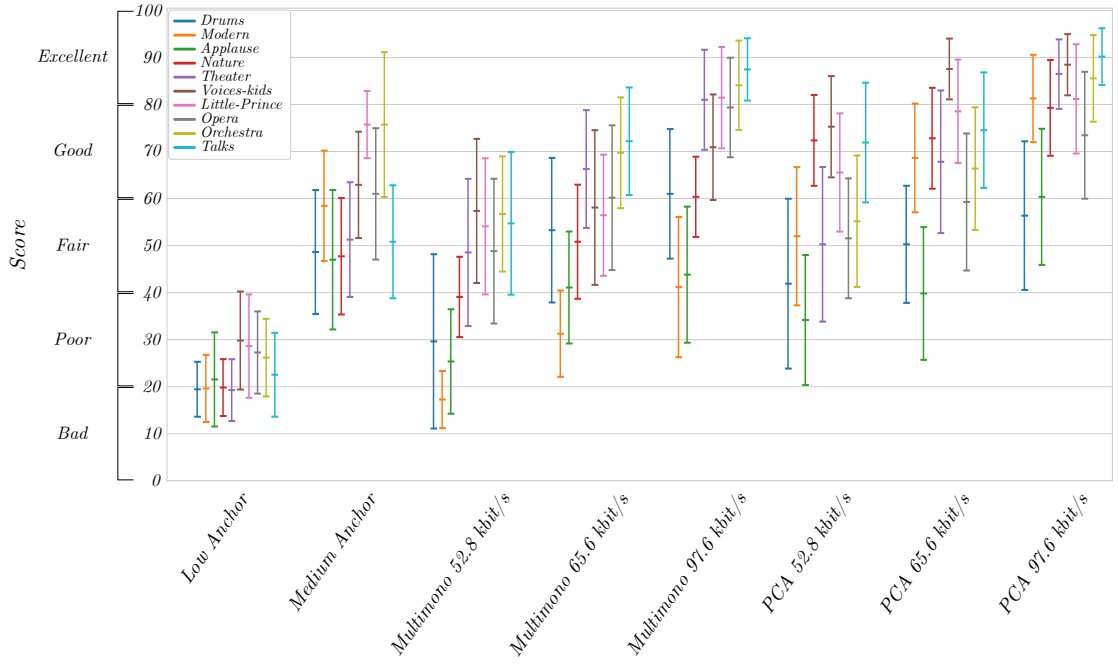


FIGURE 4.10 – Score MUSHRA pour chaque échantillon pour les différentes conditions avec l'intervalle de confiance à 95%.

En regardant en détail la figure 4.10, pour certains échantillons comme *Applause* ou *Drums*, les scores obtenus par notre méthode ne sont pas meilleurs que les scores du multimono. Ces deux scènes sonores sont des scènes considérées comme complexes, c'est-à-dire avec un grand nombre

de sources simultanées.

Le nombre de sources jouées sur la même trame et leurs positions dans l'espace très différentes, la recombinaison des composantes, ne permettent pas d'obtenir des signaux significativement moins corrélés que les composantes d'origines.

De plus, une partie du débit total est utilisée pour transmettre les métadonnées, ce qui laisse un débit moins important pour le codage des canaux audio et crée une dégradation plus importante des canaux audio par rapport à la méthode multimono. L'amélioration de qualité audio apportée par une décorrélation partielle n'arrive pas à contrebalancer la perte de qualité liée au débit moins important alloué aux canaux audio. Ceci expliquerait cette perte de qualité audio pour ces échantillons. Dans une moindre mesure, le phénomène est également visible pour des échantillons avec un champ diffus important. Cela peut être observable pour l'échantillon *Opera*, échantillon correspondant à un enregistrement d'une chanteuse lyrique dans une salle de concert avec une forte réverbération.

4.3.3 Analyse de la méthode par l'étude du vecteur intensité

À l'écoute des échantillons de test, notre méthode limite l'apparition des artefacts spatiaux et les altérations spatiales. Il serait intéressant de quantifier cette amélioration spatiale apportée par notre traitement dans un but de mesurer les performances et d'optimiser les paramètres de la méthode. Les métriques objectives conventionnelles, telles que : le SNR, PEAQ... ne permettent pas de faire une distinction entre le bruit de codage spatial par rapport aux bruits de codage temps-fréquences classiques. Pourtant, pour des signaux simples, il est assez aisé pour des auditeurs de déceler leur présence. Dans cette section, nous proposons, une première approche pour visualiser l'amélioration apportée par notre méthode en terme d'artefacts spatiaux.

Dans [Rudzki *et al.*, 2019], les auteurs étudient l'altération subjective de la spatialisation apportée par le codage des signaux ambisoniques. Dans cette étude, l'hypothèse est que la qualité de la spatialisation et la capacité de localisation sont équivalentes. Lors de tests subjectifs, les participants devaient localiser la position de la source principale dans l'échantillon audio. Cette même tâche a été réalisée pour différents débits évalués. Des résultats de ces tests, les auteurs estiment l'impact du débit sur les erreurs de localisation. Même si cette hypothèse est discutable pour des scènes dites complexes (non prise en compte des sources secondaires, de l'enveloppement de la scène sonore...), pour des scènes simples, l'hypothèse peut être considérée comme valide.

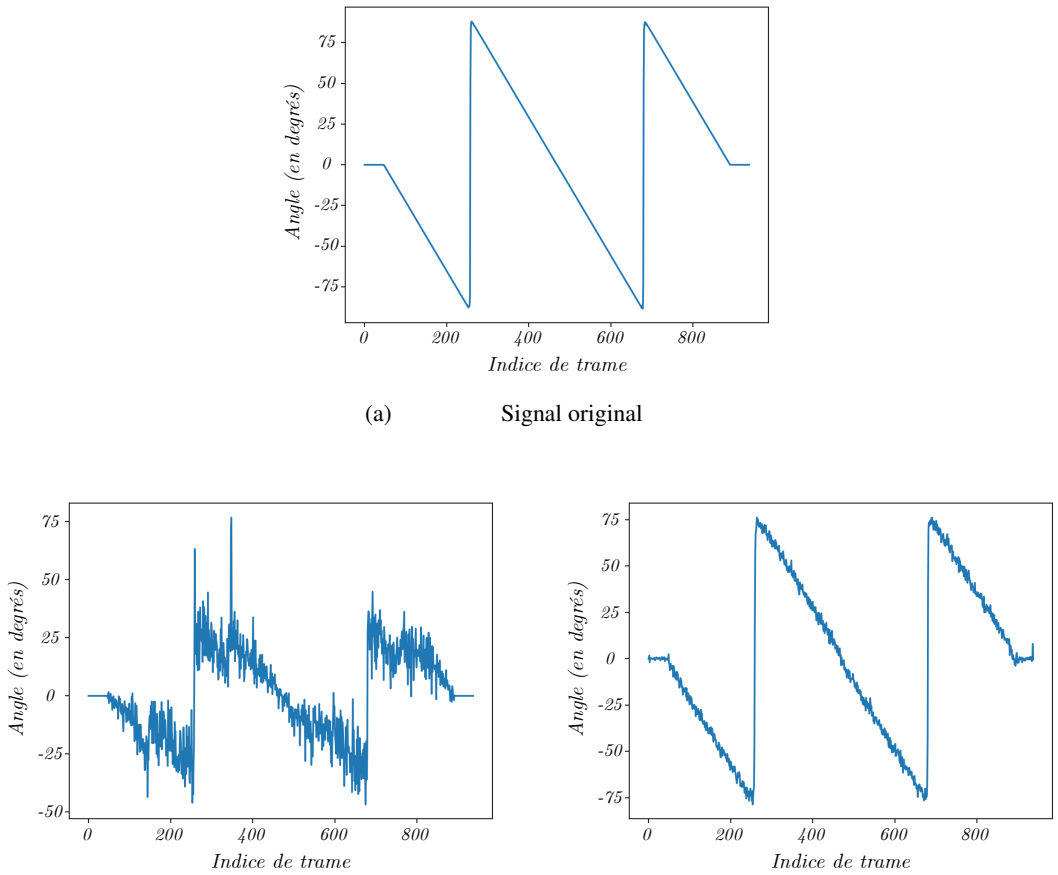
Nous nous sommes inspirés de la même idée pour faire une visualisation des dégradations générées par les artefacts spatiaux. L'estimation de la localisation de la source principale est remplacée par le calcul de la DOA de la scène sonore. En comparant l'estimation de la direction de la source principale du signal original et l'estimation de la direction du signal codé, il est possible de visualiser et mesurer la présence de dégradation spatiale. Pour des signaux simples, plus l'estimation de la direction par la DOA du signal codé diffère de l'estimation du signal original, plus la présence d'artefact spatial est important.

Pour déterminer la localisation de la source, une méthode simple de DOA consiste à calculer le vecteur intensité $\mathbf{I}(f,t)$ pour chaque trame t. Pour estimer la direction uniquement de la source principale, la moyenne du vecteur $\mathbf{I}(f,t)$ est réalisée selon l'axe des fréquences f pour chaque

trame. La direction de la source principale (θ, ϕ) peut être déduite, selon la formule :

$$DOA_t(\theta, \phi) = \angle E_f[-\mathbf{I}(t,f)] \tag{4.10}$$

où $\angle$ donnent l'azimut et l'élévation (θ, ϕ) du vecteur et E_f correspond à l'espérance mathématique selon l'axe des fréquences. Pour augmenter la stabilité de l'estimation dans le temps, un lissage temporel de la valeur $\mathbf{I}$ est effectué. Ce lissage correspond à la valeur moyenne de $\mathbf{I}$ calculé sur les 3 dernières trames.



(a) Signal original

(b) Signal codé avec la méthode multimono

(c) Signal codé avec notre méthode

FIGURE 4.11 – Estimation de la position de la source θ par le Vecteur Intensité en fonction du temps.

Le résultat de la localisation produite par le calcul de la DOA peut être tracé pour les différentes méthodes de codage. Pour notre exemple, nous avons utilisé un signal ambisonique simple, un bruit rose faisant une rotation autour de l'auditeur dans le plan horizontal en 10 secondes. La rotation du bruit rose est faite uniquement dans le plan horizontal, seul l'angle θ change de valeur. La figure 4.11 montre l'estimation de l'azimut θ de la source en fonction du temps. Deux méthodes de codage sont comparées : la méthode multimono et notre méthode par PCA. La figure 4.11(a) correspond à la trajectoire estimée pour le signal original. La trajectoire de la source est clairement

visible, la source démarre face à l'auditeur puis fait une rotation autour de lui pour revenir à son emplacement d'origine en face de l'auditeur. L'estimation étant lissée temporellement, la valeur extrême de $-75 \leq \theta \leq 75$ au lieu des $90°$ attendu. L'estimation de θ pour la méthode multimono, figure 4.11(b), est beaucoup plus dégradée. Avec une estimation très fortement bruitée par rapport au signal original. La difficulté à estimer la position de la source peut être expliquée par la présence de bruit de codage spatialisé qui perturbe l'estimation. Pour la méthode par décorrélation, figure 4.11(c), l'estimation de l'angle θ est bien plus proche du signal d'origine, même si un léger bruit est ajouté à l'estimation.

La visualisation de cette estimation semble être un outil intéressant pour l'étude objective de la présence d'artefacts. L'estimation de la position de la source par le calcul de la DOA est très sensible à la présence de sources perturbatrices. La présence de ces dernières, même à un faible niveau, dégrade fortement la localisation de la source principale. Cette forte sensibilité fait de cette mesure un indicateur intéressant pour l'estimation la quantité des artefacts spatiaux. Malheureusement, l'estimation de la position de la source à l'aide de la DOA ne peut être appliquée que pour des scènes simples avec une unique source active simultanément. Pour les signaux complexes, un raffinement de la méthode serait nécessaire. D'autres méthodes d'estimation de la DOA pourrait être mises en place [Adavanne *et al.*, 2018].

Même si au regard de la figure 4.11, la spatialisation du codage multimono semble complètement altérée, un auditeur peut lui localiser et suivre la trajectoire de la source principale sans difficulté. Le système auditif peut facilement distinguer la contribution de la source principale par rapport aux artefacts spatiaux. De plus, la source principale a une trajectoire plus cohérente et une grande constance dans le temps, alors que les artefacts spatiaux sont plus intermittents et avec des trajectoires plus erratiques. Ce phénomène aide l'auditeur à localiser la source principale par rapport au bruit et permet d'avoir une faible erreur de localisation malgré de nombreux artefacts spatiaux.

4.4 Résumé et perspectives

Dans ce chapitre, nous avons présenté une extension basée sur la méthode multimono pour le codage des signaux FOA. L'idée de notre extension consiste à faire un rematriçage des composantes pour supprimer la corrélation des signaux avant de les coder par les codecs cœur mono. Cette décorrélation permet de limiter les dégradations spatiales liées à la déformation de la base ambisonique par les codecs cœur mono. Nous avons voulu créer une extension indépendante du codec cœur utilisé dans la méthode multimono. Pour cela, la méthode travaille sur les signaux FOA dans le domaine temporel, de plus, il n'y a pas de recouvrement entre les trames audio. Pour effectuer cette décorrélation, la méthode s'appuie sur la décomposition en éléments propres pour calculer une matrice de transformation pour chaque trame. Cette matrice de transformation peut être interprétée comme une formation de faisceaux, il est donc possible de visualiser la manière dont la méthode va recombiner les composantes. Parce qu'il n'y a pas de recouvrement entre les trames, deux mécanismes ont dû être mis en place pour ne pas créer de discontinuité des signaux entre les trames et garantir une transition fluide. Ces deux mécanismes se basent sur la comparaison de la direction et du sens des vecteurs de la trame courante et précédente. Le but de ces mécanismes est de trouver un affectation optimale pour faire correspondre les vecteurs des deux

trames. Ces mécanismes permettent de supprimer les discontinuités liées aux permutations des vecteurs. Malgré cela, des changements brusques de la scène sonore peuvent faire varier fortement les vecteurs et créer des discontinuités dans les signaux. Pour pallier ce problème, notre méthode propose une interpolation des coefficients de la matrice de vecteurs propres d'une trame à l'autre pour garantir une transition progressive entre le contenu des canaux. Pour garantir une interpolation constante des coefficients de la matrice de transformation entre la trame précédente et la trame courante, cette interpolation est faite dans le domaine des doubles quaternions. Nous avons également proposé un mécanisme d'allocation dynamique du débit entre les composantes rematricées dans le but d'allouer plus de débit aux signaux les plus énergétiques.

Par la suite, nous avons présenté un test subjectif conduit pour évaluer les performances de notre méthode par rapport à la méthode multimono de base. Les résultats ont montré qu'à débit équivalent, notre méthode avait une qualité significativement meilleure que l'approche multimono classique. La raison qui semble expliquer cette amélioration de qualité serait la suppression des artefacts spatiaux. Enfin, nous avons présenté une exploration sur la manière de visualiser et mesurer objectivement l'amélioration de la spatialisation à l'aide du vecteur intensité. Ces travaux ont donné lieu à deux publications en conférence [Mahé *et al.*, 2019b,Mahé *et al.*, 2019a] ainsi qu'une dépôt de brevet [Ragot et Mahé, 2019].

CHAPITRE 5

Codage spatial ambisonique par correction de l'image spatiale

Sommaire du chapitre

Au chapitre 4, nous avons étudié une méthode pour limiter les artefacts spatiaux par recombinaison dynamique des composantes ambisoniques. Dans cette méthode, nous avons proposé une amélioration de l'approche multimono tout en conservant intact le codec cœur existant. Pour y arriver, notre méthode a modifié le format des signaux d'entrée de l'approche multimono en passant des composantes à 4 canaux audio équivalents de manière adaptative.

Dans ce chapitre, nous avons voulu explorer une approche qui utiliserait la méthode multimono sans modifier le codec cœur, ni le contenu d'entrée du codage multimono. Pour cela, notre méthode ajoute un post-traitement après le codage de la méthode multimono pour corriger les défauts de la spatialisation apportés par le codage.

L'idée principale de notre méthode est de calculer la répartition spatiale de l'énergie du signal d'origine pour pouvoir restaurer l'image spatiale des signaux décodés. Pour cela, la méthode calcule l'image spatiale d'origine et l'image spatiale obtenue après décodage du signal ambisonique. De ces images, une matrice de correction est calculée. Cette matrice a pour but d'appliquer une correction aux signaux ambisoniques décodés pour recréer, le plus fidèlement possible, la spatialisation d'origine. La matrice de correction est calculée dans le codeur et transmise en tant que métadonnée au décodeur. Dans l'approche, la plus grand partie du coût de traitement est effectué

dans le codeur qui procède à un décodage local des signaux codés en multimono par le codec cœur et calcule la matrice de correction. Le décodeur se contente d'appliquer la correction sur les signaux décodés.

Cette méthode étend le codage multimono sans en modifier le fonctionnement, cela permet une compatibilité avec un codage multimono classique, tout en proposant une amélioration de la spatialisation pour les décodeurs intégrant l'extension.

5.1 Présentation de la méthode

5.1.1 Calcul de la cartographie

Dans le domaine ambisonique, il existe de nombreuses méthodes qui étudient comment visualiser la scène sonore. Une méthode simple consiste à calculer une cartographie de la puissance provenant d'un ensemble de directions de l'espace sonore. Cette cartographie est appelée SRP [McCormack *et al.*, 2019, Jarrett *et al.*, 2010]. Il existe d'autres méthodes pour faire des cartographies de l'espace, comme les méthodes utilisant des filtrages spatiaux de type MVDR [Rafaely, 2019], ou des méthodes comme CroPaC [Delikaris-Manias et Pulkki, 2013]. Dans notre méthode, nous nous concentrerons sur le calcul de carte d'énergie par la méthode SRP dans son implémentation la plus basique.

Dans la méthode SRP, le signal provenant d'un ensemble de directions de l'espace est extrait par un filtrage spatial, aussi appelé formation de voie (*beamforming*). Pour une trame de taille L, la puissance du signal provenant de chaque direction va être calculée comme :

$$s_i(l) = \sum_{n=1}^{N} \mathbf{d}_i^T(n)\mathbf{b}_n(l) \tag{5.1}$$

où $s_i(l)$ est le signal extrait par le $i^{\text{ième}}$ faisceau, $\mathbf{d}_i$ est le vecteur de poids codant pour la direction (θ_i, ϕ_i) du faisceau et $\mathbf{b}_n(l)$ la $n^{\text{ième}}$ composante ambisonique. Puis, l'énergie de chaque signal extrait est calculée comme :

$$\mathbf{P}_i = \frac{1}{L}\sum_{t=1}^{L} s_i(l)^2 \tag{5.2}$$

où l est l'indice de l'échantillon allant de $[1, L]$. La cartographie de la puissance peut être visualisée en traçant la valeur de $\mathbf{P}_i$ en fonction de (θ, ϕ). Dans la plupart des cas, pour être visualisée, la cartographie de la puissance est projetée dans une représentation équirectangulaire. Un exemple de cette projection a déjà été présenté dans la figure 4.3.

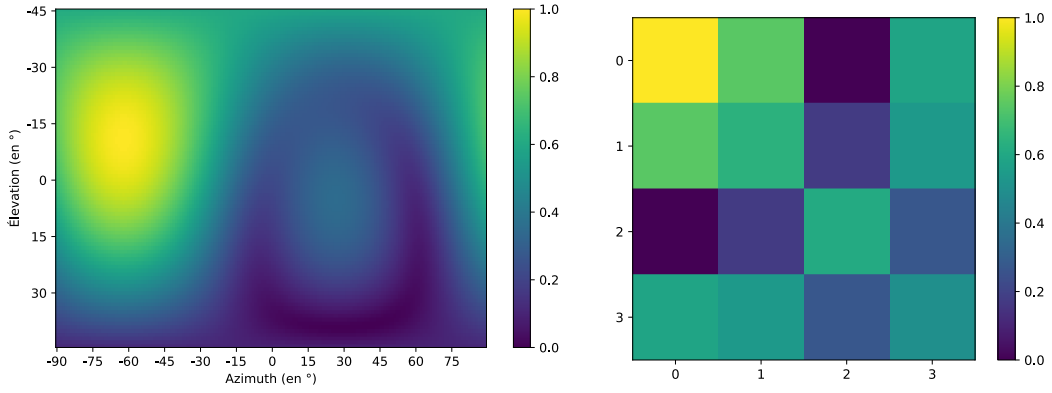
Le calcul des signaux intermédiaires s_i peut-être assez coûteux en temps de calcul. En réécrivant les équations, il est possible de calculer directement $\mathbf{P}_i$ sans avoir à extraire les signaux intermédiaires. Cela est rendu possible par l'utilisation de la matrice de covariance :

$$\mathbf{C} = \mathbf{B}\mathbf{B}^t \tag{5.3}$$

où $\mathbf{B}$ est une matrice de taille $N \times L$ composée des L échantillons des N composantes ambisoniques. Ce qui permet d'obtenir $\mathbf{P}_i$ par :

$$\mathbf{P}_i = \mathbf{d}_i^T \mathbf{C} \mathbf{d}_i \tag{5.4}$$

Le détail du calcul permettant de passer de l'équation 5.2 à l'équation 5.4 peut être trouvé à l'annexe B. Cette nouvelle équation permet un calcul de la cartographie moins coûteux, puisque les signaux intermédiaires provenant de chaque direction n'ont pas besoin d'être calculés. Ce n'est pas l'unique avantage de cette formule, la matrice de covariance $\mathbf{C}$ de dimension $N \times N$ permet d'avoir la représentation la plus compacte de l'image spatiale. Comme les valeurs de $\mathbf{d}_i$ sont indépendantes de la matrice $\mathbf{C}$ et ne dépendent que de la direction i, il est possible de calculer l'énergie du signal provenant de n'importe quelle direction i de l'espace à partir de la matrice $\mathbf{C}$ en calculant simplement l'équation 5.4. La figure 5.1 montre la cartographie d'une trame d'un signal FOA obtenu par la méthode SRP ainsi que la matrice de covariance de cette même trame. Pour faciliter la lecture de la figure, les représentations ont été normalisées par l'énergie maximale.



(a) Cartographie du signal. (b) Matrice de covariance.

FIGURE 5.1 – Cartographie de la puissance du signal FOA ainsi que la matrice de covariance associée.

5.1.2 Calcul de la correction spatiale

Pour la trame courante de taille L, les signaux ambisoniques d'entrée sont agrégés dans une matrice $\mathbf{B}$ telle que $\mathbf{B} = [[b_1(1), \ldots, b_1(L)], \ldots, [b_n(1), \ldots, b_n(L)]]$. Les signaux ambisoniques en sortie du codage multimono sont notés $\tilde{\mathbf{B}}$ et la matrice de covariance associée $\tilde{\mathbf{C}} = \tilde{\mathbf{B}}\tilde{\mathbf{B}}^t$. Notre méthode propose de trouver une matrice $\mathbf{T}$ telle que :

$$Carto(\mathbf{T}\tilde{\mathbf{B}}) \equiv Carto(\mathbf{B}) \tag{5.5}$$

On définit : $\mathbf{B}_{cor} = \mathbf{T}\tilde{\mathbf{B}}$ comme les composantes ambisoniques corrigées. Cette équivalence d'énergie (eq. 5.5) selon toute les directions i peut être exprimée par le biais des matrices de covariance, telle que :

$$\forall i, \mathbf{d}_i \mathbf{C} \mathbf{d}_i = \mathbf{d}_i \mathbf{C}_{cor} \mathbf{d}_i \tag{5.6}$$

La matrice de covariance corrigée est exprimée comme :

$$\begin{aligned} \mathbf{C}_{cor} &= \mathbf{T}\tilde{\mathbf{B}}(\mathbf{T}\tilde{\mathbf{B}})^t \\ \mathbf{C}_{cor} &= \mathbf{T}\tilde{\mathbf{C}}\mathbf{T}^t \end{aligned} \tag{5.7}$$

En remplaçant l'expression de $\mathbf{C}_{cor}$ dans l'équation 5.6, le problème revient à trouver une matrice de correction $\mathbf{T}$ qui permet de satisfaire :

$$\mathbf{T}\tilde{\mathbf{C}}\mathbf{T}^t = \mathbf{C} \tag{5.8}$$

Pour résoudre cette équation, notre méthode utilise la factorisation de Cholesky. Cette factorisation permet de décomposer une matrice réelle $\mathbf{C}$ définie positive en 2 matrices triangulaires :

$$\mathbf{C} = \mathbf{L}\mathbf{L}^t \tag{5.9}$$

De la même manière, il est possible de décomposer $\tilde{\mathbf{C}} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^t$. À partir de ces deux décompositions, il est possible de trouver une solution à $\mathbf{T}$ dans l'équation 5.8 par :

$$\begin{aligned} \mathbf{T}\tilde{\mathbf{L}}\tilde{\mathbf{L}}^t\mathbf{T}^t &= \mathbf{L}\mathbf{L}^t \\ (\mathbf{T}\tilde{\mathbf{L}})(\mathbf{T}\tilde{\mathbf{L}})^t &= \mathbf{L}\mathbf{L}^t \end{aligned} \tag{5.10}$$

Ce qui donne une solution :

$$\mathbf{T} = \mathbf{L}\tilde{\mathbf{L}}^{-1} \tag{5.11}$$

où $(.)^{-1}$ représente la matrice inverse. La factorisation de Cholesky ne peut être appliquée que sur des matrices définies positives. Comme les matrices de covariance $\mathbf{C}$ et $\tilde{\mathbf{C}}$ sont uniquement des matrices semi-définies positives, une valeur de conditionnement ε est ajoutée à la diagonale des matrices de covariance pour garantir la positivité des matrices. Dans notre méthode, la valeur de ε est fixée à 10^{-9}.

5.2 Description détaillée du codec

La figure 5.2 montre l'architecture du codec complet de notre méthode. Dans la méthode, le codec cœur fonctionne avec des trames de 20 ms. Pour un signal ambisonique échantillonné à 32 kHz, la longueur de trame est de $L = 640$. Dans sa conception, la méthode n'impose pas de contrainte sur l'ordre ambisonique du signal d'entrée. La description du codec ci-dessous sera donc une implémentation générique fonctionnant pour n'importe quel ordre ambisonique. Cependant, le nombre de composantes augmente fortement avec l'ordre ambisonique, pour la transmission d'un signal HOA, une approche de codage avec extraction des sources prédominantes et de l'ambiance sonore, type MPEG-H, pourrait permettre d'atteindre une meilleure qualité pour un débit équivalent à une approche où toutes les composantes sont transmises.

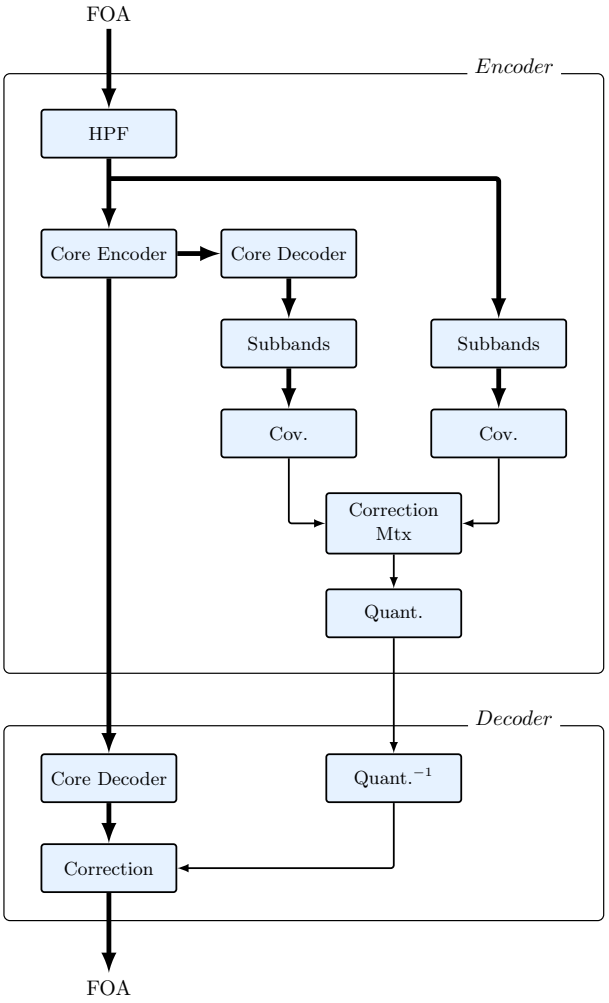


FIGURE 5.2 – Schéma complet du codage FOA par post-traitement.

5.2.1 Description du codeur

Chaque composante du signal ambisonique d'entrée est tout d'abord filtrée par un filtre passe-haut de fréquence de coupure 20 Hz. Comme pour la méthode par rematriçage PCA, les coefficients du filtre IIR sont repris du filtre utilisé dans le codec EVS décrit dans [3GPP TS 26.445, 2019]. Ce filtrage a pour but de supprimer tout biais possible lors du calcul des matrices de covariance. Dans le codeur, chaque composante est ensuite codée et décodée par un codec cœur mono.

Le signal ambisonique d'origine $\mathbf{B}_t$ et le signal décodé $\tilde{\mathbf{B}}_t$ sont ensuite découpés en J sous-bandes. Notre méthode utilise un découpage en bandes critiques de Bark, présentée dans le chapitre 2.1.3. Les bandes critiques de Bark sont constituées de 24 sous-bandes, chacune définie par une fréquence centrale et une largeur fréquentielle. Envoyer une matrice de correction $\mathbf{T}$ pour chaque sous-bande nécessiterait une quantité trop importante de métadonnées. Pour limiter le débit des métadonnées, le nombre de sous-bandes a dû être limité. Les bandes de Bark sont regroupées par 4 pour obtenir un total de 7 sous-bandes. Le tableau 5.1 dresse la liste des 7 sous-bandes utilisées

dans notre méthode ainsi que la correspondance avec les bandes de Bark originales. Pour chaque sous-bande une matrice de correction $\mathbf{T}$ est calculée et transmise au décodeur.

Tableau 5.1 – Découpage en sous-bandes du signal FOA

Numéro de la bande	Correspondance bandes de Bark	Fréquence centrale (Hz)	Fréquence de coupure (Hz)	
1	$1-4$	325	20	630
2	$5-8$	855	630	1080
3	$9-12$	1400	1080	1720
4	$13-15$	2210	1720	2700
5	$16-18$	3550	2700	4400
6	$19-21$	6050	4400	7700
7	$22-24$	11600	7700	15500

Les matrices de covariance sont calculées sur des trames d'analyse de 40 ms correspondant à 2 trames audio t et $t-1$. Pour garantir une transition progressive entre les corrections, un fenêtrage de Hann est appliqué à la trame d'analyse avec un recouvrement de 50% entre 2 trames d'analyse consécutives. La matrice de correction appliquée à la trame audio t est donc calculée à partir d'une matrice de covariance basée sur les trames audio t et $t-1$.

Pour chaque sous-bande j, les matrices de covariance $\mathbf{C}_t(j)$ et $\tilde{\mathbf{C}}_t(j)$ sont calculées par :

$$\mathbf{C}_t(j) = \mathbf{B}_t(j)\mathbf{B}_t(j)^t \tag{5.12}$$

Une fois les matrices de covariance calculées, le codeur en déduit les matrices de correction spatiale pour chaque sous-bande $\mathbf{T}_t(j)$ grâce à l'équation 5.10. La matrice $\mathbf{T}_t(j)$ est normalisée pour préserver un niveau sonore identique pour la composante omnidirectionnelle (W) décodée avant et après post-traitement. Cette normalisation est là pour assurer que la correction corrige uniquement les déformations spatiales et que la correction appliquée ne compense pas une réponse en fréquence du codec cœur. Pour certains codecs, la réponse en fréquence peut avoir une certaine coloration notamment dans les plus hautes fréquences.

5.2.2 Détails des métadonnées transmises

Pour chaque sous-bande j de la trame t, la matrice de correction $\mathbf{T}_t(j)$ doit être quantifiée avant d'être transmise au décodeur. Comme les matrices $\mathbf{L}$ et $\tilde{\mathbf{L}}$, définies précédemment sont triangulaires, la matrice de correction $\mathbf{T}_t(j)$ est aussi triangulaire. Seuls les coefficients du triangle inférieur de la matrice $\mathbf{T}_t(j)$ ont besoin d'être quantifiés et transmis. Pour le cas d'un signal FOA, 10 coefficients de la matrice $\mathbf{T}_t(j)$ doivent être transmis par sous-bande. La matrice de correction $\mathbf{T}_t(j)$ ne modifie pas la composante W, le coefficient $\mathbf{T}_t(j)[0][0]$ de la matrice corrigeant la composante W est toujours égal à 1. Il n'a donc pas besoin d'être transmis. Pour chacune des 7 sous-bandes, 9 coefficients sont transmis chacun codé sur 8 bits. La totalité des informations de correction de l'image spatiale nécessite un débit total de $25{,}2$ kbit/s.

Les valeurs des coefficients n'étant pas uniformément distribuées, une quantification scalaire utilisant la loi-μ [ITU-T G.711, 1988] a été mise en place. Ce type de quantification permet de faire une transformation non-linéaire des valeurs avant la quantification uniforme pour permettre une quantification plus fine des valeurs les plus fréquentes. Cette transformation non-linéaire est pilotée par un facteur de compression. Ce facteur découle de deux variables, la plage de valeurs que l'on veut pouvoir quantifier et la variance supposée de la distribution des valeurs des coefficients. Dans notre méthode, ce facteur est commun pour l'ensemble des coefficients des matrices $\mathbf{T}$ de toutes les sous-bandes. Ce facteur est déterminé comme :

$$c = \sqrt{2} \times \frac{v_{max}}{3\sigma^2} \tag{5.13}$$

avec v_{max} la valeur absolue maximale que peut prendre un coefficient et σ^2 la variance de la distribution des valeurs des coefficients fixée de manière empirique. Pour v_{max} la valeur a été fixée à 10, la valeur des coefficients peut aller de $]-10, 10[$. Les coefficients en dehors de ces bornes sont tronqués à la valeur maximale (resp. minimale). D'après nos observations sur un large ensemble d'échantillons audio, nous avons fixé la variance pour l'équation à $\sigma^2 = 3{,}5$ pour notre méthode.

$$F(x) = v_{max} \times sgn(x) \frac{1 - e^{-|x|\frac{c}{v_{max}}}}{1 - e^{-c}} \tag{5.14}$$

où x est un coefficient de la matrice $\mathbf{T}$, $sgn(.)$ indique le signe de x : -1 si $x < 0$, $+1$ si $x \geq 0$. Une fois la transformation appliqué aux coefficients, ils sont quantifiés par une quantification uniforme avec 8 bits par coefficient, soit 256 valeurs différentes. La figure 5.3 montre la projection des coefficients entre l'espace d'origine et l'espace quantifié. Les valeurs extrêmes sont quantifiées avec une moins grande précision que les plus courantes (les valeurs proches de 0).

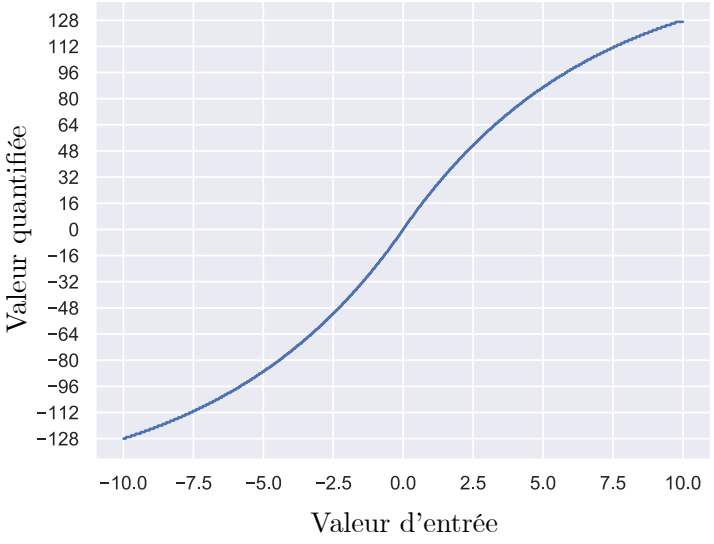


FIGURE 5.3 – Projection des coefficients entre l'espace d'entrée vers l'espace quantifié.

Il est à noter que les valeurs de la diagonale $\mathbf{T}_t(j)$ sont toujours positives, il serait donc possible de réduire la plage des valeurs quantifiées pour ne considérer que l'intervalle de $[0, 10]$; Cela permet d'économiser 1 bit pour 3 coefficients pour chaque sous-bande, soit une économie total de $1{,}05$ kbit/s. L'économie de débit étant relativement faible, l'optimisation n'a pas été intégrée à l'implémentation évaluée.

5.2.3 Description du décodeur

Le fonctionnement du décodeur est relativement similaire au codeur. Les signaux reçus sont décodés par le codec cœur mono. Une fois décodées, les composantes sont découpées en $J = 7$ sous-bandes. Ces sous-bandes sont les mêmes que celles utilisées dans l'encodeur.

En parallèle, les métadonnées sont décodées et les coefficients de la matrice $\mathbf{T}$ sont reconstituées. La transformation inverse est appliquée pour retrouver les coefficients d'origine :

$$F^{-1}(x) = \frac{-v_{max}}{c} \times sgn(x) \ln\left(1 - \frac{1 - e^{-c}}{v_{max}} |x|\right) \tag{5.15}$$

Une fois les matrices reconstituées, elles sont ensuite appliquées à chaque sous-bande du signal décodé pour corriger la spatialisation et ainsi retrouver la même répartition de l'énergie dans l'espace que le signal ambisonique original.

5.3 Test subjectif

5.3.1 Conditions du test Ref AB

Le test subjectif du chapitre 4 nous a déjà renseigné sur la qualité atteignable par l'approche multimono. Pour cette évaluation, l'important est de mesurer si le post-traitement apporte une amélioration par rapport à la méthode multimono. Pour évaluer la performance relative des deux méthodes, un test subjectif RefAB est réalisé. Dans ce type de test, il est demandé aux participants de comparer 2 conditions (A et B) entre elles. L'une des deux conditions correspond à l'échantillon codé par le multimono basique, l'autre condition correspond à la méthode avec post-traitement. Les deux conditions sont généralement codées avec le même débit total. Une présentation détaillée de la méthodologie RefAB peut être trouvée dans le chapitre 2.2.1.3.

Pour chaque échantillon, le participant doit noter la paire de conditions selon une échelle discrète allant de -3 à $+3$. Un score positif signifie que la condition A est perçue comme de meilleure qualité que la condition B. De manière analogue, un score négatif signifie que la condition B est de meilleure qualité que la condition A. L'échelle de notation est une échelle graduée, à chaque graduation un terme est associé. Le tableau 5.2 indique le terme associé à chaque graduation. Pour s'assurer que le participant ne note pas une préférence, mais réellement la fidélité audio, une référence explicite est ajoutée à côté des 2 autres conditions. La référence explicite dans notre test est le signal FOA d'origine. Le test est effectué sur casque audio, les conditions de test sont toutes binauralisées par les mêmes filtres binauraux.

Tableau 5.2 – Échelle de notation pour le test Ref AB.

Échelon	Comparaison
$+3$	A is much better than B
$+2$	A is better than B
$+1$	A is slighly better than B
0	A is the same as B
-1	B is slighly better than A
-2	B is better than A
-3	B is much better than A

Lors des écoutes préparatoires à ce test nous avons constaté que les filtres de binauralisation utilisés par le moteur de rendu Resonance Audio [Gorzel *et al.*, 2019] apportaient une coloration importante aux signaux binauralisés. Cette coloration est plus fortement perçue en hautes fréquences, où les filtres binauraux apportent une forte atténuation. Les artefacts de codage étant généralement plus présents dans les parties extrêmes du spectre, une telle coloration pourrait masquer une partie des artefacts hautes fréquences. Dans le contexte de test sur la qualité audio, nous avons décidé de nous tourner vers un autre moteur de rendu utilisant des filtres binauraux plus neutres en termes de coloration.

Malgré des recherches, nous n'avons pas trouvé de littérature sur la comparaison de coloration apportée par les filtres HRTF non personnalisés et l'impact que cela pouvait avoir sur la qualité perçue. Le choix d'un moteur de rendu binaural a donc été guidé par un ensemble d'écoutes informelles pour estimer quel moteur de rendu semblait être le plus neutre. Notre choix s'est porté sur le moteur de rendu *SPARTA* [McCormack et Politis, 2019], dans sa configuration par défaut. En plus d'apporter une coloration moins importante, ce moteur est *open source* et possède une documentation détaillée.

Notre méthode étant conçue pour fonctionner indépendamment du codec cœur utilisé, nous avons mené deux tests subjectifs indépendants. Un premier avec le codec cœur EVS et un second avec Opus dans le but de vérifier que notre méthode apporte effectivement une amélioration de qualité quel que soit le codec utilisé.

Les signaux FOA d'entrée utilisés pour les tests sont échantillonnés à 32 kHz. Les échantillons ambisoniques critiques utilisés pour le test ont été sélectionnés pour représenter une variété de contenu : musique, voix, milieu réverbérant, ainsi que des signaux difficiles à coder comme des sons percussifs, bruits colorés. Le détail des échantillons utilisés pour les tests peut être trouvé à l'annexe A. Le codec EVS est paramétré pour coder les signaux avec une largeur de bande de 16 kHz, soit du *Super Wideband* (SWB). Le codec Opus ne prend pas en charge les signaux échantillonnés à 32 kHz. Pour garder les mêmes échantillons et obtenir des résultats comparables pour le codec Opus, les signaux échantillonnés à 32 kHz sont sur-échantillonnés à 48 kHz, codés en *Full-band* (FB) puis sous-échantillonnés à 32 kHz.

Pour chaque codec, deux débits sont utilisés : un débit faible R_1 et un débit moyen R_2. La qualité à bas débit n'étant pas exactement la même entre les deux codecs. Les débits R_1 et R_2 ne sont pas identiques pour les deux codecs. Pour EVS, $R_1 = 97{,}6$ ($4 \times 24{,}4$) kbit/s et $R_2 = 128$

(4×32) kbit/s. Pour Opus, $R_1 = 128$ (4×32) kbit/s et $R_2 = 160$ (4×40) kbit/s. Ce débit est un débit global, comprenant le débit pour coder les composantes ainsi que de les métadonnées pour le cas de notre méthode. Toutes les conditions de test sont récapitulées dans le tableau 5.3. Le codec EVS ayant un nombre limité de débits possibles (9,6, 13,2, 16,4, 24,4, 32,0 kbit/s...), pour certaines conditions, une partie du débit total n'est pas consommé. Ce débit disponible aurait pu être alloué à une plus grande quantité de métadonnées, cependant nous avons préféré garder les mêmes métadonnées pour les deux tests (25,2 kbit/s) pour une meilleure comparabilité entre les deux tests.

Ces tests subjectifs ont été menés sur deux panels de sujets différents : un panel d'experts, composé d'individus professionnels du domaine de la compression ou du traitement audio et un panel de sujets naïfs, composé d'individus volontaires, sans perte d'audition connue. Ce choix a été fait pouvoir évaluer la capacité de personnes non-sensibilisées aux artefacts spatiaux à déceler les dégradations liées à leur présence. De plus, la comparaison des résultats des deux panels a permis d'étudier si l'impact des artefacts spatiaux sur la qualité globale est la même, quelle que soit la familiarité du sujet avec le domaine du codage audio. Le nombre de participants a été respectivement de 11 pour le panel expert et de 13 pour le panel naïf.

Tableau 5.3 – Paires de conditions utilisées pour le test CCR ainsi que leurs débits (en kbit/s).

Codec cœur utilisé	**débit multimono**	**débit pour la méthode proposée**
EVS	$4 \times 24{,}4$ (97,6)	$4 \times 16{,}4 + 25{,}2$ (90,8)
EVS	4×32 (128)	$4 \times 24{,}4 + 25{,}2$ (122,8)
Opus	4×32 (128)	$4 \times 25{,}7 + 25{,}2$ (128)
Opus	4×40 (160)	$4 \times 33{,}7 + 25{,}2$ (160)

Les écoutes ont été réalisées sur casque dans une salle traitée acoustiquement. Le casque utilisé est le *Sennheiser HD 650* et la carte son *Focusrite Scarlett 6i6*. Le matériel a été le même pour tous les participants. Le test étant fait sur des sujets naïfs, une session de familiarisation au test de qualité audio a été réalisée. Cette phase de familiarisation se compose d'une brève introduction sur les sons spatiaux, suivie d'un test d'entrainement constitué de 4 échantillons, ceci dans le but de faire entendre quelques exemples de dégradation possible. Ce test est réalisé seul, sans la présence de l'expérimentateur. La même session de familiarisation a été menée pour le panel de sujets expérimentés.

5.3.2 Résultat des tests

La figure 5.4 montre le résultat échantillon par échantillon pour chacun des tests (Opus et EVS). Les résultats agrègent les résultats des deux panels et les différents débits. En moyenne, la méthode proposée apporte une amélioration par rapport à l'approche multimono de base, quel que soit le codec cœur utilisé (Opus et EVS), ainsi que pour les deux débits testés (R_1, R_2). Le post-traitement ne corrigeant pas les artefacts fréquentiels, l'amélioration de qualité est donc imputable uniquement à la correction de l'image spatiale.

Les résultats montrent qu'en moyenne notre post-traitement apporte une amélioration de la

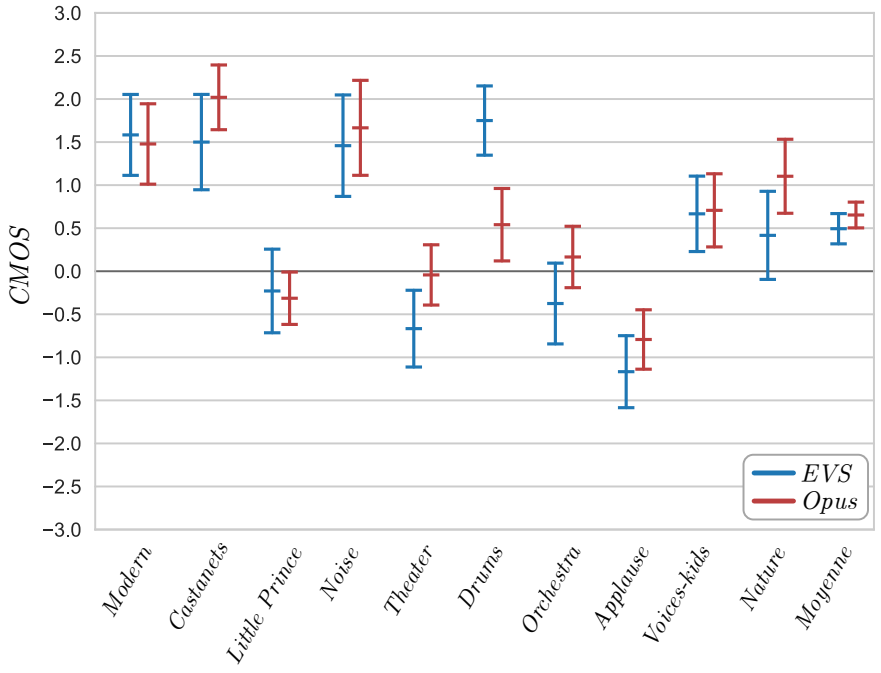


FIGURE 5.4 – Moyenne des scores obtenus avec l'intervalle de confiance à 95%. Un score positif (resp. négatif) correspond à une amélioration (resp. dégradation) apportée par notre méthode par rapport à la méthode multimono classique.

qualité et ce pour les deux codecs. Seuls trois échantillons voient leur qualité significativement plus basse que sans le post-traitement. Ces trois des échantillons sont : *Applause*, *Theater*, *Little Prince*. Plusieurs pistes peuvent apporter des éléments d'explication de cette dégradation de leurs qualités. La correction spatiale est faite sur la cartographie moyenne du contenu sur les trames t et $t-1$, soit 40 ms. Si durant ce même intervalle de temps plusieurs événements sonores se produisent dans des directions différentes, la cartographie moyenne calculée ne permettra pas d'obtenir la position correcte des sources. Par conséquent, la correction spatiale ne permet pas de retrouver l'image spatiale d'origine. Pour l'échantillon *Applause*, le nombre important d'impacts et leur dispersion produiraient une matrice de correction incohérente, ce qui pourrait expliquer cette dégradation de la qualité globale. Pour les échantillons *Theater* et *Little Prince*, le même phénomène serait en cause, ici ce serait la forte réverbération du signal ambisonique qui pourrait perturber le calcul de la cartographie du signal.

Une seconde explication pour la dégradation serait que cette amplification dans une zone de l'espace par la matrice $\mathbf{T}$ sur l'ensemble de la durée de la trame provoquerait un effet de pré/post-écho spatial. Sur le même principe que le pré/post-écho fréquentiel pour les sons percussifs, le bruit de fond dans une zone de l'espace se retrouverait amplifié avant et après un impact.

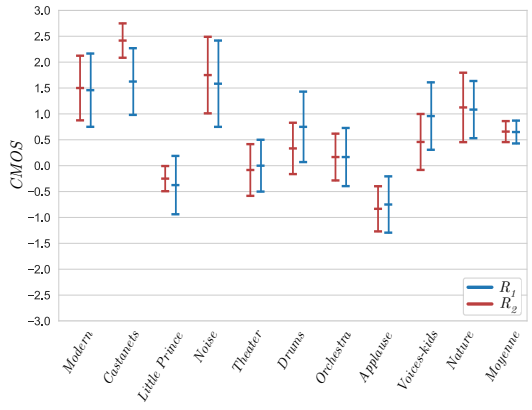
La dernière piste pour expliquer la dégradation des scores de certains échantillons serait que notre méthode alloue moins de débit au codage des composantes, par rapport au codage multimono classique, la métadonnée consommant une partie du débit total. Ce débit moins important entraine mécaniquement une dégradation plus importante de la qualité audio des composantes par rapport à

l'approche multimono. La qualité perdue doit normalement être compensée par le gain de qualité obtenu grâce à la correction spatiale. Ce qui devrait permettre d'obtenir une qualité globale plus importante. Cependant, il est envisageable que, pour certains contenus, la dégradation du signal soit plus importante que le gain de qualité spatiale. Les échantillons du type *Applause* sont connus pour être des sons difficiles à traiter pour les codecs. Il est possible que pour ces contenus percussifs, retirer du débit au codec cœur dégrade plus fortement la qualité que pour les autres types de contenu.

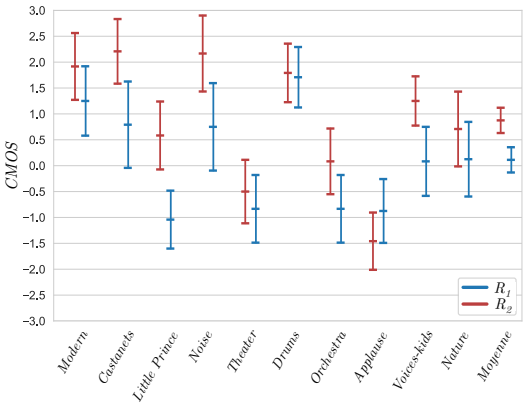
Résultats détaillés pour la méthode par post-traitement

La figure 5.5 montre le résultat détaillé pour chaque échantillon en fonction d'un facteur pris en particulier. Les figures 5.5(a) et 5.5(b) montrent les résultats obtenus au test pour le codec Opus et pour le codec EVS selon le débit utile R_1 ou R_2. Pour le codec Opus, les scores sont très similaires pour les deux débits, là où pour le codec EVS, le score de certains échantillons est différent entre le débit R_1 et R_2, comme pour l'échantillon *Little Prince* ou *Noise*. Cette différence de résultats entre les deux débits pourrait être expliquée par la grande variété de modes de codage du codec EVS, la sélection du mode de codage dépendant du contenu du signal audio et du débit. Pour certains échantillons, les modes utilisés pour coder l'échantillon au débit R_1 n'ont peut-être pas été les mêmes que pour le débit R_2. Cela semble indiquer que le post-traitement n'apporte pas le même gain pour tout le mode de codage. Il serait intéressant d'étudier le gain apporté par le post-traitement en fonction du mode de codage. En fonction de ce gain, il serait possible de déterminer pour quel mode le post-traitement devrait être activé. Pour Opus, le codec est beaucoup homogène dans son fonctionnement et le nombre de modes internes est plus réduit, ce qui se traduit dans des résultats plus consistants entre R_1 et R_2. Les figures 5.5(c) et 5.5(d) montrent les mêmes résultats que les deux figures précédentes, cette fois-ci en les regroupant en fonction du débit.

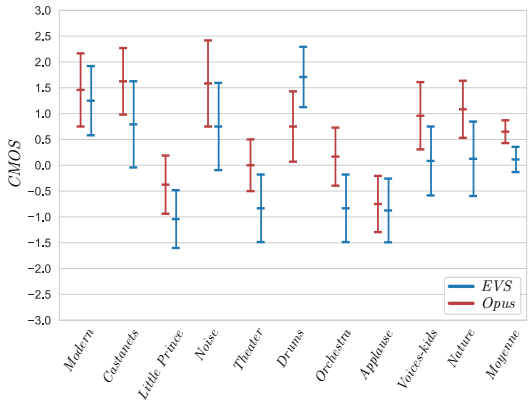
Les figures 5.5(e) et 5.5(f) présentent les résultats obtenus pour les débits R_1 et R_2 en fonction de l'expertise du panel de sujet : *Naïf* ou *Expert*. La forme générale des résultats est similaire entre les deux figures. Par exemple, le gain apporté pour le post-traitement est plus important pour l'échantillon *Drum* que pour *Orchestra* pour les deux échantillons. Cette similarité concerne également la différence de gain apporté par la méthode selon le débit. Par exemple, la relation entre le gain de R_1 et R_2, pour l'échantillon *Noise*, suit la même tendance, quel que soit le panel. Par ailleurs, comme il était possible de s'y attendre, les intervalles de confiance sont plus importants pour le panel *Naïf* que pour le panel *Expert*. Cela peut être expliqué par la familiarité du panel *Expert* avec l'ambisonique et les artefacts de compression audio.



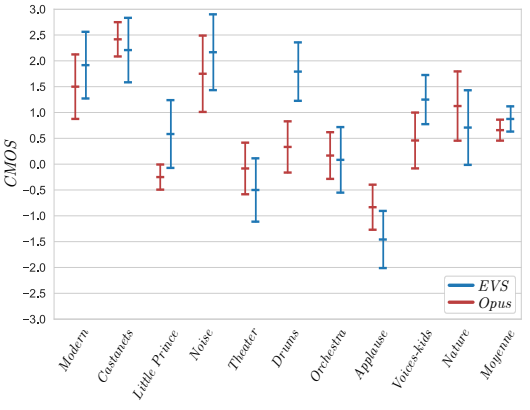
(a) Résultats pour le codec Opus selon les deux débits R_1 et R_2



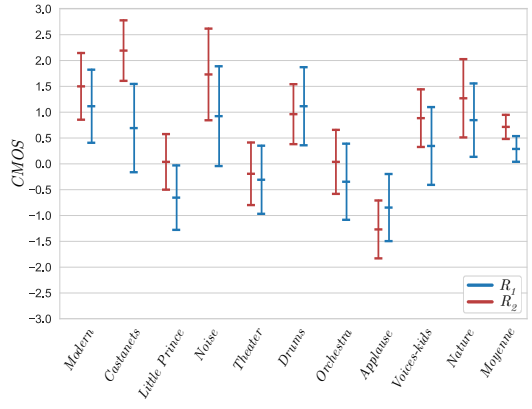
(b) Résultats pour le codec EVS selon les deux débits R_1 et R_2



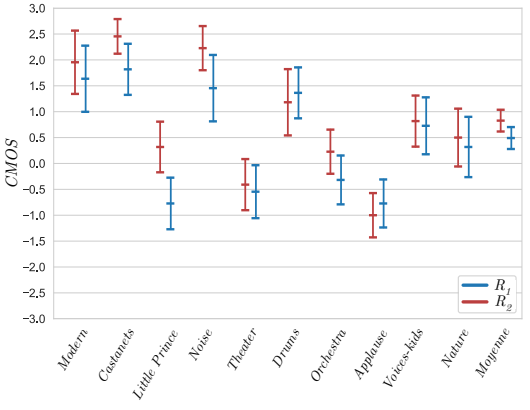
(c) Résultats pour le débit R_1 selon les deux codecs EVS et Opus.



(d) Résultats pour le débit R_2 selon les deux codecs EVS et Opus.



(e) Résultats pour le panel Naïf selon les deux débits R_1 et R_2



(f) Résultats pour le panel Expert selon les deux débits R_1 et R_2

FIGURE 5.5 – Résultats obtenus en fonction des différents facteurs. Score moyen avec l'intervalle de confiance à 95%.

5.3.3 Analyse statistique par ANOVA

Pour analyser plus en détail l'influence des différents facteurs (codecs cœur, débits…) sur l'amélioration de qualité audio mesurée lors des tests, une analyse de variance (ANOVA) a été conduite [Miller Jr, 1997]. Les différentes variables pouvant influer la qualité ont été regroupées en un ensemble de facteurs : le facteur *Codec*, regroupant les deux codecs cœur, le facteur *Débit* regroupant les deux débits, le facteur *Expertise*, correspondant aux deux panels, le facteur *Échantillon* correspondant aux 10 échantillons différents. Une première analyse de variance a été faite pour comparer l'importance du facteur *Expertise* sur la qualité audio estimée. Pour cela, l'analyse a été faite en comparant le score CMOS, selon les facteurs intragroupes *Codec*, *Bitrate*, *Échantillon* et le facteur intergroupe *Expertise*. Cette première analyse a montré qu'il n'y avait pas d'effet significatif du facteur *Expertise* sur les résultats. De plus, ce facteur n'a pas d'interaction significative sur les autres facteurs. Cela suggère que les deux panels de sujets ont un système de notation similaire et qu'il n'est pas pertinent de conserver ce facteur dans les analyses suivantes. Cette analyse valide l'hypothèse que la perception et l'impact des artefacts spatiaux sont identiques pour les participants, quelle que soit la familiarité avec le domaine du codage et son impact sur la notation.

Une seconde analyse de variance a été réalisée en excluant le facteur *Expertise*, mais en gardant les intergroupes : *Codec*, *Débit*, *Échantillon*. Le tableau 5.4 récapitule les résultats détaillés de l'analyse. Cette figure montre que le facteur *Débit* et *Échantillon* ont tous les deux une influence significative sur la qualité audio, ainsi qu'une interaction bidirectionnelle entre ces deux facteurs. Ces effets et interactions sont illustrés par la figure 5.6. Le résultat montre qu'il n'y a pas de dépendance au facteur *Codec*. Ces résultats suggèrent que le post-traitement devrait fonctionner pour un codage multimono utilisant d'autres codecs cœur que les deux codecs évalués.

Tableau 5.4 – Résultats de l'analyse *ANOVA* pour les facteurs intragroupes "Codec", "Débit" et "Échantillon".

Effet	**Deg. de liberté**	**Ratio F**	**Valeur p**
Codec	1	3,71	0,067
Bitrate	1	10,12	0,004
Item	9	27,64	0,000
Codec * Bitrate	1	28,39	0,000
Codec * Item	9	3,43	0,001
Bitrate * Item	9	2,63	0,007
Codec * Bitrate * Item	9	1,22	0,283

5.4 Limites de la méthode

5.4.1 Retard spatial lié à la méthode

Il est possible de reprendre l'exemple utilisé à la section 4.1.1 pour mettre en lumière la déformation spatiale, pour cette fois visualiser l'effet du post-traitement sur la cartographie du signal. Pour rappel, dans l'exemple précédent, nous avions utilisé un signal ambisonique HOA contenant

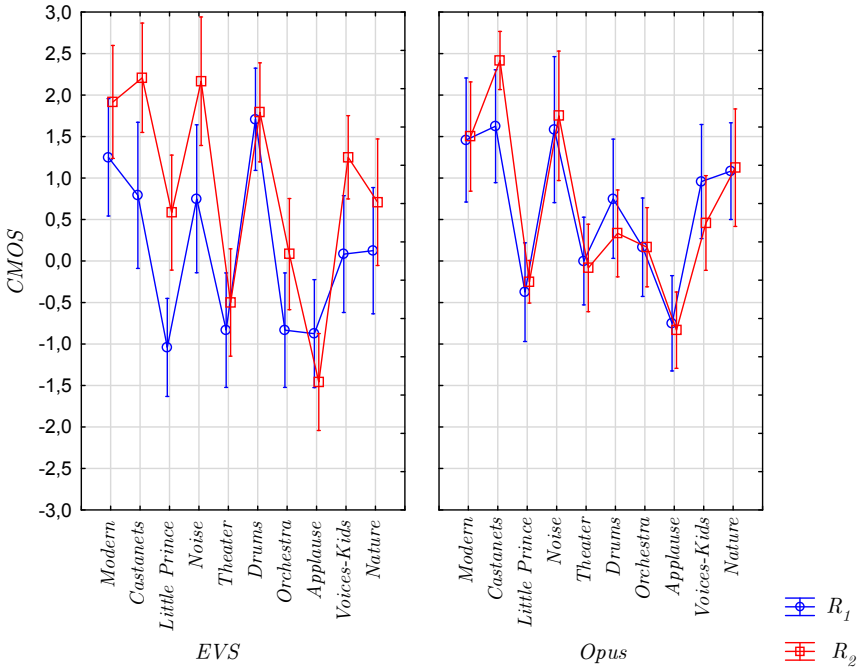


FIGURE 5.6 – Résultat du test en fonction du codec utilisé : EVS, Opus et du débit : R_1, R_2 pour chaque échantillon.

une unique source : un bruit rose qui fait le tour de l'auditeur sur le plan horizontal.

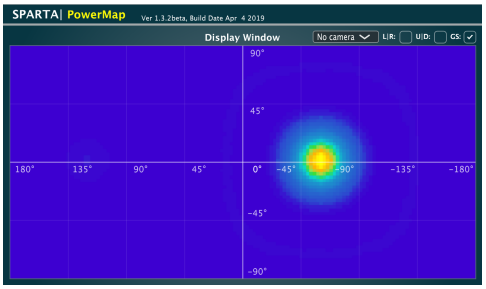
Les figures 5.7(a) et 5.7(b) montrent la cartographie de la puissance du signal ambisonique original, quand la source est dans la direction $(\theta, \phi) = (-60°, 0°)$, et le même signal codé avec une approche multimono classique à 16 kbit/s par composante. Dans cette deuxième figure, les artefacts spatiaux sont clairement visibles sur la cartographie.

La figure 5.7(c), quant à elle, représente le signal codé avec l'approche par post-traitement. Les artefacts spatiaux ont été atténués au point qu'ils ne soient plus visibles sur la cartographie. Pour autant la cartographie n'est pas parfaitement identique à la cartographie originale. La forme de la source n'est plus parfaitement circulaire, mais possède une trainée. Cette trainée peut être expliquée par le fonctionnement de la méthode combinée avec le déplacement de la source au cours du temps. Le calcul de la matrice de covariance est fait sur deux trames t et $t-1$ de 20 ms, soit une trame d'analyse de 40 ms. Comme la source est en mouvement, la covariance va capturer la position moyenne de la source sur cette période. Cette matrice de covariance moyenne est ensuite utilisée pour calculer la matrice de correction. La matrice de correction va être appliquée uniquement à la trame courante t, ce qui va déformer la forme de la source en ajoutant un retard spatial.

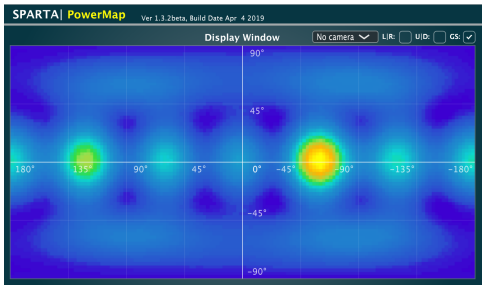
Pour les sources fixes, ce fonctionnement ne devrait pas impacter la localisation des sources. Il serait même possible d'exploiter ce phénomène pour améliorer la précision spatiale, comme cela a été présenté dans [McCormack *et al.*, 2019].

Pour les sources en mouvement et plus particulièrement pour les sources très rapides, l'impact devrait être beaucoup plus important, avec un étalement important de l'énergie dans l'espace.

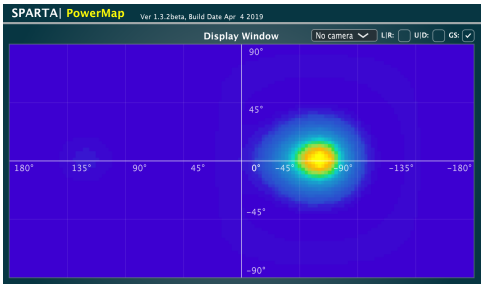
Ce phénomène pourrait également expliquer la différence de résultats lors des tests entre l'échantillon *Applause* et *Drums*, et pourquoi sur l'échantillon *Applause* la qualité est dégradée par le post-traitement alors que sur l'échantillon *Drums* la qualité est améliorée. L'échantillon *Applause* contient un grand nombre de sons percussifs, des claquements de mains, provenant de directions proches de l'aléatoire. L'échantillon *Drums* contient quant à lui un grand nombre de sons percussifs, des frappes sur une batterie, mais la direction de chaque source est relativement fixe dans le temps, correspondant aux positions des fûts de la batterie.



(a) Original



(b) Codé par la méthode multimono classique



(c) Codé par notre approche

FIGURE 5.7 – Cartographie de la puissance du signal ambisonique.

5.4.2 Cas particulier d'Opus mode ambisonique

Certains codecs, comme Opus, proposent une approche de codage dite ambisonique. Le mode ambisonique *family mapping 3* d'Opus a été présentée à la section 3.3.1.2. Dans ce mode, les composantes ambisoniques sont rematricées vers une représentation de l'espace équivalent s'apparentant à un format-A théorique. Dans le format-A théorique, chacun des 4 canaux générés correspond au signal capté par 4 microphones parfaitement coïncidents. Ces 4 microphones sont positionnés sur les faces d'un tétraèdre. À partir de ces 4 canaux, 2 paires de canaux sont formées. Le codec Opus code chacune des paires de canaux avec un codec cœur stéréo indépendant.

Dans cette section, nous explorons l'efficacité de notre post-traitement sur cette approche de codage ambisonique, aussi appelé multistéréo. Nous avons adapté notre post-traitement pour le faire fonctionner avec une approche multistéréo pour des signaux FOA. La figure 5.8 montre le fonctionnement de l'approche multistéréo avec post-traitement. Comme avec l'approche multimono, notre traitement est placé en amont et en aval du codage multistéréo.

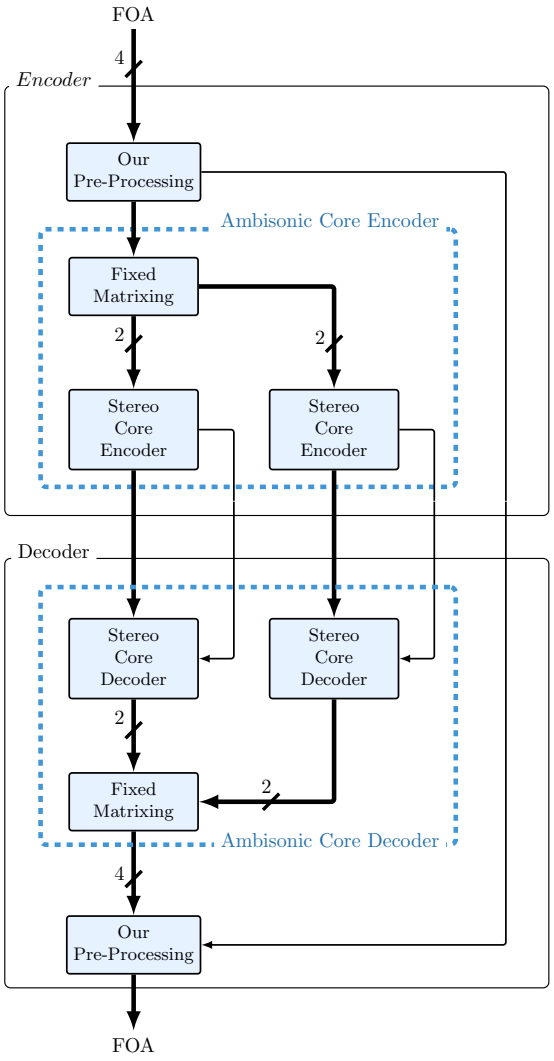


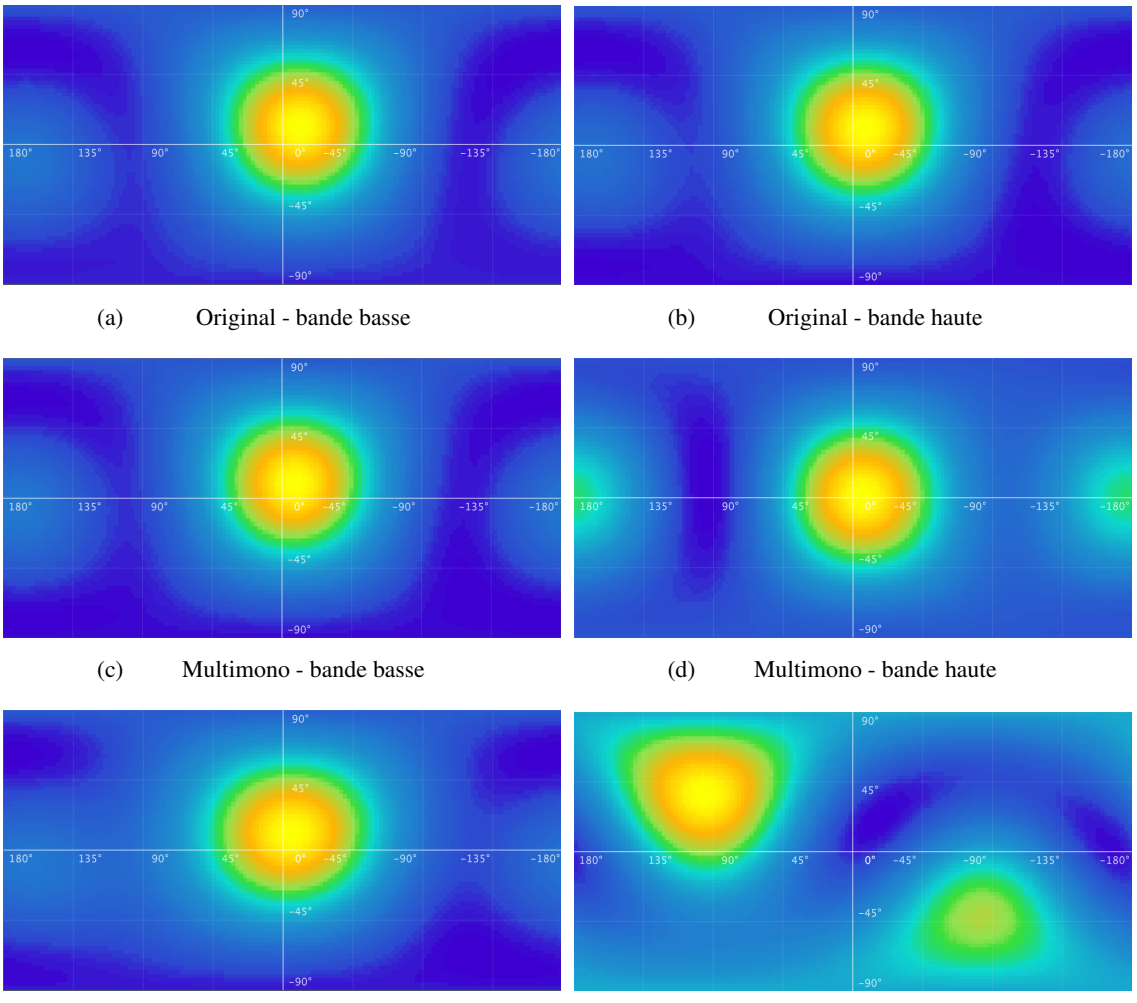
FIGURE 5.8 – Schéma du codec par post-traitement avec le mode multistéréo d'Opus

Résultat du post-traitement pour l'approche multistéréo

En observant les images spatiales entre les signaux d'origine et codés par l'approche multistéréo puis corrigés, le post-traitement semble fonctionner correctement. Cependant lors de l'écoute de ces signaux, même si les directions des sources sont bonnes, la qualité audio des sources est très dégradée par rapport à l'original. Des craquements sont présents dans le signal ainsi que des bruits larges bandes dans les hautes fréquences. A cause de ces dégradations, la qualité audio s'en trouve grandement diminuée. Nous avons analysé les mécanismes qui empêchent le post-traitement d'améliorer la qualité globale des signaux.

Pour la bande basse, le codec cœur stéréo reproduit la forme d'onde des quatre canaux. Les déformations spatiales ajoutées aux composantes sont similaires aux déformations apportées lors du codage par l'approche multimono. La méthode par post-traitement arrive donc à corriger les déformations du signal FOA et à retrouver une image spatiale proche de l'original.

Pour la bande haute, le codec cœur stéréo fait un panoramique d'intensité pour chaque paire de canaux. La dégradation audio viendrait donc de la bande haute. Le panoramique d'intensité semble produire un signal ambisonique avec une répartition de l'énergie spatiale trop éloignée de la répartition d'énergie d'origine pour être corrigée par le post-traitement. La matrice de correction devrait amplifier fortement des zones de l'espace où l'énergie du signal codé était faible. Cette forte amplification expliquerait la dégradation de la qualité audio.



(a) Original - bande basse (b) Original - bande haute

(c) Multimono - bande basse (d) Multimono - bande haute

(e) Multistéréo - bande basse (f) Multistéréo - bande haute

FIGURE 5.9 – Image spatiale, pour la bande basse et la bande haute, de l'échantillon *Voices-Kids* au temps $t = 6{,}13s$ codé avec différentes conditions codage.

La figure 5.9 montre différentes images spatiales du même signal FOA. Les lignes correspondent aux conditions différentes : le signal d'origine, le signal codé par l'approche multimono et codé par l'approche multistéréo, avant le post-traitement. La première colonne sont les images spatiales de la bande basse, allant de 50 Hz à 6 kHz. La seconde colonne correspond aux images spatiales de la bande haute, allant de 6 kHz à 22 kHz. Le signal utilisé est *Voices-Kids*. Pour rappel, dans cet échantillon, deux personnes discutent au milieu d'un brouhaha de cour d'école. La des-

cription détaillée peut être trouvée à l'annexe A. La cartographie a été prise à un moment où une seule des personnes parle, un fort bruit ambiant est présent. Pour la bande basse, les cartographies des signaux multimono et multistéréo (figure 5.9(c) et 5.9(c)) sont assez proches de la cartographie d'origine (figure 5.9(a)). Pour la bande haute, la cartographie du signal multimono (figure 5.9(d)) est plus éloignée de la cartographie d'origine (figure 5.9(b)), cependant avec une matrice de correction adaptée, l'image spatiale d'origine peut être restaurée. Pour le signal multistéréo (figure 5.9(f)), la cartographie est vraiment très éloignée de la répartition d'origine. Il n'y a pas un simple ajout d'énergie dans certaines directions de l'espace, mais un changement total de cette répartition de l'espace, au point où la direction $(\theta, \phi) = (-10°, 0°)$ d'où devrait provenir le plus d'énergie est l'un des endroits où la quantité d'énergie est minimale. Ce manque d'énergie demande une forte amplification du signal venant de cette direction de la part de la matrice de correction et donc une augmentation du bruit dans le signal.

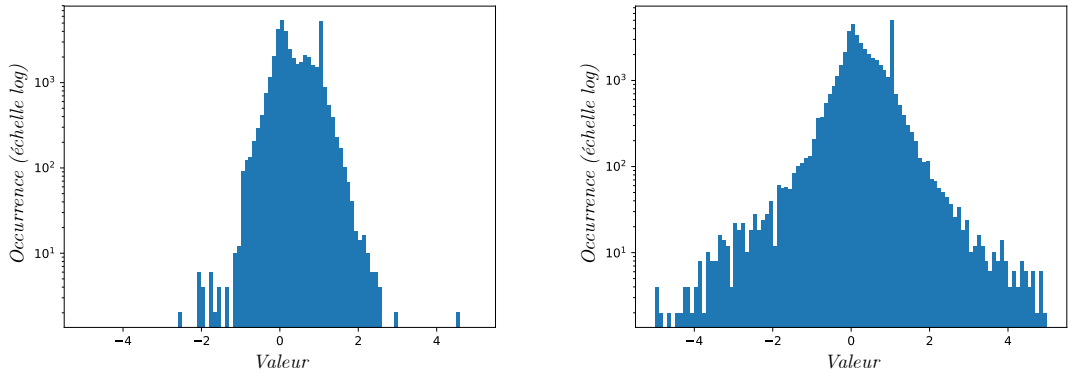
Cette plus forte amplification est visible en observant les coefficients de la matrice de correction $\mathbf{T}$ pour les signaux multimono et multistereo. La figure 5.10 montre l'histogramme des valeurs des coefficients de la matrice $\mathbf{T}$ obtenu par le multimono (5.10(a)) et multistéréo (5.10(b)) sur l'entièreté de l'échantillon *Voices-Kids*. Les artefacts étant présents dans la partie haute du spectre, seules les valeurs des matrices des deux dernières bandes ont été utilisées, là où la fréquence centrale de la bande est supérieure à 6 kHz. Pour la matrice de correction du signal multimono, la moyenne est de $0{,}39$ avec un écart-type de $0{,}5$ avec des valeurs minimale et maximale $-2{,}50$ à $5{,}29$, alors que pour la matrice de correction du signal multistéréo, la moyenne est de $0{,}31$, un écart-type de $0{,}75$ avec des valeurs minimale et maximale $-8{,}89$ à $7{,}30$.

Les moyennes des deux distribution sont similaires, cependant l'étalement de la distribution est plus important pour les valeurs de la matrice de correction issue du signal multistéréo. Cela suggère une modification plus conséquente des composantes et de la scène sonore.

Lors du calcul des histogrammes, pour ne pas ajouter de biais au calcul, seule la matrice triangulaire inférieure a été prise en compte, la matrice triangulaire supérieure étant égale à 0. De plus, l'énergie de W étant normalisée, le coefficient $\mathbf{T}_j[0, 0] = 1$ a été retiré. Sur les histogrammes un pic peut être observé autour de la valeur 1, ce pic correspond au cas où la matrice de correction utilise une composante sans modifier son énergie.

D'après un ensemble d'écoutes informelles suivies de notre analyse sur les coefficients de la matrice $\mathbf{T}$, il semblerait que la méthode par post-traitement ne permet pas d'améliorer la qualité sur les méthodes multistéréo, telle que l'approche de Opus *family mapping 3*. Des changements doivent être intégrés dans le calcul de la matrice de correction $\mathbf{T}$ pour s'adapter à la manière dont les codecs stéréo traitent le signal. Une modification simple consisterait à ajouter une limitation des coefficients de $\mathbf{T}$ pour éviter une trop forte amplification. Cette limitation réduirait du même coup la capacité de la matrice de correction à restaurer l'image spatiale. La restauration de l'image spatiale ne serait donc que partielle. Il est cependant possible que cette restauration partielle n'ait en pratique que peu d'impact sur la qualité, l'acuité du système auditif étant moins précise en hautes fréquences.

De plus, nos expérimentations faites avec l'approche multistéréo suggèrent qu'à partir d'une certaine quantité de dégradation ajoutée au signal codé, la spatialisation du signal ne peut plus être restaurée par le post-traitement. Il serait intéressant d'essayer de définir les conditions et les seuils à partir desquels le post-traitement ne permet plus la restauration de la spatialisation du signal.



(a) Multimono (b) Multistéréo

FIGURE 5.10 – Histogramme des coefficients des matrices de corrections pour la bande haute de l'échantillon *Voices-Kids*.

Malheureusement, du fait du grand nombre de facteurs à prendre en compte, cette exploration n'a pas pu être menée dans le cadre de cette thèse.

5.5 Résumé et perspectives

Dans ce chapitre, nous avons présenté un post-traitement qui permet réduire les dégradations spatiales d'un signal après codage multimono. Pour ce post-traitement, nous voulions proposer une méthode pour améliorer le multimono sans changer le fonctionnement interne de la méthode. Ce post-traitement se base sur l'utilisation de l'image spatiale du signal et du signal codé pour élaborer une matrice de correction. Cette matrice de correction est appliquée aux signaux ambisoniques décodés pour reproduire une image spatiale identique à l'image spatiale d'origine. Les images spatiales sont estimées par le biais des matrices de covariance du signal ambisonique. La matrice de correction est calculée à partir des matrices de covariance du signal original et du signal codé grâce à la factorisation de Cholesky. La matrice de correction doit être calculée dans le codeur puis transmise au décodeur qui l'applique aux signaux décodés. Pour pouvoir calculer la matrice de correction, le codeur doit procéder à un décodage local des composantes ambisoniques. Une architecture alternative pour la méthode est présentée dans le chapitre 6. Cette architecture permet d'éviter le décodage local en faisant une quantification directe des matrices de covariance. Cette méthode de codage par post-traitement a donné lieu à une publication en conférence [Mahé *et al.*, 2021] et au dépôt d'un brevet [Mahé *et al.*, 2019].

Dans la suite de ce chapitre, nous avons mené des tests subjectifs pour évaluer les performances de notre post-traitement. Pour ces tests, l'objectif a été de comparer les échantillons avec et sans post-traitement et de mesurer l'amélioration apportée par ce dernier. Pour vérifier que ce traitement fonctionne indifféremment du codec et débit utilisés, le même test a été conduit avec 2 codecs et 2 débits globaux différents. Les résultats de ces tests ont montré que notre post-traitement apportait en moyenne une amélioration de qualité, notamment grâce à une meilleure spatialisation.

Une analyse *ANOVA* a montré que le contenu de l'échantillon avait une réelle influence sur l'amélioration de la qualité apportée par notre méthode, ce qui suggère que selon le type de signaux, le post-traitement n'apporte pas la même quantité d'amélioration. De nouveaux paramétrages ou mécanismes auraient besoin d'être ajoutés à notre méthode pour garantir une amélioration quel que soit le contenu codé.

Dans la dernière partie de ce chapitre, nous avons effectué une expérimentation avec des signaux codés avec une approche multistéréo. Cette expérimentation a montré que les hypothèses faites en bande haute par les codecs cœur stéréo produisaient des images spatiales trop éloignées par rapport aux images spatiales d'origine. Pour être utilisable avec des codecs cœur stéréo, le post-traitement a donc besoin d'un raffinement. Une première piste pourrait être de limiter des gains de la matrice de correction pour les bandes les plus hautes. De plus, cette expérimentation suggère qu'à partir un seuil de dégradation, notre post-traitement n'est plus en mesure de reproduire une image spatiale. Il serait intéressant d'étudier ce seuil pour délimiter la zone de fonctionnement pour le post-traitement.

CHAPITRE 6

Codage ambisonique paramétrique par restauration de l'image spatiale

Sommaire du chapitre

Dans le chapitre 5, nous avons étudié l'utilisation de l'image spatiale pour mettre au point un post-traitement de restauration de l'image spatiale. Dans ce chapitre, nous nous sommes intéressés à une méthode de codage paramétrique pour l'ambisonique en utilisant l'image spatiale.

Contrairement aux méthodes de codage de type *waveform matching*, les méthodes paramétriques ne cherchent pas à reconstruire le signal d'entrée à l'identique. Ces méthodes cherchent à extraire les caractéristiques prédominantes du signal d'origine dans le but de recréer un signal contenant les mêmes caractéristiques principales. Cette stratégie permet d'obtenir une meilleure qualité subjective du signal à bas débit.

Pour le domaine de l'audio multicanal, les méthodes paramétriques ne vont transmettre qu'un nombre réduit de signaux, en envoyant un sous-ensemble ou une recombinaison des signaux d'entrée. Pour permettre de recréer la scène spatiale originale, un ensemble de paramètres de reconstruction sont extraits du signal d'origine. Ils sont ensuite transmis au décodeur en tant que métadonnées. Cet ensemble de paramètres permet une représentation plus compacte de l'information de la scène et nécessite une moins grande quantité de données que le signal qu'ils modélisent. C'est pour cela que les méthodes paramétriques se retrouvent plus efficaces pour coder les signaux multicanaux à bas débit. Cependant, pour les débits plus importants, les simplifications faites par la modélisation de la source ne permettent pas d'obtenir une reconstruction transparente du signal d'origine. Cela se traduit par un gain important de la qualité sur une certaine gamme de débits puis, à partir d'un seuil la qualité atteint un plateau de qualité. Pour chaque méthode de codage, il est possible de tracer une courbe représentant la qualité audio en fonction du débit. Selon les

caractéristiques et la richesse du modèle, le plateau peut-être à un niveau de qualité plus ou moins élevé. Les modèles les plus complets pouvant permettre d'obtenir une qualité asymptotique plus élevée. Cependant, plus le modèle sera gros, moins les performances dans le très bas débit seront bonnes.

Dans ce chapitre, nous proposons une méthode de codage paramétrique bas débit. Cette méthode permet de coder un signal FOA à partir d'un signal mono et d'utiliser l'image spatiale pour recréer la spatialisation du signal FOA d'origine. Notre méthode vise une gamme de débit allant de 48 à 64 kbit/s. Notre idée est d'utiliser la composante W comme unique signal audio transmis. L'hypothèse a été faite que cette composante omnidirectionnelle, possède toute l'information fréquentielle de la scène sonore. L'information spatiale sera quant à elle représentée par les matrices de covariance du signal origine. Ces images spatiales seront transmises à intervalles réguliers au décodeur sous forme de métadonnées. À partir de cette information spatiale et de la composante W, le décodeur va recréer l'ensemble de composantes ambisoniques. L'opération de passer d'un nombre réduit de signaux à un nombre plus important est appelé *upmix*. Pour projeter la composante omnidirectionnelle reçue dans l'ensemble des 4 composantes, notre *upmix* repose sur une méthode similaire au calcul de la matrice de correction présentée à la section 5.1.

6.1 Présentation de la méthode

6.1.1 Calcul de la correction spatiale dans le cadre de l'*upmix*

À la section 5.1.2, nous avons présenté la cartographie de la puissance d'un signal, aussi appelée image spatiale. Pour une trame audio t donnée, une image spatiale peut être représentée par la matrice de covariance $\mathbf{C}$ du signal ambisonique. Cette matrice $\mathbf{C}$ peut être calculée sur signal plein bande ou par sous-bande. Pour un signal FOA, la matrice $\mathbf{C}$ est de taille 4×4. À partir de la matrice de covariance du signal original $\mathbf{C}$, telle que définie dans l'équation (5.3), et du signal décodé $\tilde{\mathbf{C}}$, une matrice de correction $\mathbf{T}$ peut être calculée pour restaurer l'image spatiale d'origine avec pour objectif :

$$Carto(\mathbf{T}\tilde{\mathbf{B}}) \equiv Carto(\mathbf{B}) \tag{6.1}$$

Comme présenté à la section 5.1.2, cet objectif peut être reformulé comme :

$$\mathbf{T}\tilde{\mathbf{C}}\mathbf{T}^t = \mathbf{C} \tag{6.2}$$

où les matrices $\mathbf{B}$ et $\tilde{\mathbf{B}}$ sont respectivement les matrices contenant les signaux ambisoniques d'entrée et de sortie d'une trame de taille L, et $\mathbf{C}$ et $\tilde{\mathbf{C}}$ les matrices de covariance des signaux d'origine et décodés. Grâce à l'équation (5.11), une solution pour la matrice $\mathbf{T}$ peut être trouvée comme :

$$\mathbf{T} = \mathbf{L}\tilde{\mathbf{L}}^{-1} \tag{6.3}$$

où $\mathbf{L}$ et $\tilde{\mathbf{L}}$ sont les matrices triangulaires obtenues par décomposition de Cholesky pour la matrice de covariance du signal d'origine $\mathbf{C}$ et du signal codé $\tilde{\mathbf{C}}$. L'opérateur $(.)^{-1}$ représente l'inverse de la matrice.

La matrice de correction $\mathbf{T}$ recombine les composantes $\tilde{\mathbf{B}}$ pour obtenir une image spatiale identique à la scène d'origine. Dans la méthode proposée, seule la composante W est codée puis transmise, la matrice $\tilde{\mathbf{B}}$ n'est composée que d'un seul vecteur non nul. Par conséquent, la matrice $\tilde{\mathbf{C}}$ ne possède qu'une valeur non nulle, ce qui ne permet pas de trouver une solution satisfaisante pour la matrice $\mathbf{T}$. Le calcul de la correction de l'image spatiale nécessite des modifications pour pouvoir être utilisée dans le contexte d'une méthode d'augmentation de canaux. Une matrice de correction $\mathbf{T}$ calculée avec uniquement une composante omnidirectionnelle ne pourra faire qu'une redistribution d'un seul signal sur les 4 composantes FOA, ce qui revient à une simple spatialisation du signal dans une direction donnée.

Pour remédier à ce problème, un filtre décorrélateur est appliqué à la composante W décodée pour recréer un signal ambisonique proche d'un champ diffus. Dans ce nouveau signal ambisonique, la matrice $\tilde{\mathbf{B}}$ possède 4 vecteurs non nuls. Pour un signal ambisonique diffus idéal, la matrice de covariance $\tilde{\mathbf{C}}$ serait une matrice diagonale. De plus, les coefficients des composantes X, Y, Z seront égaux. En pratique, les filtres décorrélateurs sont ajustés manuellement pour éviter l'apparition de certains artefacts et effets trop artificiels de la décorrélation, comme des effets métalliques... Cet ajustement produit une décorrélation partielle du signal. Les coefficients hors de la diagonale de $\tilde{\mathbf{C}}$ ne sont donc pas exactement nuls. Grâce à cette décorrélation, un plus grand nombre de degrés de liberté est ajouté pour trouver une solution pour la matrice $\mathbf{T}$. Cette nouvelle matrice $\mathbf{T}$ peut cette fois recréer l'image spatiale d'origine.

De nombreuses méthodes existent pour calculer des filtres de décorrélation [Bouéri et Kyriakakis, 2004, Zotter *et al.*, 2011]. La plupart des méthodes s'intéressent à création de signaux de décorrélation à partir d'un signal mono. Cependant, dans le cas ambisonique, cette décorrélation doit répondre à une certaine structure pour garantir la création d'un champ diffus. Dans [Zotter *et al.*, 2014], une solution d'élargissement des sources sonores par dispersion des fréquences, basée sur l'estimation de la DOA, est présentée. Cette méthode de décorrélation, bien qu'intéressante, a dû être écartée car elle nécessite un signal ambisonique "complet", c'est-à-dire avec 4 composantes, pour fonctionner.

Dans notre méthode, le signal ambisonique "complet" doit être produit à partir d'une unique composante. Les filtres ambisoniques de décorrélation élaborés pour le *framework* Ambisonic Toolkit (ATK) [Lossius et Anderson, 2014] permettent de résoudre ce problème. Ces filtres permettent de transformer un signal omnidirectionnel en un champ diffus FOA. Plusieurs jeux de filtres sont mis à disposition, chacun des jeux permettant une décorrélation du signal plus ou moins importante. Pour chaque jeu de filtres, plusieurs réponses impultionnelles sont disponibles chacune avec une taille allant de 512 à 16384 échantillons. La documentation du *framework* ne donne pas d'information quant à l'influence de la longueur des réponses impulsionnelles sur la qualité audio. Pour nos expérimentations, nous avons utilisé une réponse impulsionnelle de 1024 échantillons. La convolution de W avec cette réponse impulsionnelle ajoute un retard d'environ 20 ms au signal. Ce retard est trop important pour être utilisé en condition réelle pour un codec conversationnel, mais dans le cadre expérimental cela nous a permis de mesurer la qualité maximale de notre méthode. Une optimisation ultérieurs des filtres décorrélateurs pourrait être faite pour limiter ce retard.

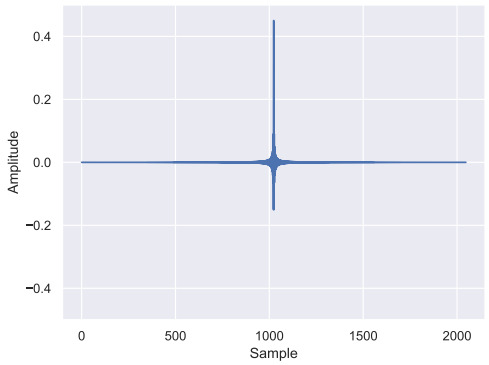
Les filtres sont publics et leurs utilisations sont sous licence *Creative Commons*, cependant, la méthode d'élaboration et l'algorithme permettant de générer ces jeux de filtres ne sont pas publics. Une analyse de ces filtres a été faite pour comprendre leur conception et leur fonctionnement.

D'après nos observations, les filtres fonctionnent ainsi :

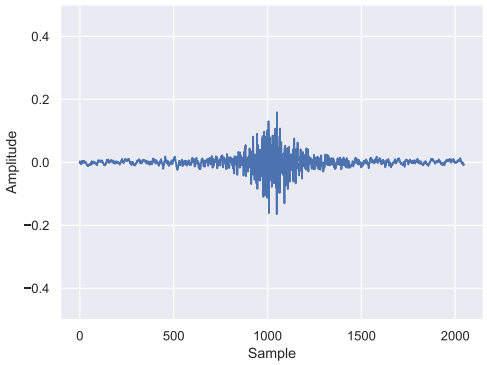
— Le filtre pour la composante W est un sinus cardinal fenêtré. Ce sinus est centré dans la trame pour être synchronisé par rapport aux autres filtres.

— Les filtres pour les composantes X, Y, Z sont des filtres passe-tout avec un retard de phase différent par chaque bande d'octaves.

Sur la figure 6.1, les deux figures supérieures montrent la réponse impulsionnelle des filtres pour les composantes W (figure 6.1(a)) et X (figure 6.1(b)). La figure 6.1(c), montre quand à elle la phase de la transformée de Fourier de la réponse impulsionnelle du filtre de la composante X. Les différents retards de phase par sous-bandes sont visibles. Malgré ces retards de phase localisés, la phase suit une pente globalement proche d'une phase linéaire. Ce retard limité la composante W et les autres composantes doit permettre de garder une synchronisation globale entre les composantes et ne pas altérer trop fortement la base ambisonique en ajoutant un retard trop important à l'une des composantes par rapport aux autres.



(a) Réponse impultionnelle du filtre pour la composante W.



(b) Réponse impultionnelle du filtre pour la composante X.

(c) Phase de la transformée de Fourier de la réponse impultionnelle du filtre pour la composante X.

FIGURE 6.1 – Analyse des filtres décorrélateurs ambisoniques du Framework Ambisonic Toolkit.

6.1.2 Description détaillée du codeur

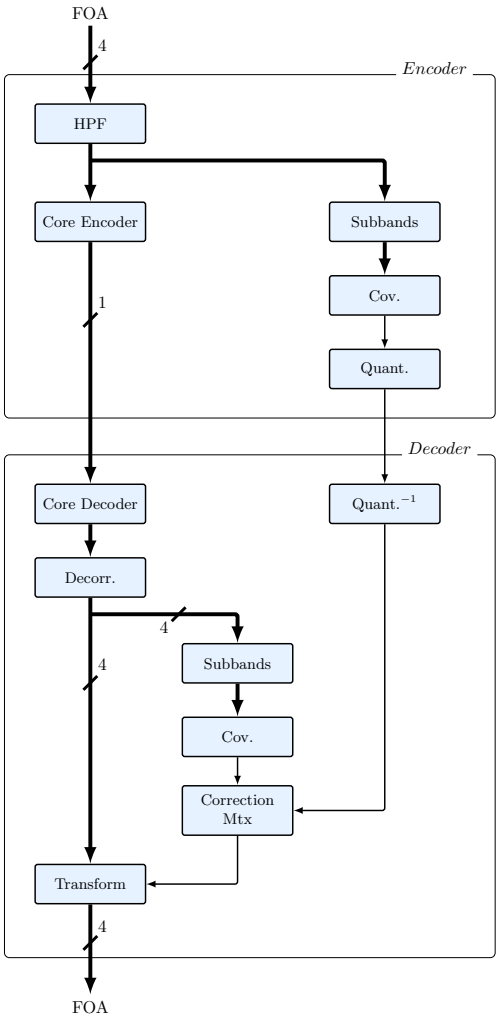


FIGURE 6.2 – Codec complet de la méthode de codage FOA par *upmix*.

La figure 6.2 présente une vue d'ensemble du fonctionnement des différents modules de la méthode de codage par *upmix*. Les traits gras représentent les flux de signaux, les traits fins représentent les transferts de métadonnées. Pour les signaux, l'indication du nombre de composantes est indiquée avec les barres obliques. Dans cette méthode, le codec cœur travaille sur des trames audio de 20 ms en *Full-Band* (FB) avec des signaux échantillonnés à 48 kHz. La taille de chaque trame est de $L = 960$ échantillons. Comme pour les méthodes précédentes, un filtrage passe-haut de fréquence de coupure 20 Hz est appliqué à chaque composante FOA dans le but de supprimer tout biais dans le calcul des matrices de covariance.

Pour déterminer l'information spatiale, le signal est découpé en trame d'analyse avec un recouvrement de 50 % entre deux trames consécutives. Contrairement aux trames audio, ces trames d'analyse ont une durée de 200 ms. Elle est composée des dernières 200 ms du signal. Les 4 composantes ambisoniques sont découpées en J sous bandes selon l'échelle perceptive de Mel [Umesh *et al.*, 1999]. Les filtres ont été conçus dans le domaine fréquentiel puis convertis dans le domaine

temporel pour obtenir les réponses impulsionnelles. Le calcul de l'échelle de Mel est détaillé à l'annexe C. Ces réponses impulsionnelles sont convoluées avec les composantes. Le domaine fréquentiel permet un prototypage facilité des filtres mais crée des filtres avec des retards importants de l'ordre de $10 - 20$ ms. La conception de filtres à faible retard serait nécessaire pour une implémentation conversationnelle de la méthode.

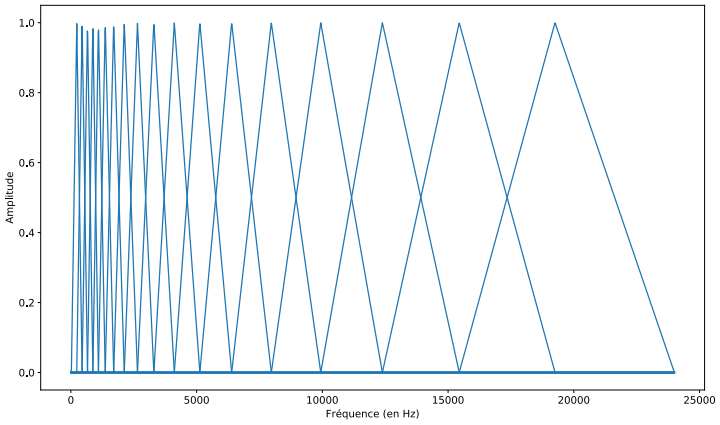


FIGURE 6.3 – Réponse en fréquence des filtres composant le banc de filtres.

Selon la quantité de débit alloué pour les métadonnées, les sous-bandes de Mel J peuvent être plus ou moins importantes pour utiliser la totalité du débit. Dans notre méthode, nous avons utilisé un total de 18 sous-bandes. La figure 6.3 montre la réponse en fréquence des différents filtres qui compose le banc de filtres utilisé pour le découpage en sous-bandes des composantes ambisoniques. Le choix d'utiliser le découpage en bandes de Mel plutôt que les bandes de Bark a été fait pour permettre une plus grande flexibilité, là où le découpage en bande de Bark ne permettait qu'un nombre limité de possibilités. De plus, contrairement aux bandes de Bark, entre deux bandes de Mel consécutives, un recouvrement fréquentiel est présent. Ce recouvrement permet une transition plus progressive entre les bandes et évite ainsi des changements trop importants en passant d'une bande à l'autre lors du calcul des matrices de covariances.

Pour chaque sous-bande j, une matrice de covariance $\mathbf{C}_j$ est calculée pour la trame d'analyse courante de 200 ms. La taille des trames d'analyse a été choisie pour permettre d'avoir une bonne stabilité de l'image spatiale, ce qui permet de limiter les perturbations de l'image spatiale par des éléments ponctuels de l'arrière-plan sonore. De plus, cela limite la quantité de métadonnées à transmettre. Pour chaque sous-bande, une matrice est transmise toutes les 100 ms. Cette taille peut paraître importante, cependant, la trame d'analyse a pour unique but de représenter l'image spatiale d'une scène sonore. Dans notre méthode, l'hypothèse est faite que dans une scène sonore, la vitesse de déplacement des sources et les modifications de la scène sont relativement lentes et progressives par rapport au signal audio. Dans le cadre de cette hypothèse, l'utilisation de trames d'analyse plus longues que les trames audio paraît cohérente.

Le retard algorithmique d'un codec conversationnel doit être relativement faible. À titre d'exemple, pour le codec EVS, ce retard a été fixé à 32 ms [Dietz *et al.*, 2015] pour les signaux échantillonnés à 32 et 48 kHz. Pour conserver une retard comparable aux spécifications actuelles, il n'est pas possible pour notre méthode d'appliquer sur la trame t la correction spatiale calculée

sur cette même trame car cela ajouterait un retard de 100 ms au codec. Pour que notre méthode reste dans un niveau de retard acceptable, l'image spatiale calculée sur la trame $t-1$ est appliquée sur la trame t. Contrairement à la méthode par post-traitement, le signal que la méthode spatialise est un signal mono. Il est donc tout à fait possible de le spatialiser avec n'importe quelle image spatiale sans générer d'artefacts spatiaux. Les sources seront simplement spatialisées à l'endroit où elles étaient présentes à la trame précédente. Cette différence de position peut être qualifiée de retard spatial. Cette technique permet donc de ne pas ajouter de latence temporelle au système, mais uniquement un retard spatial.

Pour un exemple tel que l'échantillon *Noise* (décrit à l'annexe A), où un bruit rose effectue une rotation autour de l'auditeur en 10 secondes, un retard spatial de l'ordre de 100 ms revient à un décalage de la source de $3{,}6°$. Cette différence d'angle est inférieure au Mimimum Audible Angle (MAA), tel que présenté dans [Daniel, 2011]. Pour un échantillon tel que *Noise*, cette modification de position sera donc imperceptible pour les auditeurs. Cependant, pour des scènes sonores avec des déplacements plus rapides, ce retard spatial pourra être perceptible.

Pour limiter ce décalage de position, un fenêtrage asymétrique est appliqué à chaque trame d'analyse pour donner plus d'importance à la fin de la trame. La figure 6.4 montre la forme de cette fenêtre asymétrique. Elle est composée d'une fenêtre exponentielle de 180 ms ainsi qu'une demi-fenêtre de Hann de 20 ms.

Une matrice de covariance $\mathbf{C}_j$ est estimée pour chacune des sous-bandes. Ces matrices sont ensuite quantifiées avant d'être transmises au décodeur.

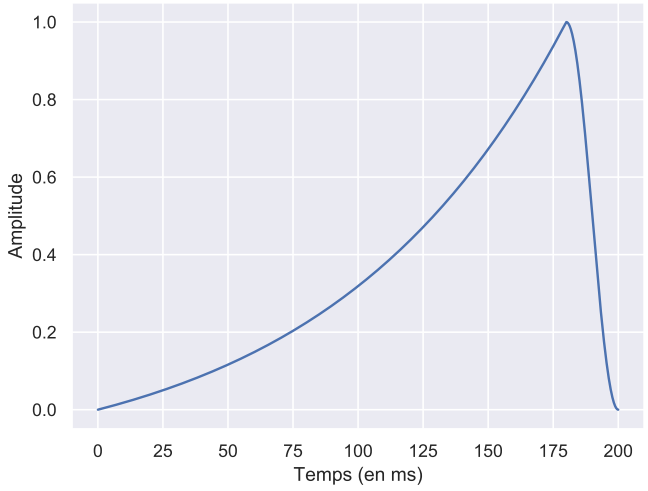


FIGURE 6.4 – Fenêtre d'analyse utilisée pour les trames d'analyse.

6.1.3 Quantification de la matrice de covariance

Dans la méthode par post-traitement, la matrice de correction $\mathbf{T}_j$ est calculée du côté du codeur. Puis cette matrice est quantifiée et transmise au décodeur. Pour déterminer les matrices $\mathbf{T}_j$, la méthode nécessite d'effectuer un décodage local des composantes pour obtenir $\tilde{\mathbf{C}}_j$. Ce décodage

local ajoute un coût important en terme de complexité pour la méthode.

Pour éviter ce décodage local, notre méthode par *upmix* quantifie directement les matrices de covariance $\mathbf{C}_j$. Le décodeur se charge de calculer les matrices de covariance $\tilde{\mathbf{C}}$ du signal décodé. Puis, pour chacune des sous-bandes, les deux matrices $\mathbf{C}_j$ et $\tilde{\mathbf{C}}_j$ vont être utilisées par le décodeur pour déterminer les matrices de correction $\mathbf{T}_j$.

Pour la quantification des matrices de covariance, une simple quantification scalaire de chacun des coefficients de la matrice n'est pas suffisante. Cette quantification ne permet pas de préserver la structure et les propriétés de la matrice de covariance. L'altération de la structure de la matrice est problématique pour le calcul de la matrice de correction, notamment lors de la décomposition de la matrice par la méthode Cholesky. Pour être utilisée, la décomposition de Choleksky a besoin que la matrice traitée soit symétrique et définie positive. Pour préserver les propriétés de la matrice de covariance, la quantification mise en place dans notre méthode a été inspirée de la quantification de la matrice de vecteurs propres dans la méthode de codage par rematriçage dynamique à la section 4.2.3.

Une décomposition en éléments propres de la matrice de covariance $\mathbf{C}$ est effectuée pour obtenir la matrice des vecteurs propres $\mathbf{Q}$ et une matrice diagonale des valeurs propres $\mathbf{\Lambda}$, telle que :

$$\mathbf{C} = \mathbf{Q}\Lambda\mathbf{Q}^{-1} \tag{6.4}$$

L'indice de sous-bande j a été omis par souci de clarté des équations. La matrice des vecteurs propres $\mathbf{Q}$ est une matrice de rotation ou de réflexion, de dimension 4×4. Si la matrice $\mathbf{Q}$ est une matrice de réflexion $det(\mathbf{Q}) = -1$, alors le signe du dernier vecteur propre est inversé. Une fois cette vérification faite, la matrice $\mathbf{Q}$ est décomposée en un ensemble de 6 angles d'Euler généralisés. Ces angles sont quantifiés avec une quantification scalaire uniforme. Sur les 6 angles, les 3 angles ont comme bornes $[-\frac{\pi}{2}, \frac{\pi}{2}[$ et les 3 autres angles de bornes $[-\pi, \pi[$. Un budget de 6 bits est consacré pour les angles $[-\frac{\pi}{2}, \frac{\pi}{2}[$ et un budget de 7 sont alloué aux angles $[-\pi, \pi[$.

Les 4 coefficients de la matrice des valeurs propres $\mathbf{\Lambda}$ sont quantifiés par une quantification logarithmique utilisant la loi-μ. Pour le calcul du facteur de compression, la variance estimée σ^2 est fixé à 3. Une matrice de covariance étant semi-définie positive : $\forall i,\ \lambda_{ii} \geq 0$. Seule la partie positive doit être quantifiée, ce qui permet d'économiser le bit de signe lors de la quantification. Les 4 valeurs propres sont codées avec un budget de 13 bits par valeurs. La transmission d'une matrice de covariance demande donc un budget de 91 bits : $3 \times (6 + 7) + 4 \times 13$.

Chaque trame de 200 ms est composée de 18 matrices, une matrice par sous-bande. Les trames ayant un recouvrement de 50 %, 10 trames sont transmises par seconde, ce qui donne un débit total pour les métadonnées de 16,38 kbit/s, arrondi par la suite à 16,4 kbit/s.

Cette approche sans décodage local et avec une quantification directe des matrices de covariances peut également être utilisée dans la méthode post-traitement. Cette approche est détaillée à la section 6.2.4.

6.1.4 Description détaillée du décodeur

La composante $\tilde{\mathbf{W}}$ transmise par le codeur est décodée par le codec cœur, puis les filtres de décorrélation ambisonique sont appliqués à $\tilde{\mathbf{W}}$ pour obtenir le signal ambisonique diffus $\tilde{\mathbf{B}}$. Le filtre décorrélateur utilisé est l'un de ceux fournis dans le framework ATK. La taille du filtre est de $L = 1024$ échantillons avec le niveau de décorrélation maximum (niveau 5). Le signal ambisonique diffus $\tilde{\mathbf{B}}$ est ensuite découpé en sous-bandes, selon les mêmes caractéristiques que la découpe en sous-bandes faite dans l'encodeur.

Pour chacune des sous-bandes j, une matrice de covariance est calculée $\tilde{\mathbf{C}}_j$. Les angles d'Euler la matrice de vecteurs propres $\mathbf{Q}$ sont extraits des métadonnées. Puis grâce aux valeurs propres $\mathbf{Q}_j$ et vecteur propre $\mathbf{\Lambda}_j$, la matrice de covariance $\mathbf{C}_j$ est reconstruite pour chacune des sous-bandes. Avec les deux matrices $\mathbf{C}_j$ et $\tilde{\mathbf{C}}_j$, la matrice de correction $\mathbf{T}_j$ est calculée. Ces matrices sont appliquées aux signaux $\tilde{\mathbf{B}}$ pour obtenir un signal ambisonique corrigé pour avoir une image spatiale similaire à l'image spatiale d'origine.

6.2 Évaluation de la méthode

Pour évaluer les performances de notre méthode, un test subjectif a été réalisé selon la méthodologie MUSHRA. Pour ce test, deux autres méthodes de codage ont été ajoutées dans le but de permettre une comparaison de la qualité audio de notre méthode par rapport à l'état de l'art pour le codage FOA. Les deux méthodes utilisées dans ce test sont : le mode ambisonique d'Opus (*Family Mapping 3*) [Skoglund, 2018], et la méthode paramétrique DirAC [Pulkki *et al.*, 2018].

L'approche d'Opus mode ambisonique permet d'avoir une méthode récente pour le codage ambisonique. L'approche DirAC permet d'avoir une méthode paramétrique de référence.

Par ailleurs, il existe une implémentation de référence pour ces deux méthodes, ce qui facilite la comparaison de nouvelles méthodes de codage et la reproductibilité de notre test. Toutefois, dans l'implémentation de référence DirAC, fourmie dans [Pulkki *et al.*, 2018], la quantification des métadonnées est manquante. À la section 6.2.2, nous présenterons en détail la quantification implémentée pour notre test.

6.2.1 Conditions expérimentales

Pour chacune des méthodes, deux débits R_1 et R_2 ont été utilisés. Pour notre méthode ainsi que la méthode DirAC, les débits sélectionnés sont $R_1 = 48$ kbit/s et $R_2 = 64$ kbit/s. La qualité des méthodes non paramétriques à très bas débit et bas débit est connue pour être plus faible que celle des méthodes paramétriques. C'est pour cela que pour la méthode Opus mode ambisonique, les deux débits utilisés sont plus importants avec : $R_1 = 64$ kbit/s et $R_2 = 96$ kbit/s. Ce choix permet une évaluation plus fine des conditions car la qualité entre chacune d'elles est plus proche. De plus, ce choix permet d'éviter d'augmenter artificiellement les scores du test MUSHRA en ajoutant des conditions de mauvaise qualité.

Pour notre méthode ainsi que la méthode DirAC, le codec cœur utilisé est le codec Opus mono. Pour toutes les conditions, le codec Opus mono et le mode ambisonique de Opus sont

Tableau 6.1 – Liste des conditions utilisées lors du test MUSHRA.

Abréviation	**Description**
Hidden Ref	Référence cachée FOA
Low Anchor	Référence FOA filtrée à 3,5 kHz avec une réduction spatiale ($\alpha = 0{,}65$)
Medium Anchor	Référence FOA filtrée à 7 kHz avec une réduction spatiale ($\alpha = 0{,}8$)
Multi-Stereo 64	FOA codé par Opus mode Ambisonique 64 kbit/s
Multi-Stereo 96	FOA codé par Opus mode Ambisonique 96 kbit/s
DirAC 48	FOA codé par la méthode DirAC à 48 kbit/s ($33 + 15$)
DirAC 64	FOA codé par la méthode DirAC à 64 kbit/s ($49 + 15$)
Upmix 48	FOA codé par la méthode proposée à 48 kbit/s ($31{,}6 + 16{,}4$)
Upmix 64	FOA codé par la méthode proposée à 64 kbit/s ($47{,}6 + 16{,}4$)

réglés pour coder les signaux en *Fullband* (FB). Les signaux FOA d'entrée utilisés pour le test sont échantillonnés à 48 kHz.

Ce test est composé de 9 échantillons en FOA représentant une variété de contenus critiques. Ces échantillons sont en partie les mêmes que dans les tests précédents, cette fois-ci dans une version 48 kHz. La description détaillée des échantillons peut être trouvée à l'annexe A. Le nombre d'échantillons a été réduit par rapport au test MUSHRA effectué pour le post-traitement afin de raccourcir la durée du test et ainsi limiter la fatigue auditive des participants liée à un nombre d'écoutes trop important.

Les 2 ancres utilisées pour ce test ont été générées avec la même méthode que pour le test évaluant la méthode PCA, présenté à la section 4.3.1. Les signaux des ancres sont filtrés fréquentiellement respectivement à 3,5 et 7 kHz. Un facteur de réduction spatiale de $\alpha = 0{,}65$ est appliqué pour l'ancre basse et $\alpha = 0{,}8$ pour l'ancre moyenne. L'ensemble des 9 conditions utilisées pour ce test sont résumées dans le tableau 6.1.

Ce test a été conduit sur 12 sujets. Tous les participants sont des professionnels du domaine audio ou des personnes familières avec les tests audio. Aucun participant n'avait de perte auditive connue. Du fait des restrictions sanitaires liées au Covid-19, les tests n'ont pas pu être réalisés dans le laboratoire de test audio d'Orange Labs. Chaque participant a réalisé le test dans un environnement domestique, en utilisant son propre matériel audio. La consigne a été d'utiliser une carte son externe et un casque de type studio. Pour le débit utilisé, les dégradations audio sont relativement importantes pour les conditions codées, l'hétérogénéité des équipements audio utilisés n'a eu qu'un impact limité sur le résultat. Lors de l'analyse des résultats, aucune corrélation entre le matériel utilisé et la notation des conditions n'a pu être établie. La version du codec cœur utilisé par toutes les méthodes paramétriques ainsi que pour le mode ambisonique d'Opus a été le codec Opus (v1.3.1).

6.2.2 Détail de la méthode DirAC utilisée

Comme présenté à la section 3.3.2.2, DirAC est une méthode de codage FOA paramétrique. La méthode repose sur l'estimation de deux caractéristiques de la scène sonore : la direction de

la source principale et le caractère diffus, *diffusness*, de la scène. Cette section se divise en deux parties : dans un premier temps, nous expliquerons le fonctionnement générale de la méthode décrite dans [Pulkki *et al.*, 2018]. Dans un second temps, nous détaillerons la quantification des métadonnées implémentées pour le test MUSHRA.

Implémentation de la méthode DirAC

La méthode découpe le signal ambisonique en trames de 40 ms, d'une taille de $L = 1920$ échantillons à 48 kHz. Un recouvrement de 50 % est fait entre deux trames consécutives.

Pour chaque trame, une transformée de Fourier est calculée pour chacune des composantes ambisoniques. À partir de ces transformées est calculé le vecteur intensité acoustique $\mathbf{I}_t(f)$ pour chaque carreau fréquentiel F. Pour plus d'information sur le vecteur intensité acoustique, le lecteur peut se référer au chapitre 1.2.5. Grâce à l'ensemble de ces directions, il est possible d'estimer la direction de la source principale dans la scène. Dans DirAC, seule la partie active est prise en compte pour l'estimation de la direction de la source principale. Le vecteur intensité active est calculé par :

$$\mathbf{I}_t(f) = \begin{bmatrix} \Re\{W_t(f)X_t^*(f)\} \\ \Re\{W_t(f)Y_t^*(f)\} \\ \Re\{W_t(f)Z_t^*(f)\} \end{bmatrix} \tag{6.5}$$

où $W_t(f)$, $X_t(f)$, $Y_t(f)$, $Z_t(f)$ sont les transformées de Fourier de chaque composante W, X, Y, Z pour la trame t. L'opérateur $(.)^*$ désigne le conjugué complexe et $\Re$ la partie réelle. Il est possible d'estimer la direction prédominante de la trame par :

$$\text{DOA} = \angle\mathbb{E}[-\mathbf{I}] \tag{6.6}$$

avec $\angle$ l'opérateur permettant de donner l'angle (θ, ϕ) du vecteur et $\mathbb{E}[.]$ l'opérateur de l'espérance mathématique.

Dans l'implémentation fournie de DirAC, pour chaque carreau temps-fréquence, la direction prédominante (θ, ϕ) ainsi que son caractère diffus Ψ sont estimés. Transmettre une direction par carreau temps-fréquence demande une très grande quantité de débit. Cependant dans les articles qui décrivent la méthode [Laitinen *et al.*, 2012, Hirvonen *et al.*, 2009], un découpage en sous-bandes ERB est réalisé. Pour chacune des sous-bandes j, à partir de l'ensemble des directions prédominantes (θ, ϕ) une direction moyenne est calculée. Selon les articles, le nombre de sous-bandes ERB peut varier. Dans notre implémentation de DirAC, le même type de découpage a été réalisé. Nous avons fait le choix d'estimer une direction pour 30 sous-bandes. Ce choix se base sur les valeurs utilisées dans l'article [Hirvonen *et al.*, 2009]. En parallèle, la composante W est codée par le codec cœur Opus mono puis transmise au décodeur.

Dans la partie décodeur, les informations de direction (θ, ϕ) et le caractère diffus de chaque sous-bande sont extraits des métadonnées. Puis ces informations de direction sont utilisées pour spatialiser la composante $\tilde{W}$. La spatialisation est réalisée par la méthode VBAP [Pulkki, 1997] sur un ensemble de haut-parleurs virtuels placés sur une sphère autour de l'auditeur. Le champs diffus de la scène est simulée par l'ajout d'une version décorrélée de la composante $\tilde{W}$ sur l'ensemble des

haut-parleurs. Cette largeur de la source est contrôlée par le caractère diffus ainsi que l'énergie de la composante $\tilde{W}$ sur la sous-bande. Dans l'implémentation fournie, le nombre et la position des haut-parleurs virtuels pour la méthode VBAP ne sont pas spécifiés. Nous avons choisi de prendre un ensemble de $L = 14$ haut-parleurs positionnés selon une grille de Lebedev. Le choix de ce nombre de haut-parleurs correspond à la plus petite solution à la quadrature de Lebedev, d'une sphère ayant plus de 8 points.

Les signaux des haut-parleurs sont ensuite projetés dans le domaine ambisonique pour obtenir un signal au format FOA par la méthode du réencodage spatial, telle que :

$$\mathbf{B}(t) = \sum_{l=1}^{L} \mathbf{d}_l s_l(t) \tag{6.7}$$

avec $\mathbf{B}$ le signal ambisonique, $\mathbf{d}_l$ le vecteur d'encodage de la direction du haut-parleur l et s_l le signal audio du $l^{\text{ième}}$ haut-parleur virtuel.

Quantification des métadonnées

Dans la méthode DirAC, les métadonnées à transmettre sont composées de la direction de la source (θ, ϕ) et du caractère diffus pour chacune des 30 sous-bandes.

Pour quantifier la direction de la source, nous avons discrétisé la surface de la sphère. Chaque point de cette surface correspondant à une direction de l'espace par rapport au centre de la sphère. La discrétisation de la sphère est faite par une quadrature de Lebedev de 770 points. Ce nombre de points a été choisi pour avoir un écart angulaire inférieur à $3{,}5°$ entre chaque point. Cet écart angulaire est du même ordre de grandeur que la différence minimum d'angle audible (MAA) pour un être humain [Daniel, 2011].

Pour chaque sous-bande, la direction (θ, ϕ) estimée de la source principale est approximée par le point le plus proche sur la discrétisation de sphère. Les points de la grille de Lebedev sont identifiés par un index allant de $[1, \ldots, 770]$. Le codage de l'index de la direction nécessite un budget de 10 bits pour chacune des sous-bandes.

Le caractère diffus, dans la manière dont il est estimé dans [Pulkki *et al.*, 2018], produit des valeurs comprises entre 0 et 1. Pour chaque bande, la valeur du caractère diffus est codée par une quantification scalaire uniforme sur 10 bits. Le budget total pour les métadonnées dans l'implémentation de DirAC utilisé dans ce test est de 15 kbit/s, soit $25 \times 30 \times (10 + 10)$ bits.

6.2.3 Résultats des tests subjectifs

La figure 6.5 montre les résultats obtenus lors du test MUSHRA en fonction des différentes conditions. Pour chaque condition, la moyenne ainsi que l'intervalle de confiance à 95 % sont indiqués. La comparaison des résultats de la méthode par *upmix* et la méthode multistéréo de Opus montre que notre méthode permet d'obtenir une qualité significativement meilleure que la méthode multistéréo à débit équivalent. La méthode d'*upmix* à 48 kbit/s obtient même des scores équivalents à la méthode multistéréo à 64 kbit/s. Les scores obtenus par la méthode DirAC sont quant à eux inférieurs aux résultats des deux autres méthodes pour les deux débits testés.

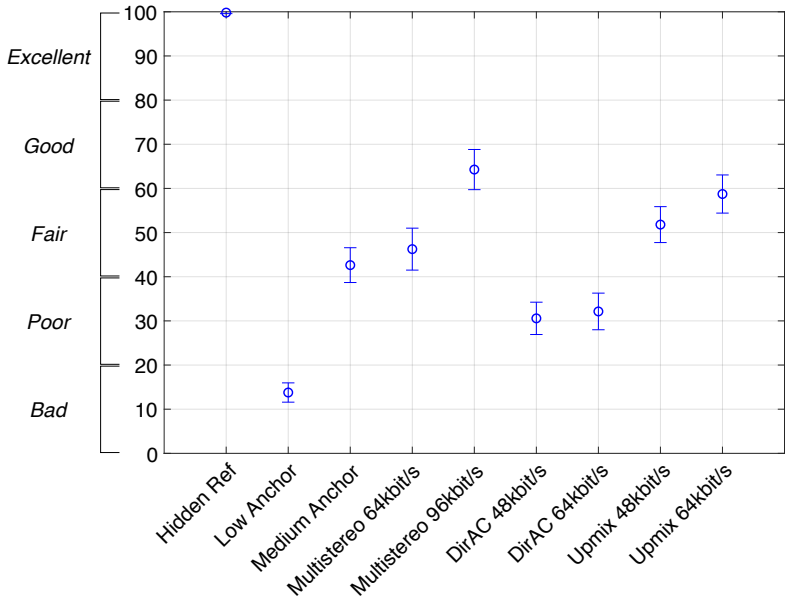


FIGURE 6.5 – Score MUSHRA pour les différentes conditions avec l'intervalle de confiance à 95%.

En observant les scores du test dans leur ensemble, il est possible de remarquer que les résultats pour toutes les conditions sont relativement bas par rapport au test réalisé à la section 5.3.2. Seule une condition arrive à atteindre un score moyen supérieur à 60, score qui peut être considéré comme une condition de qualité acceptable. Il serait possible d'expliquer cette baisse de score simplement par un débit de codage trop bas. Cependant, des scores plus bas sont également visibles sur les conditions qui ne sont pas codées comme les deux ancres. Pour essayer de comprendre cette notation plus sévère des échantillons, il est important de passer en revue un certain nombre de facteurs qui auraient pu expliquer ces scores.

Du point de vue de la méthodologie de test, le logiciel utilisé pour effectuer le test ainsi que les instructions l'accompagnant ont été les mêmes que pour le test à la section 4.3. De même, les ancres ont subi les mêmes détériorations. Du côté des sujets, sur le panel de 12 participants, 8 avaient participé aux deux tests précédents ce qui aurait dû limiter la variabilité de la notation des participants entre les deux tests.

De plus, l'hétérogénéité du matériel et l'utilisation de matériel de moindre qualité lors de l'écoute auraient dû légèrement augmenter les scores (certains artefacts sont masqués) plutôt que l'inverse. L'utilisation d'une salle domestique, non traitée acoustiquement aurait également dû avoir une influence positive sur les scores, le bruit ambiant pouvant masquer en partie les défauts codés des échantillons audio.

Le dernier facteur qui peut expliquer ces scores pourrait être l'utilisation d'un moteur de rendu binaural différent entre les deux tests. Lors du test pour la méthode PCA (section 4.3.1), le moteur de rendu utilisé était *Resonance Audio* [Gorzel *et al.*, 2019]. Alors que pour le test Ref AB pour le post-traitement (section 5.3.1) ainsi que ce test, le moteur de rendu a été changé pour le moteur *SPARTA* [McCormack et Politis, 2019]. De notre point de vue, les filtres binauraux utilisés dans

Resonance Audio apportaient une coloration trop importante au signal codé. Nous avons alors changé le moteur de rendu pour *SPARTA* utilisant, selon nous, des filtres binauraux plus neutres. Ce changement de filtres a pu faire ressortir de manière plus importante certains artefacts de codage, notamment en haute fréquence, ce qui pourrait expliquer une notation plus sévère dans ce test. Comme évoqué à la section 5.3.1, les notions de coloration et de fidélité audio dans le cas des filtres binauraux sont des notions difficiles à mesurer, notamment lors d'utilisation non personnalisée. Il est donc difficile d'aller plus loin dans l'explication de l'impact des filtres sur l'évaluation de la qualité audio perçue.

Cette problématique du choix des filtres binauraux utilisés n'est pas un problème spécifique à notre test et peut être étendue à toute méthode de codage audio ambisonique. Pour résoudre cette problématique dans le futur, deux directions différentes seraient envisageables.

La première consisterait à considérer la méthode de codage et le moteur de rendu comme deux éléments ne pouvant pas être dissociés. Lier ces deux éléments pourraient permettre de tirer parti du fonctionnement de moteur de rendu (approximation, coloration...) pour masquer le bruit de codage ajouté par la méthode de codage. Dans ce cas, la qualité audio de la méthode ne peut pas être étudiée séparément du moteur de rendu utilisé. Ce qui induirait que toute modification des filtres binauraux pourrait remettre en cause la qualité audio de toute la méthode de codage.

La seconde solution consisterait à rendre la méthode de codage complètement indépendante du moteur de rendu. Dans ce cas pour évaluer la qualité d'une méthode, des tests subjectifs devraient être conduits avec de multiples moteurs de rendus, que ce soit avec des filtres binauraux mais également sur haut-parleurs.

Résultats détaillés du test pour la méthode de l'upmix spatial

La figure 6.6 montre des scores obtenus par chaque échantillon selon les différentes méthodes de codage. Pour la même méthode, tous les échantillons n'obtiennent pas le même score. Les méthodes ont plus de difficulté plus importante pour coder certains échantillons que d'autres. Par exemple, *Talks* obtient le meilleur score des échantillons codés par la méthode multistéréo alors qu'il obtient un score moyen quand il est codé par les autres méthodes. L'échantillon *Applause*, quant à lui obtient l'un des scores les plus bas, quel que soit la méthode. En observant les résultats de la méthode multistéréo, il est possible de constater que la différence des résultats entre les échantillons est importante. Certains échantillons ont un rendu de bonne qualité quand d'autres ont un rendu médiocre. Cela semble indiquer une plus forte dépendance au contenu codé par rapport aux autres méthodes. Pour la méthode DirAC, les résultats sont plus homogènes que pour la méthode multistéréo. Les échantillons *Drums*, *Theater*, *Talks*, se détachent des autres et obtiennent de meilleurs scores, quel que soit le débit. Pourtant, il ne semble pas y avoir de points communs entre ces trois échantillons. Certains sont des captations réelles quand d'autres sont synthétiques. L'un contient de la musique, les deux autres sont des signaux de parole, l'un avec une seule source dans des directions bien établies, l'autre avec plusieurs sources concurrentes réparties dans l'espace. Pour le cas de l'*upmix*, mis à part l'échantillon *Applause*, les résultats entre les échantillons sont assez homogènes. Cela indique que les hypothèses faites par notre méthode ne sont pas adaptées au contenu similaire à cet échantillon.

Quand on observe les résultats pour le débit 48 kbit/s et 64 kbit/s, il est possible de remarquer

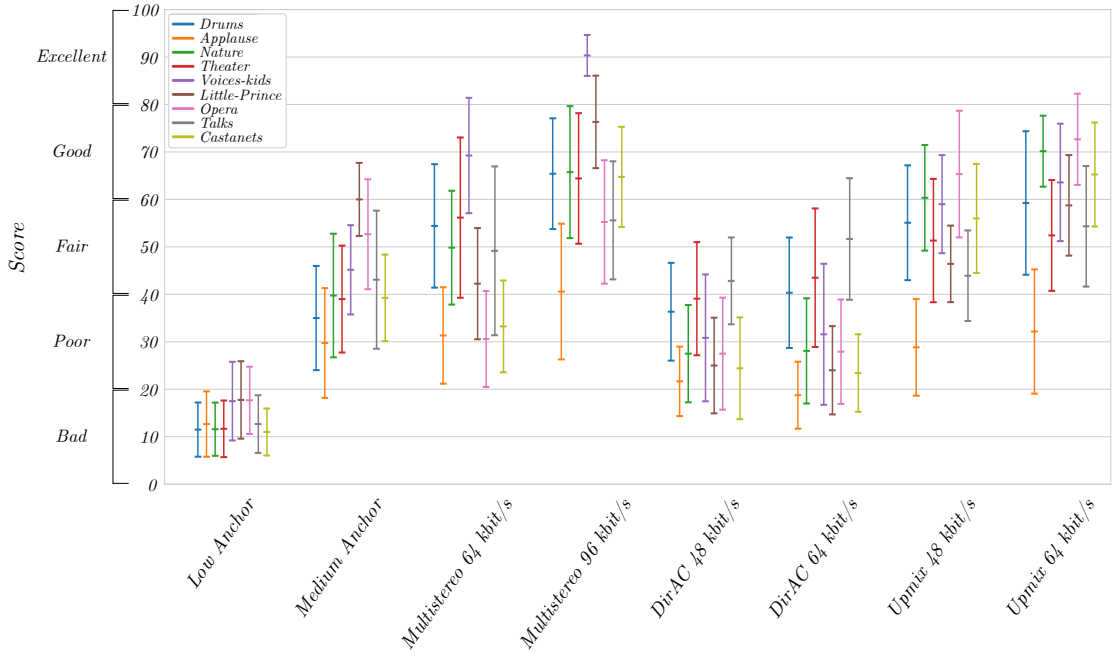


FIGURE 6.6 – Résultats détaillés du test MUSHRA, pour chaque échantillon pour les différentes conditions, avec l'intervalle de confiance à 95 %.

que les relations entre échantillons sont similaires. Par exemple, le score de l'échantillon *Drums* et plus bas que l'échantillon *Nature* et le score de *Theater* sont plus bas que les deux autres échantillons. Cette hiérarchie est présente pour les deux débits, ce qui permet plus facilement de prédire le résultat qu'obtiendra un échantillon pour un autre débit par rapport aux résultats des autres échantillons à ce débit.

Dépendance de la méthode au contenu

Lors de la phase de préparation du test subjectif, des écoutes informelles ont été menées dans le but de fixer les différents paramètres de la méthode : nombre de bandes, fréquence de transmission de l'image spatiale.... Ces écoutes ont pu montrer que différents paramétrages de la méthode permettaient d'obtenir de meilleurs résultats selon le type de contenu présent dans la scène ambisonique codée. Pour des scènes avec de nombreux sons percussifs ou des sources très mobiles, comme l'échantillon *Applause*, une fréquence de transmission plus importante des images spatiales semble améliorer la qualité, alors que le nombre de sous-bandes b utilisé pour le découpage semble avoir peu d'influence sur le rendu global. Pour les scènes avec de nombreuses sources concurrentes, comme l'échantillon *Talks* ou *Nature*, il est préférable de diminuer la fréquence de transmission au profit d'un découpage en sous-bandes plus importantes dans le but de limiter le nombre de sources se chevauchant sur la même sous-bande. La proportion entre le débit alloué au signal W et débit alloué à l'information spatiale peut également être posée. Pour certains échantillons où le contenu du signal W est particulièrement difficile à coder, un choix pourrait être de favoriser le débit alloué au signal au détriment de la métadonnée spatiale. Il serait intéressant d'étudier la question d'un point de vue perceptif : est-il préférable d'avoir une spatialisation gros-

sière d'un signal avec un contenu fréquentiel de meilleure qualité ou une spatialisation fine avec un signal avec un contenu fréquentiel fortement dégradé ? Des écoutes informelles sur un grand nombre d'échantillons nous ont permis de fixer les valeurs finales pour les différents paramètres de notre méthode, ces valeurs permettent un compromis entre tous les points évoqués précédemment.

6.2.4 Comparaison avec la méthode DirAC

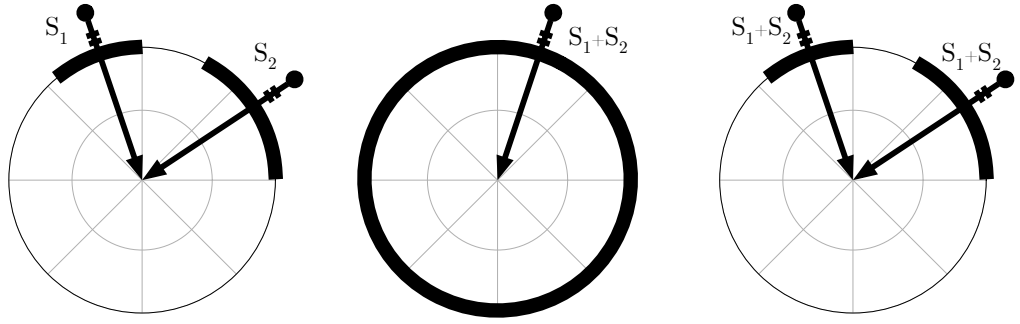
Dans le test que nous avons conduit, les performances de notre méthode ont été évaluées par rapport à la méthode DirAC. Cette méthode a été choisie car elle est une référence dans le codage ambisonique paramétrique. Il serait intéressant d'analyser les similitudes et les différences de fonctionnement pour dresser des parallèles entre les deux méthodes.

Pour les deux méthodes, la composante W est transmise avec une information spatiale dans le but de recréer le signal ambisonique d'origine. Pour chacune, une extraction de l'information spatiale est faite par sous-bande perceptive. Même si l'information spatiale extraite n'est pas de même nature, l'un analyse la scène sonore pour en extraire la source principale, quand l'autre capture une répartition de l'énergie. Cependant des analogies entre les deux méthodes peuvent être faites. Dans le cas d'un signal avec au plus une source dans chaque bande de fréquences, le résultat obtenu par les deux approches est similaire. Pour les signaux contenant plusieurs sources concurrentes dans la même bande, les approximations faites par les deux méthodes produisent des résultats très différents.

Dans la méthode DirAC, l'information spatiale provient d'une analyse de la scène pour déterminer la direction de la source et le caractère diffus. Le fonctionnement de l'analyse donne des contraintes explicites sur le contenu capturé du signal. Pour chacune des sous-bandes, une seule direction est extraite. Cette direction correspond à la direction moyenne de l'énergie. Le champ diffus est assimilé à une énergie repartie de façon homogène dans toutes les directions de l'espace. Si plusieurs sources sont présentes, la direction extraite est la direction moyenne des sources pondérées par l'énergie. Cette direction peut ne pas correspondre à la direction d'une des sources présente mais à un point tout autre de l'espace. De plus, l'écart des directions des sources d'origine par rapport à la direction moyenne contribue à une surévaluation du caractère diffus, ce qui entraine l'ajout d'un fort champ diffus au signal décodé.

Dans notre méthode, l'information spatiale est représentée par la matrice de covariance. Cette manière de représenter l'image spatiale pourrait être perçue au premier abord comme agnostique sur le contenu de la scène sonore mais cette représentation engendre également des approximations implicites sur la scène sonore. Notre méthode recréée un signal ambisonique avec la même répartition de l'énergie que le signal d'original. Cependant, une même répartition de l'énergie peut être produite par des signaux ambisoniques différents. Dans une même sous-bande, si deux sources sont présentes, notre méthode reproduira une répartition d'énergie dans l'espace identique à celle d'origine, pourtant le contenu des composantes n'est pas identique, les deux sources sonores seront mélangées.

La figure 6.7 montre la manière dont les deux méthodes spatialisent un signal contenant deux sources concurrentes S_1 et S_2, provenant de directions différentes. Ces deux sources sont actives dans la même bande de fréquences. La direction de chaque source est matérialisée par une flèche et la largeur de la source par un trait épais. Dans DirAC, le signal produit sera la somme de $S_1 + S_2$



(a) Original (b) DirAC (c) Notre méthode

FIGURE 6.7 – Modélisation du champ sonore pour un signal composé de 2 sources S_1 et S_2 fait par la méthode DirAC et notre méthode. Les flèches représentent les directions des sources, la partie du cercle unitaire en gras représente la largeur de la source.

localisée dans une direction de l'espace. Un champ diffus est ajouté dans toutes les directions de l'espace pour simuler la largeur des sources. Dans notre méthode, les directions des sources ainsi que la largeur des sources sont représentées par la matrice de covariance. Le signal produit sera la somme de $S_1 + S_2$, spatialisé, cette fois, dans la direction de la source S_1 ainsi que de la source S_2. Pour notre méthode, la décorrélation, appliquée avant la spatialisation, permet de produire deux sources distinctes et non une seule, comme dans le cas où les signaux auraient été corrélés. Cette décorrélation permet également de reproduire la largeur des sources.

La stratégie de spatialisation mise en place dans notre méthode devrait apporter une plus grande stabilité par rapport à l'estimation d'une unique source. Cette stratégie évite des erreurs d'estimation de la direction ainsi que des changements brutaux de direction pour les signaux composés de plusieurs sources. De plus, cela ajoute une plus grande cohérence, en terme de direction, pour les sources présentes sur plusieurs bandes. Ces sources restent dans la même direction même pour les bandes où elles ne sont pas la source prédominante.

Du point de vue de l'extraction de l'information spatiale, les deux méthodes ont mis en place un lissage temps-fréquence pour garantir une meilleure cohérence de la scène sonore d'une trame à l'autre. Ce lissage est effectué de manière différente par chacune des méthodes. Le lissage temporel est assuré par un filtrage exponentiel dans le cas de DirAC et par l'utilisation d'une trame d'analyse longue pour notre méthode. Dans le domaine fréquentiel, un lissage est effectué selon l'axe temporel pour DirAC. Dans notre méthode, une certaine continuité est assurée grâce au recouvrement des bandes de Mel.

Une limite des deux méthodes est qu'aucune des deux représentations spatiales ne permet de prendre en compte les effets champs proches [Daniel, 2003]. Les deux représentations assimilent les sources à des ondes planes, les sources sont situées à l'infini. Dans le cas d'une source située à une distance faible par rapport à l'auditeur, cette source produira une onde sphérique. Même si la source devrait pouvoir être correctement capturée par les deux représentations, lors de la spatialisation les méthodes ne possèdent pas de mécanisme pour reproduire l'effet de champ proche. La

source sera donc respatialisée comme une onde plane et l'effet de distance sera gommé.

Architecture alternative de la méthode par post-traitement

La quantification de l'image spatiale, proposé la méthode d'*upmix* à la section 6.1.3 peut être adapté à la méthode de codage par post-traitement. La figure 6.8 présente une architecture alternative du codec par post-traitement. Dans cette architecture, ce n'est pas la matrice de transformation $\mathbf{T}$ qui est transmise de l'encodeur vers le décodeur, mais directement la matrice de covariance originale $\mathbf{C}$. Cela permet d'économiser le coût de calcul et le retard lié au décodage local des composantes ambisoniques dans la partie codeur et le découpage en sous-bande de ces composantes. La même quantification est utilisée pour la méthode d'*upmix*. Une décomposition de la matrice de covariance $\mathbf{C}$ en vecteurs et valeurs propres est effectuée. Les vecteurs propres sont convertis en angles d'Euler généralisés puis quantifiés par une quantification scalaire. Les valeurs propres sont directement quantifiées par une quantification scalaire.

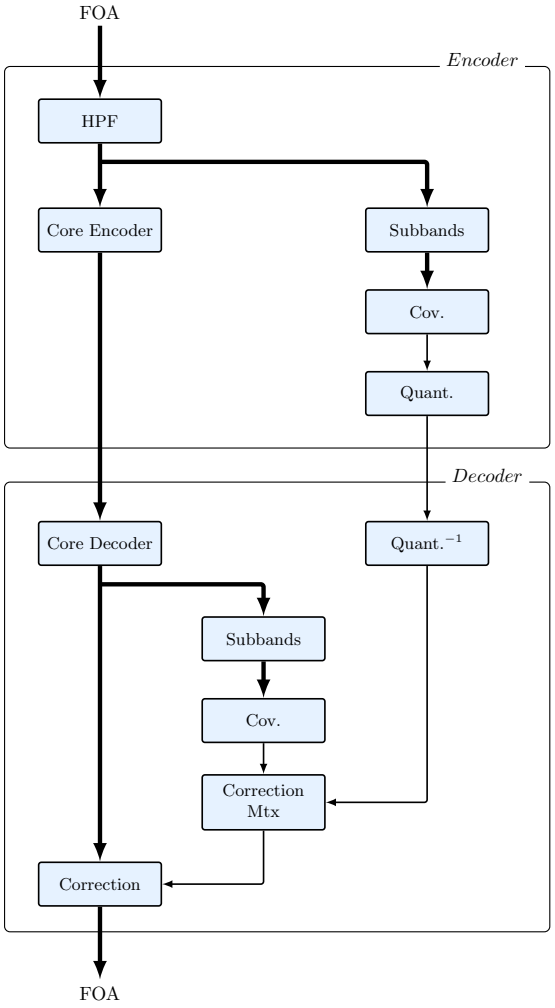


FIGURE 6.8 – Architecture alternative du codage FOA par post-traitement.

6.3 Résumé et perspectives

Dans ce chapitre, nous avons présenté une méthode de codage paramétrique pour l'Ambisonie d'ordre 1. Cette méthode se base sur l'idée qu'il est possible de reconstituer l'ensemble du signal ambisonique à partir de la composante omnidirectionnelle W et de l'information spatiale. L'hypothèse a été faite que la composante W possède toute l'information fréquentielle nécessaire et que l'information spatiale peut être approximée par la répartition de l'énergie dans la scène d'origine. Comme une seule composante a besoin d'être codée par le codec cœur puis transmise au décodeur, cette méthode peut fonctionner à des débits bien moins importants que les méthodes non paramétriques.

Le signal ambisonique d'entrée est découpé selon les sous-bandes de Mel. Puis de manière analogue à la méthode par post-traitement, pour chaque trame, l'image spatiale est capturée par le biais du calcul des matrices de covariance $\mathbf{C}_j$.

Dans notre méthode par *upmix*, une nouvelle quantification a été mise permettant de quantifier directement la matrice de covariance $\mathbf{C}$ et ainsi permettre d'éviter le décodage local des composantes. Cette quantification consiste en une décomposition de la matrice $\mathbf{C}$ en éléments propres.

Dans le décodeur, la composante $\tilde{\mathbf{W}}$ est décodée par le codec cœur. Pour recréer la spatialisation d'origine, des filtres décorrélateurs vont être appliqués à la composante $\tilde{\mathbf{W}}$ pour former un signal ambisonique proche du champ diffus $\tilde{\mathbf{B}}$. À partir de ce signal ambisonique $\tilde{\mathbf{B}}$, une matrice de covariance est calculée. En utilisant la matrice d'origine $\mathbf{C}$ et la matrice du signal diffus, une matrice de correction $\mathbf{T}$ va être calculée. Cette matrice $\mathbf{T}$ est ensuite appliquée au signal $\tilde{\mathbf{B}}$ pour produire le signal ambisonique de sortie.

Dans la suite de ce chapitre, un test MUSHRA a été conduit pour comparer notre méthode par *upmix* à deux autres méthodes de l'état de l'art : le mode ambisonique d'Opus et la méthode paramétrique DirAC. Pour chacune des méthodes, deux débits ont été testés. Les résultats ont montré que notre méthode permet d'obtenir une qualité audio significativement meilleure que la méthode multistéréo à débit équivalent. De plus, notre méthode à 48 kbit/s a obtenu un score de qualité équivalent à celui de la méthode multistéréo à 64 kbit/s. La méthode DirAC a quant à elle eu des résultats inférieurs aux autres méthodes.

Dans la dernière partie du chapitre, une comparaison entre le fonctionnement de la méthode DirAC et notre méthode a été faite. Cette comparaison a permis de confronter les hypothèses faites par chacune des méthodes et d'analyser comment ces hypothèses ont pu se traduire dans la mise en œuvre de traitements relativement similaires. Nous avons vu dans ce chapitre que notre méthode permettait de faire un codage paramétrique à bas débit. Cette méthode a pris le parti de faire un minimum d'hypothèses sur le contenu de la scène (nombre de sources, type de son…). Cependant, des tests d'écoute informels ont montré que l'adaptation des valeurs des paramètres de la méthode en fonction du contenu de la scène sonore permettrait d'obtenir de meilleurs résultats. Notre méthode pourrait être étendue pour la rendre adaptative. Une analyse sommaire de la scène sonore pourrait piloter les paramètres de la méthode afin d'obtenir la meilleure qualité possible. Des expérimentations avec le vecteur intensité et une détection de transitoire dans la composante omnidirectionnelle ont été menées. Malheureusement, l'analyse rudimentaire de la scène ainsi que la difficulté de fixer les seuils appropriés pour chaque type de contenus n'ont pas permis d'intégrer cette amélioration dans notre méthode. Les travaux présentés dans ce chapitre ont donné lieu au

dépôt d'un brevet sur le codage paramétrique de l'ambisonique où les paramètres de l'approche sont pilotés par une analyse du contenu de la scène sonore [Ragot et Mahé, 2020].

CHAPITRE 7

Compression audio par auto-encodeur variationnel

Sommaire du chapitre

Ce chapitre est consacré à la compression audio par réseau de neurones. Notre travail est basé sur les travaux en compression d'image proposés par Ballé et Minnen [Ballé *et al.*, 2017, Minnen *et al.*, 2018]. Dans un codec classique, le but est de transformer les données d'entrée en une représentation nécessitant moins d'information que les données d'origine pour être transmises. L'encodeur, par l'intermédiaire de multiples traitements, doit traiter les données pour obtenir une représentation plus adapté pour être transmise avec un débit moins important. Le décodeur, quant à lui, doit inverser cette transformation. Ces traitements sont généralement des transformations linéaires et élaborées à la main. Par ailleurs, chaque traitement est optimisé comme un bloc indépendant ce qui peut rendre les traitements sous-optimaux une fois mis ensemble. Les approches par réseaux de neurones peuvent permettre d'avoir une approche bout en bout, ce qui permet d'avoir une optimisation globale du traitement. De plus, les réseaux de neurones peuvent apprendre des transformations non linéaires ce qui peut permettre d'obtenir des solutions plus optimisés que les méthodes traditionnelles.

Dans les chapitres précédents, nous avons cherché à étudier des méthodes de codage ambisonique qui utilisent des codecs cœurs mono. Dans ce chapitre, nous nous sommes intéressés à l'élaboration d'un codec mono par réseau de neurones pour remplacer ces codecs cœurs. Ce choix a été fait pour pouvoir donner la possibilité de créer des méthodes plus facilement utilisables pour d'autres méthodes de codage. De plus, ce choix permet de bénéficier des nombreuses bases de données d'enregistrements réels de parole et de bruit mono [Zen *et al.*, 2019, Veaux *et al.*, 2017] pour les entrainements de nos modèles de réseaux de neurones. En comparaison, la taille actuelle des bases de données d'enregistrements ambisoniques réels reste trop petite encore pour permettre l'entrainement de modèle. Le choix pourrait être fait d'utiliser des bases d'enregistrements ambisoniques synthétiques, cependant, ces enregistrements semblent encore trop rudimentaires pour refléter fidèlement une prise de son ambisonique réelle.

Dans la première partie de ce chapitre, nous adapatons les modèles proposés dans les articles [Ballé *et al.*, 2017, Minnen *et al.*, 2018] pour réaliser une méthode de compression audio mono. Nous avons tout d'abord étudié la capacité d'un tel modèle à reproduire un signal audio sans contrainte de débit. Ces expérimentations ont permis de mesurer les caractéristiques (temps d'exécution, taille du réseau...) du modèle en fonction de ses hyperparamètres. Cela a permis d'étudier la faisabilité d'une telle compression dans le domaine audio. Dans un second temps, la contrainte de débit a été ajoutée pour mesurer les performances de compression du modèle. Les expérimentations ont été menées sur un corpus d'enregistrements de parole mono échantillonné à 48 kHz. Par la suite, une extension de ce modèle a été réalisée. Cette extension consiste en l'ajout d'une prédiction de l'activation de l'espace latent. Cette prédiction est faite par le biais de la modélisation de l'espace latent appelé représentation hyper-latente. Une évaluation de ce modèle a montré de meilleures performances de compression par rapport au modèle initial. Les performances de ce modèle avec représentation hyper-latente ont été comparées à deux codecs audio traditionnels MP3 et Opus. Enfin, nous présentons les limites des modèles présentés et des pistes d'amélioration.

7.1 Codage audio par Deep Learning

Il y a plusieurs manières de tirer parti de l'apprentissage automatique pour le codage audio. La première manière est d'utiliser l'apprentissage automatique pour réaliser une analyse du signal d'audio d'entrée afin de piloter le codec (choix du mode de fonctionnement, des paramètres...). Par exemple, dans son module de la détection de voix, ou Voice Activity Detection (VAD), le codec Opus [Valin *et al.*, 2016] a d'abord utilisé des modèles de Markov cachés pour déterminer si le signal d'entrée était un signal de parole ou de musique. Depuis sa version v1.3, un nouveau module de classification à base d'un réseau de neurones récurrent a été implémenté améliorant la qualité de la classification.

Une seconde manière consiste à utiliser l'apprentissage automatique comme un module de synthèse sonore permettant de générer directement le signal audio. Cette synthèse sonore peut être utilisée comme une extension des codecs traditionnels dans le but de corriger les dégradations apportées par la compression. Cette synthèse peut être utilisée pour générer l'ensemble du signal à partir d'une représentation ou d'un ensemble de paramètres.

7.1.1 Amélioration des codecs traditionnels

Correction des altérations de codage

Dans l'article [Biswas et Jia, 2020], les auteurs proposent d'utiliser un réseau antagoniste génératif, ou Generative Adversarial Network (GAN), pour rehausser la qualité audio du signal codé par un codec mono traditionnel. Le modèle présenté est composé de 2 parties. Un premier réseau, appelé générateur, est chargé de corriger les dégradations produites par le codec mono. Un second réseau est là pour permettre l'entrainement du premier réseau. Dans le premier réseau, le signal d'entrée $\mathbf{x}$ est codé puis décodé par un codec mono. Le signal décodé $\hat{\mathbf{x}}$ est ensuite fourni à un réseau de neurones, g, qui devra corriger les altérations du signal pour le rendre plus proche du signal d'origine $\mathbf{x}$, tel que :

$$\min \|\mathbf{x} - g(\hat{\mathbf{x}}, \phi)\| \tag{7.1}$$

avec g la fonction appliquée par le réseau sur le signal $\hat{\mathbf{x}}$ et ϕ les paramètres du réseau.
Le second réseau consiste en un réseau, appelé le discriminateur. Lors de l'entrainement du modèle, des signaux vont être fournis au discriminateur. Ces signaux peuvent être soit un signal non codé, soit un signal restauré. Le discriminateur doit alors déterminer si le signal qui lui a été fourni en entrée est un signal non codé ou un signal restauré. Au cours de l'entrainement, le discriminateur aura une analyse de plus en plus fine des signaux, ce qui va pousser le réseau générateur à produire des signaux de plus en plus convaincants aux yeux du discriminateur. Au fil du temps, les signaux générés seront indifférenciables par rapport à des signaux non-codés, ce qui signifie que les dégradations du signal auront été corrigées. Une fois l'entrainement terminé, le premier réseau est utilisé comme post-traitement pour corriger les dégradations générées par le codec cœur.

Extension de bande audio

Dans l'article [Miron et Davies, 2018], les auteurs proposent de faire une extension de bande audio à l'aide d'un auto-encodeur. À partir d'un signal à bande étroite, allant de 0 à $7{,}5$ kHz, l'auto-encodeur génère une bande haute pour obtenir un signal pleine bande, allant de 0 à $22{,}05$ kHz. Cependant, le signal généré semble posséder une certaine forme de flou fréquentiel. Ce flou fréquentiel rend difficile la génération d'éléments très localisés en fréquence et en temps, ce qui semble limiter l'utilisation de la méthode pour des signaux avec une bande de fréquence plus réduite. Dans l'article [Bosca *et al.*, 2021], le même type de flou a pu être constaté dans le cas de la génération de masque fréquentiel pour de la séparation de sources. Les auteurs proposent de corriger ce défaut en remplaçant l'architecture de l'auto-encodeur par une architecture de type U-Net.

7.1.2 Synthèse du signal par réseaux de neurones

Génération du résidu de prédiction LPC

Pour la synthèse vocale et les applications de *text-to-speech*, le modèle appelé *Wavenet* [Oord *et al.*, 2016], s'est imposé comme un modèle de référence. La particularité de ce modèle est de gé-

nérer le signal échantillon par échantillon. Ce modèle est un modèle récursif, c'est-à-dire qu'il génère l'échantillon $x[t]$ à partir des N échantillons précédents tels que : $x[t] = g(x[t-N], \ldots, x[t-1], \phi)$, avec g le modèle et ϕ les paramètres du modèle. Pour synthétiser le signal, en plus des N échantillons précédents, le réseau prend en entrée un ensemble de paramètres. Selon l'utilisation faite du modèle, ces paramètres peuvent être de nature différente. Par exemple, pour une application de *text-to-speech*, les paramètres pourront être des caractéristiques de la voix cible ou les mots du texte à générer sous la forme de *word embedding*.

Dans l'article [Skoglund et Valin, 2020], les auteurs proposent une méthode de codage de parole utilisant un modèle similaire à *Wavenet*. Leur méthode est basée sur le codage LPC, présenté succinctement à la section 3.1.1. Une analyse LPC est effectuée sur le signal d'entrée pour extraire les paramètres de la voix. Ces paramètres seront utilisés pour piloter le modèle de synthèse vocale. Contrairement au codage LPC classique, dans la méthode proposée le signal d'excitation, aussi appelée résidu de prédiction, n'est pas transmise au décodeur. Ce signal d'excitation est généré dans le décodeur par le modèle *WaveRNN*, une variante du modèle *Wavenet*. Pour créer cette excitation, le modèle *WaveRNN* utilise comme paramètres d'entrée les coefficients LPC et ainsi que les caractéristiques fines extraites du signal d'origine à l'aide d'un second réseau de neurones. Seuls les paramètres LPC et les caractéristiques extraits par le second réseau sont transmis du codeur vers le décodeur, la méthode permet un codage de la parole avec un débit de 1,6 kbit/s avec une qualité proche de celle obtenue par le codec Opus à un débit de 9 kbit/s [Valin et Skoglund, 2019]. Cependant, le filtre LPC est une méthode dédiée au traitement de la parole, ce qui rend l'utilisation du codec difficile pour d'autres types de signaux comme la musique.

Compression de la représentation du signal

Un autre type de méthode basé réseaux de neurones cherche à trouver une représentation du signal nécessitant moins d'informations à transmettre pour décrire le signal. Une première partie du modèle est chargée de trouver une représentation du signal, cette partie est appelée la partie analyse. Une seconde partie est chargée de retrouver le signal d'origine à partir de cette représentation, cette partie est appelée la partie synthèse. Dans le domaine du codage d'image, l'approche [Ballé *et al.*, 2017] propose d'entrainer un auto-encodeur variationnel, ou VAE, pour optimiser cette représentation des données. Cette architecture est bien connue pour les problèmes de réduction de dimensionalité. Un auto-encodeur a une forme de sablier, la taille des couches de neurones dans le réseau diminue jusqu'à arriver au *bottleneck*, là où la représentation est la plus compacte, aussi appelé espace latent, puis la taille des couches augmente pour avoir en sortie la même taille que la couche d'entrée. La partie en amont du *bottleneck* est la partie analyse, la partie en aval est la partie synthèse.

L'architecture du VAE est entrainée pour reproduire le signal d'entrée en passant par le *bottleneck*, qui force le réseau à trouver une nouvelle représentation des données. La partie analyse de l'auto-encodeur est utilisée comme un codeur qui prend le signal d'entrée et le transforme en un espace latent. L'espace latent produit est quantifié et transmis à la partie synthèse qui joue le rôle du décodeur.

Lors de l'entrainement, le réseau cherche à maximiser une fonction de coût composé de deux termes : une contrainte de reconstruction, qui correspond à la qualité du signal reproduit et une

contrainte de débit qui correspond à la quantité de données nécessaires à transmettre pour reconstruire le signal. Dans le domaine de la compression d'image et vidéo, ce type d'approche a montré des résultat permettant de concurrencer les approches de codage de l'état de l'art, comme HEVC et VTM [Cheng *et al.*, 2020, Ladune *et al.*, 2021].

Upmix ambisonique par réseau de neurones

Une approche intéressante de codage paramétrique ambisonique a été proposée [Morgado *et al.*, 2018]. Les auteurs proposent de générer un signal FOA complet à partir d'un signal audio mono et d'un signal vidéo 360°. Le modèle présenté propose d'utiliser une approche multimodale, le réseau de neurones s'appuie à la fois sur le signal audio et la vidéo. Dans une partie du réseau, un auto-encodeur de type U-net est chargé de réaliser une séparation des sources audio principales contenues dans le signal, puis une seconde partie cherche à trouver les corrélations entre les signaux séparés et des portions spatiales de la vidéo. Les signaux sont ensuite spatialisés dans la direction associée à la portion de la vidéo. Les auteurs présentent la méthode comme une manière de créer un son plus immersif pour des vidéos 360° fait avec une captation sonore mono. Mais cette même méthode pourrait être adaptée pour un codage *upmix* ambisonique en remplaçant la vidéo par de l'information spatiale plus compacte.

7.2 Compression mono bout en bout

7.2.1 Fonctionnement de l'approche

Le but de tout système de compression est de trouver une représentation du signal qui soit la plus adaptée possible pour être quantifiée et transmise avec le débit le plus réduit possible. La représentation essaye de capturer la structure du signal pour diminuer la redondance dans les données et les rendre le plus homogène possible. Généralement, pour les codecs avec perte, c'est lors de ce changement de représentation que les simplifications sont réalisées. Une fois cette représentation obtenue, elle est quantifiée. C'est-à-dire que les valeurs initiales de cette représentation sont approximées par la valeur la plus proche dans l'espace des valeurs quantifiées. Un exemple de quantification simple peut être de transformer de valeurs dans $\mathbb{R}$ vers $\mathbb{Z}$, par l'opération d'arrondi $\lfloor . \rceil$ à l'entier le plus proche. La représentation quantifiée est ensuite convertie en un train binaire, ou *bitstream*, par un codeur entropique, comme le codage Huffman. Le codage entropique, contrairement à la quantification, est sans perte.

Pour un codec audio par transformée type, un ensemble de traitements est appliqué pour transformer le signal d'entrée x vers un espace où les éléments seraient plus homogènes. De manière générale dans les codecs par transformée modernes, des prétraitements sont appliqués au signal d'entrée, comme une préaccentuation (*pre-emphasis*) du signal, puis pour chaque trame, une transformée (MDCT, QMF ...) est calculée pour obtenir le signal $\mathbf{x}$. Le signal est ensuite découpé en sous-bandes. Pour chacune de ces sous-bandes, plusieurs normalisations de l'énergie peuvent être appliquées pour rendre le signal de chaque bande plus homogène. Cette représentation $\mathbf{y}$ est ensuite quantifiée, ce qui produit une représentation approximée $\hat{\mathbf{y}}$. Cette représentation est ensuite transmise sous forme d'un train binaire. Dans la partie du décodeur, le train binaire est décodé

pour récupérer la représentation $\hat{\mathbf{y}}$, puis, le traitement inverse de celui réalisé par l'encodeur est appliqué à $\hat{\mathbf{y}}$ pour retrouver une approximation du signal d'origine $\hat{\mathbf{x}}$.

Dans l'approche [Ballé *et al.*, 2017], les auteurs proposent d'utiliser un réseau de neurones pour trouver une représentation des données d'entrée plus efficace que celle utilisée par les approches traditionnelles, c'est-à-dire une représentation qui permet, à qualité de reconstruction égale, un débit moins important. Le modèle proposé est un modèle bout en bout (*end-to-end*) avec pour objectif de proposer une architecture relativement légère et ne nécessitant pas un temps de calcul trop important. L'idée est d'utiliser un auto-encodeur variationnel (VAE) pour se substituer à la fois au codeur et au décodeur. La partie analyse permet de transformer le signal d'origine dans la représentation latente. La représentation latente qui correspond au *bottleneck* du réseau correspond à la donnée transmise entre le codeur et le décodeur. Cette représentation est quantifiée $\hat{\mathbf{y}}$, puis elle est codée par un codeur entropique. Dans leur approche, le codeur utilisé est un codeur arithmétique. Cette représentation est ensuite transmise à la partie synthèse.

La partie synthèse permet d'inverser cette transformation et de retrouver le signal d'origine à partir de cette représentation latente. L'objectif du réseau de neurones est donc de trouver une représentation $\mathbf{y}$ avec une densité de probabilité $p_{\mathbf{y}}$ optimale pour être codée par un codage arithmétique.

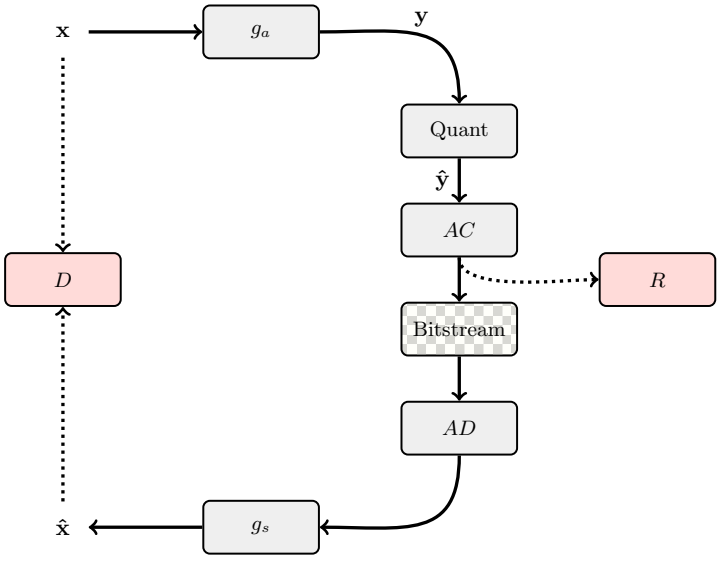


FIGURE 7.1 – Principe de l'approche de compression par réseau de neurones.

La figure 7.1 montre le fonctionnement du modèle. Le signal $\mathbf{x}$ est transformé en une représentation $\mathbf{y} = g_a(\mathbf{x}; \theta_e)$, avec g_a la fonction appliquée par la partie analyse du réseau et θ_e les paramètres du réseau. Ces paramètres seront ceux appris pendant l'entrainement du modèle. La représentation latente $\mathbf{y}$ est ensuite quantifiée :

$$\hat{\mathbf{y}} = \lfloor \mathbf{y} \rceil \tag{7.2}$$

où $\lfloor . \rceil$ est l'opérateur qui arrondit la valeur à l'entier le plus proche. La représentation de l'espace latent $\hat{\mathbf{y}}$ une fois quantifiée est codée par un codeur entropique. Puis cette représentation est trans-

mise au décodeur. À partir de la représentation $\hat{\mathbf{y}}$, la partie synthèse du réseau g_s reproduit le signal d'origine $\hat{\mathbf{x}} = g_s(\hat{\mathbf{y}}, \theta_d)$. Lors de l'entrainement du réseau, les paramètres θ_e et θ_d sont optimisés selon la fonction de coût :

$$\mathcal{L}(\lambda) = \lambda R(\hat{\mathbf{y}}) + D(\mathbf{x}, \hat{\mathbf{x}}) \tag{7.3}$$

où D est la mesure de la distorsion entre le signal d'entrée $\mathbf{x}$ et le signal de sortie $\hat{\mathbf{x}}$ et R l'estimation du débit nécessaire pour transmettre l'espace latent. Le compromis entre fidélité de la reconstruction et débit est paramétrable par la valeur λ. Un λ petit favorisera la qualité de reconstruction au détriment du débit, une λ grand favorisera le débit, mais la qualité du signal audio en sortie sera dégradée.

Pour mesurer la distorsion D, il est possible d'utiliser diverses métriques. Selon la métrique utilisée, les résultats de reconstruction peuvent être différents. Dans le domaine de l'image, l'article [Ballé *et al.*, 2018] compare les résultats obtenus par le même modèle entrainé avec plusieurs métriques objectives utilisées comme mesure de distorsion D : l'erreur quadratique moyenne, ou Mean Squared Error (MSE), le rapport signal sur bruit de crête, ou Peak Signal-to-Noise Ratio (PSNR), et la similarité structurelle multi échelle, ou MultiScale Structural SIMilarity (MS-SSIM) [Wang *et al.*, 2003]. Pour chaque métrique, de légères différences sont présentes dans les images reconstruites par le modèle. Certaines métriques favorisent les contours des objets dans l'image quand d'autres favorisent la reproduction des textures.

Dans le domaine de l'audio, plusieurs métriques objective existent. Les principales métriques sont présentées dans le chapitre 2.2.2. Les plus élaborées prennent en compte des aspects perceptifs, comme POLQA [ITU-T P.863, 2018] ou *ViSQOL* [Chinen *et al.*, 2020]. Cependant, ces métriques sont souvent très coûteuses en temps de calcul, ce qui les rend peu adaptées pour être utilisées durant l'entrainement d'un réseau de neurones. Pour notre modèle de compression audio, nous avons donc utilisé la MSE :

$$D(\mathbf{x}, \hat{\mathbf{x}}) = E_{\mathbf{x}\sim p_{\mathbf{x}}}[\|\mathbf{x} - \hat{\mathbf{x}}\|^2] \tag{7.4}$$

Lors de l'entrainement du modèle, le débit $R(\hat{\mathbf{y}})$ nécessaire pour transmettre l'espace latent est approximé par l'entropie de Shannon :

$$R(\hat{\mathbf{y}}) = E_{\mathbf{x}\sim p_{\mathbf{x}}}[-\log_2 p_{\hat{\mathbf{y}}}(\hat{\mathbf{y}})] \tag{7.5}$$

avec $p_{\hat{\mathbf{y}}}$ la distribution de probabilité de $\hat{\mathbf{y}}$. L'entropie de Shannon est utilisée, car elle permet d'avoir une expression mathématique simple pour estimer le débit nécessaire pour $\mathbf{y}$. Les expressions de la distorsion et du débit permettent d'obtenir :

$$\begin{aligned} \mathcal{L}(\lambda) &= \lambda D(\mathbf{x}, \hat{\mathbf{x}}) + R(\hat{\mathbf{y}}) \\ &= \lambda E_{\mathbf{x}\sim p_{\mathbf{x}}}[\|\mathbf{x} - \hat{\mathbf{x}}\|^2] + E_{\mathbf{x}\sim p_{\mathbf{x}}}[-\log_2 p_{\hat{\mathbf{y}}}(\hat{\mathbf{y}})] \end{aligned} \tag{7.6}$$

Dans le modèle, deux hypothèses sont faites sur les probabilités de l'espace latent $p_{\mathbf{y}}$. La première est que les valeurs de chaque couche latente $p_{\mathbf{y}_i}$, qui compose l'espace latent, sont indépendantes d'une couche à l'autre. La seconde est qu'à l'intérieur de la $i^{\text{ième}}$ couche latente, tous les éléments suivent la même distribution $p_{\mathbf{y}_i}$:

$$p_{\mathbf{y}} = \prod_{i,j,k} p_{\mathbf{y}_i}(\mathbf{y}_{i,j,k}) \tag{7.7}$$

où j et k sont les coordonnées de l'élément de la carte d'activation 2D $i^{\text{ième}}$, ou *featuremap* de l'espace latente. Chaque couche $p_{\mathbf{y}_i}$ est modélisée par une distribution gaussienne telle que :

$$p_{\mathbf{y}_i} = \mathcal{N}(0, \sigma_i^2) \tag{7.8}$$

La valeur de la variance σ_i^2 est apprise par le réseau de neurones lors de son entrainement.

Dérivabilité du modèle de compression

Pour être entrainable, un réseau de neurones a besoin que chacune de ses opérations soit dérivable. Cela est nécessaire pour effectuer la rétropropagation du gradient au travers des différentes couches du réseau. De plus, les dérivées doivent être non-nulles, une dérivée nulle rendrait la descente de gradient inopérante. Pour l'opération de quantification $\hat{\mathbf{y}} = \lfloor \mathbf{y} \rceil$, est continue par morceaux, sa dérivée peut être exprimée comme :

$$\frac{\partial Q}{\partial y}(y) = \begin{cases} \delta(y) & \text{si } y \in \mathbb{Z} \\ 0 & \text{sinon} \end{cases} \tag{7.9}$$

où Q est la fonction de quantification et δ est la fonction de Dirac. La dérivée de la quantification est nulle presque partout, ce qui rend la rétropropagation du gradient impossible. En terme de probabilités, la quantification $\hat{\mathbf{y}} = \lfloor \mathbf{y} \rceil$ peut être modélisée par un peigne de Dirac où la pondération de chaque Dirac est donnée par une fonction de masse de probabilités q_i :

$$P_{q_i}(n) = \int_{n-\frac{1}{2}}^{n+\frac{1}{2}} p_{\mathbf{y}_i} dy \qquad \forall n \in \mathbb{Z} \tag{7.10}$$

où un P_{q_i} est calculé pour chaque carte d'activation i de l'espace latent $\mathbf{y}$. Pour permettre la rétropropagation du gradient lors de l'entrainement, la quantification est remplacée par une fonction approximant cette dernière. Cette approximation doit répondre à deux contraintes : elle doit produire une fonction de densité proche de la fonction de masse $p_{\hat{\mathbf{y}}}$, être dérivable et non nulle.

Dans [Ballé *et al.*, 2018], les auteurs approximent la quantification par l'ajout d'un bruit uniforme réparti i.i.d. $\Delta\mathbf{y}$ à la source $\mathbf{y}$. La loi uniforme qui produit ce bruit a une densité de probabilité $P_{\delta\mathbf{y}}$ centré sur 0 et d'une largeur égale au pas de quantification. Dans notre cas, le pas de quantification est de largeur 1 :

$$\tilde{\mathbf{y}} = \mathbf{y} + \Delta\mathbf{y} \tag{7.11}$$

avec $\Delta\mathbf{y} \sim \mathcal{U}[-\frac{1}{2}; \frac{1}{2}]$. Pour la suite, il est important de bien distinguer : $\mathbf{y}$ les valeurs de la distribution de l'espace latent que le modèle cherche à approximer, $\hat{\mathbf{y}}$ les valeurs de la distribution produite par la quantification de l'espace latent et $\tilde{\mathbf{y}}$ les valeurs de la distribution produite par l'approximation de la quantification. La densité de probabilité de l'approximation de la quantification est :

$$p_{\tilde{\mathbf{y}}_i} = p_{\mathbf{y}_i} * \mathcal{U}[-\frac{1}{2}; \frac{1}{2}] \tag{7.12}$$

L'ajout du bruit uniforme produit une convolution de la densité de probabilité $p_{\mathbf{y}}$ par une fonction porte d'une largeur $[-\frac{1}{2}, \frac{1}{2}]$, ce qui permet d'obtenir une version lissée de $p_{\mathbf{y}}$.

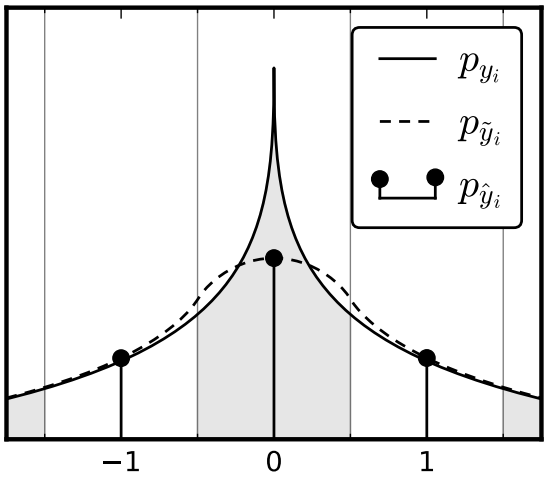


FIGURE 7.2 – Illustration de la densité de probabilité d'une valeur de l'espace latent $\mathbf{y}_i$ de la latente quantifiée $p_{\hat{\mathbf{y}}_i}$ et l'approximation de cette quantification $p_{\tilde{\mathbf{y}}_i}$, figure issue de [Ballé *et al.*, 2017].

La figure 7.2 montre une illustration de l'approximation de la quantification pour une latente $\mathbf{y}_i$. Cette illustration permet de voir la différence entre la densité de probabilité de l'espace latent $p_{\mathbf{y}_i}$, la densité discrète de l'espace latent quantifié $p_{\hat{\mathbf{y}}_i}$ et la densité de probabilité $p_{\tilde{\mathbf{y}}_i}$ de l'approximation de la quantification rendue continue par l'ajout du bruit uniforme. La validité de l'hypothèse d'approximer la quantification par l'ajout d'un bruit uniforme a été vérifiée expérimentalement et présentée de manière détaillée dans [Ballé *et al.*, 2017, Section 4]. Grâce à cette relaxation du problème, il donc est possible de rétropropager le gradient et ainsi entrainer le modèle. Lors de la phase d'inférence, l'approximation de la quantification est remplacée par la véritable quantification dans le modèle.

7.2.2 Architecture détaillée du réseau

Dans la section précédente, nous avons présenté le fonctionnement du modèle de compression d'image par réseau de neurones. Cependant, il est possible de remarquer que le fonctionnement de la méthode n'est pas spécifique au domaine de l'image et qu'une adaptation peut être réalisée pour la compression audio. La compression audio par réseau de neurones est un sujet nouveau qui se confronte encore à de nombreux problèmes et limitations. C'est pour cela que nous avons fait le choix de ne pas prendre en compte les questions de latence et le découpage en trame du signal dans notre modèle.

Pour la phase d'entrainement, notre modèle prend un signal mono échantillonné à 48 kHz. Un segment de 1 seconde est découpé aléatoirement de ce signal. Une transformée MDCT $\mathbf{X}$ est calculée pour le segment de signal audio $\mathbf{x}$. C'est ce signal $\mathbf{X}$ qui sera fourni en entrée du réseau

de neurones. Les trames de la MDCT sont de 2^8 échantillons avec un recouvrement de 50 %. Pour chaque trame, un fenêtrage en cosinus est appliqué.

Pour garantir des transformées avec des amplitudes comparables entre les segments, une normalisation des valeurs de $\mathbf{X}$ est faite :

$$\mathbf{X}_{\text{norm}} = \frac{\mathbf{X}}{\max |\mathbf{X}|} \tag{7.13}$$

Cette normalisation permet que les données soient comprises entre $[-1, 1]$ quel que soit le contenu du signal et la puissance du signal. Un seul gain $\max |\mathbf{X}|$ est calculé par segment de 1 seconde. La valeur du gain est transmise comme métadonnée au décodeur. Le débit pour transmettre cette valeur au décodeur est très faible (< 16 bits/s).

Dans l'article [Caracalla et Roebel, 2020], les auteurs proposent de compresser les valeurs de la transformée de Fourier fournie à l'entrée de leurs modèles. Cette compression a pour but de pondérer la représentation en donnant plus d'importance aux événements moins énergétiques, comme les hautes fréquences et les sons faibles. La compression est appliquée de manière indépendante à la partie réelle et à la partie imaginaire de la transformée de Fourier. Pour notre modèle, la même compression a été appliquée sur la transformée MDCT normalisée $\mathbf{X}_{norm}$:

$$\mathbf{X}_{\text{comp}} = 2 \times S(c\mathbf{X}_{\text{norm}}) - 1 \tag{7.14}$$

avec $S(.)$ est la fonction sigmoïde $S(x) = \frac{1}{1+e^{-x}}$ et c le facteur de compression fixé à $c = 10$.

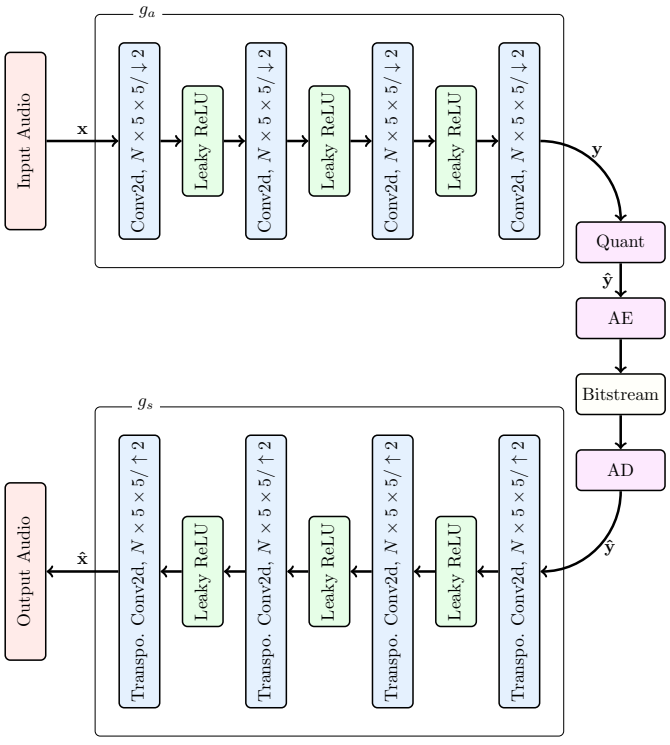


FIGURE 7.3 – Architecture du réseau de neurones.

La figure 7.3 montre l'architecture complète du VAE. La partie analyse g_a est composée de quatre couches. Chacune des couches sont constituées d'une convolution 2D avec des filtres de dimension 5×5, suivie d'une décimation par 2 de la taille des cartes d'activation. La taille des cartes d'activation est donc de plus en plus petite à mesure que l'on avance dans la partie analyse. À la suite de chacune des 3 premières couches, une fonction d'activation *Leaky ReLU* est utilisée. Pour la dernière couche, il n'y a pas de fonction d'activation pour ne pas limiter les valeurs que peut prendre $\mathbf{y}$ à la sortie de la couche. La présence d'une fonction d'activation de type sigmoïde ou tangente hyperbolique, contraindrait les valeurs de sortie dans un intervalle $]-1, 1[$. La distribution $p_{\mathbf{y}_i}$ devant approximer une loi normale, l'utilisation d'une fonction d'activation asymétrique, de type *Leaky ReLU*, ne rendrait que la convergence plus longue et difficile.

À la sortie de la partie analyse, selon que le réseau soit en entrainement ou en inférence, la quantification $\lfloor . \rceil$ ou l'approximation de la quantification est réalisée pour obtenir $\hat{\mathbf{y}}$ ou $\tilde{\mathbf{y}}$. Les valeurs de l'espace latent sont ensuite codées par un codeur arithmétique, ou Arithmetic Encoder (AE), puis le train binaire est transmis vers le décodeur.

Dans le décodeur, le train binaire sera décodé par le décodeur arithmétique, ou Arithmetic Decoder (AD), pour obtenir $\hat{\mathbf{y}}$ ou $\tilde{\mathbf{y}}$. Pour la partie synthèse, g_s à une architecture construite en miroir par rapport à la partie analyse g_a. Il y a 4 couches successives de convolution transposée 2D (*Transposed Convolution*). Dans l'architecture originale [Ballé *et al.*, 2017], une simple convolution 2D puis une interpolation linéaire étaient utilisées. L'ajout de convolution transposée permet une interpolation non linéaire plus riche qu'une simple pondération linéaire des valeurs [Minnen *et al.*, 2018]. Comme pour la partie analyse, à la suite de chacune des 3 premières couches, une fonction d'activation *Leaky ReLU* est utilisée.

Le nombre de cartes d'activation N permet de donner plus ou moins de degrés de liberté au modèle pour représenter les signaux d'entrée. Pour un entrainement fait uniquement avec une contrainte de distorsion sans contrainte de débit ($\lambda = 0$), plus la valeur de N sera élevée plus la qualité de reconstruction du modèle sera importante. Pour un N donné, un entrainement fait uniquement avec une contrainte de distorsion ($\lambda = 0$) permet d'avoir une estimation de la qualité de reconstruction maximale que peut attendre le modèle avec N carte d'activation. Avec l'introduction de la contrainte de débit ($\lambda > 0$), la qualité de reconstruction sera forcement moins importante que cette qualité maximale.

L'analyse de l'espace récepteur, ou *receptive field*, permet de connaître la quantité d'informations de $\mathbf{x}$ accessibles pour chaque élément $\mathbf{y}_{i,j,k}$ de l'espace latent. Dans notre architecture, une case d'une couche de l'espace latent $\mathbf{y}_{i,j,k}$ a un espace récepteur de taille 33×33. Cet espace récepteur correspond à une portion de la MDCT représentant une partie du signal d'entrée d'une longueur de 90,6 ms et d'une largeur de 6,19 kHz.

7.2.3 Détails de mise en œuvre et d'optimisations

Le réseau utilise l'algorithme d'optimisation Adam [Kingma et Ba, 2015] pour apprendre les paramètres θ_a et θ_s. Le taux d'apprentissage, ou *learning rate*, est initialisé à $\alpha = 10^{-4}$, il est divisé par 5 tous les 100 époques, ou *epoch*, jusqu'à atteindre $\alpha = 8 \times 10^{-7}$. Chaque entrainement est réalisé sur un ensemble de 300 époques. La taille du *batch* est fixée à 8. Lors de la phase d'initialisation, les différentes cartes d'activation sont initialisées selon l'initialisation

de Glorot [Glorot et Bengio, 2010], aussi appelé *Xavier initialization*. Les entrainements ont été réalisés sur des machines équipées d'une carte graphique *Nvidia GeForce RTX 2080 Ti* et d'un processeur *Intel Core i9-9940X*. Lors de l'inférence, les résultats et les temps d'exécution ont été mesurés sur une machine identique aux machines d'entrainement. L'implémentation a été réalisée en *Python* avec le *framework Pytorch v1.8*.

La valeur du paramètre λ qui pondère l'importance du débit de la fonction de coût est compris entre $\left[10^1, 10^5\right]$ selon le débit ciblé. Quand la contrainte de débit est très forte (λ est petit) le modèle peut rencontrer des difficultés à converger vers une solution satisfaisante. Le modèle peut rester bloqué dans un optimal local qui consiste à transmettre un espace latent $\mathbf{y}$ nul, pour minimiser le coût R sans chercher à reproduire le signal. Pour pallier ce problème, lors des entrainements avec $\lambda \geq 10^2$, les 3 premières époques sont réalisées avec un $\lambda = 10^4$ avant d'être remises à la valeur du lambda initial choisi. Des hypothèses peuvent être formulées quant à la raison de cette instabilité. Certaines pistes laissent penser que cela pourrait venir de l'initialisation des cartes d'activation qui pourrait être sous-optimale pour notre problème [Kumar, 2017]. Néanmoins, le problème n'a pas été étudié plus en détail dans le cadre de cette thèse.

7.2.4 Expérimentations

Pour évaluer la performance de notre modèle, un ensemble d'expérimentations a été conduit. Lors de ces tests, nous nous sommes concentrés sur la compression de la parole. Certaines méthodes récentes de compression de parole par réseau de neurones [Valin et Skoglund, 2019] ont déjà montré des performances supérieures à l'état de l'art à très bas débit (< 10 kbit/s). Nous avons donc fait le choix de nous intéresser à la compression sur une gamme de débit plus importante, allant de 16 kbit/s à 128 kbit/s pour des signaux échantillonnés à 48 kHz.

7.2.4.1 Base de données

La base de données utilisée pour nos expérimentations est la base de corpus VTCK [Veaux *et al.*, 2017]. Ce corpus est composé d'enregistrements de 400 phrases en anglais prononcé par 109 locuteurs (femmes et hommes), pour un total d'environ 40000 extraits audio. Les enregistrements sont échantillonnés à 48 kHz. La durée des phrases est comprise entre 2 secondes et 20 secondes.
Dans chaque enregistrement, un silence plus ou moins long peut être présent avant et après chaque phrase. Pour certains échantillons, ce silence peut être de plusieurs secondes. Ces segments de silence ont été supprimés des échantillons d'origine pour ne pas ajouter de biais lors de l'apprentissage du réseau. La détection des silences a été faite par une simple détection du passage du signal au-dessus d'un certain niveau crête. Ce niveau crête a été fixé à -40 dB. Une fois, le découpage des échantillons réalisé, le niveau RMS de l'ensemble des échantillons a été normalisé à un niveau de -23 dB.

Le corpus a été divisé en 3 parties, pour former la base d'entrainement, de validation et de test. Les échantillons ont été séparés aléatoirement selon la proportion de 80 %, 10 %, 10 % respectivement pour la base d'entrainement, de validation et de test.

7.2.4.2 Capacité de reconstruction du modèle

Des expérimentations ont été menées pour évaluer la capacité du modèle à reproduire un signal d'entrée sans prendre en considération la contrainte de débit dans la fonction de coût ($\lambda = 0$). Dans ce premier test, l'idée est de faire varier le nombre de cartes d'activation N et de mesurer la qualité de reconstruction maximale que peut obtenir le modèle.

Tableau 7.1 – Qualité de reconstruction et temps d'exécution en fonction du nombre N de *feature maps* (résultats moyens ainsi que les écarts-types).

Nombre de cartes d'activation N	***MSE* moyen ($\times 10^{-7}$)**	***SegSNR* moyen (en dB)**	**temps moyen d'exécution (en ms)**
32	1360,53 ($\pm$1390,56)	12,28 ($\pm$2,67)	68,3 ($\pm$4,3)
64	9,82 ($\pm$10,46)	28,50 ($\pm$2,33)	72,3 ($\pm$5,7)
128	1,00 ($\pm$0,80)	36,03 ($\pm$2,15)	74,4 ($\pm$4,5)
256	0,74 ($\pm$0,97)	37,85 ($\pm$2,21)	86,0 ($\pm$4,8)

Le tableau 7.1 rassemble les scores de reconstruction du modèle, c'est à dire la fidélité de reconstruction entre le signal d'origine et le signal codé, selon la valeur de N. Deux métriques ont été utilisées : la MSE et le SegSNR. Les scores indiqués dans le tableau correspondent à la moyenne et l'écart-type des scores obtenus pour du corpus de test.

Pour la métrique MSE, plus le signal codé est proche du signal d'origine plus la valeur est petite. La MSE étant utilisé comme critère de distorsion dans la fonction de coût, il semble normal que la valeur obtenue soit relativement basse quelque soit le modèle. Il est à noter que plus le nombre de cartes d'activation N augmente, plus la MSE diminue.

Pour la métrique SegSNR, contrairement à la MSE, plus le signal codé est proche de l'original plus le score sera important. Les modèles ayant été entrainés avec la métrique MSE, il est intéressant d'observer les résultats obtenus par les modèles pour une autre métrique. Comme pour la MSE, plus le nombre de cartes d'activation N augmente, plus la fidélité entre le signal d'origine et le signal codé augmente. Cependant, alors que pour la MSE un plateau semble être atteint à partir de $N = 128$, le SegSNR semble continuer à augmenter au-delà de $N = 128$. Cela peut être expliqué par le fait que le modèle n'est pas optimisé par le même critère que celui utilisé pour l'évaluation. Il est toutefois intéressant de noter que pour deux métriques objectives similaires, l'optimisation selon un critère permet d'obtenir des résultats proches pour l'autre métrique. Comme aucune contrainte de débit n'a été intégrée à la fonction de coût ($\lambda = 0$), les modèles se sont entrainés sans prendre en compte ce critère et par conséquent la distribution des valeurs sur chacune des cartes d'activation, l'observation a posteriori du débit utilisé par les modèles n'est pas pertinente.

Ce tableau indique également le temps d'exécution moyen pris par le modèle pour coder et décoder une seconde de signal. Les temps mesurés montrent qu'il est possible de coder un échantillon $12\times$ plus vite que le temps réel, ce qui permet d'imaginer l'utilisation d'un tel modèle pour des codecs audio conversationnels. Néanmoins, deux éléments sont à prendre en considération. Le premier est que seul le temps d'exécution du réseau a été pris en compte, la durée de codage et du décodage entropique doivent être ajoutées pour connaître le temps d'exécution total du co-

dage. Le second, le temps d'exécution a été mesuré pour notre implémentation en *Python* sur un ordinateur de bureau équipé d'une carte graphique. Les auteurs de l'article [Valin et Skoglund, 2019], en implémentant en *C* leurs modèles de compression de parole très bas débit < 16 kbit/s sur un processeur *ARM Snapdragon 845* ont mesuré des temps d'exécution à $1{,}5\times$ le temps réel, ce qui laisse penser que l'utilisation des réseaux de neurones pour les codecs conversationnels est envisageable.

Lors de nos expérimentations, le nombre de paramètres et l'occupation mémoire ont également été mesurés pour les différents modèles. Le tableau 7.2 présente le nombre de paramètres selon N, ainsi que la taille des paramètres et l'occupation mémoire totale du réseau pendant l'entrainement. La première information qui ressort de ce tableau est que le nombre de paramètres et leurs tailles sont relativement faibles par rapport à des modèles de synthèse audio. À titre de comparaison, le modèle *WaveRNN* utilisé dans [Valin et Skoglund, 2019] comme une brique dans le processus de compression bas débit, ou le modèle de *text-to-speech Tacotron 2* [Shen *et al.*, 2018], sont composés respectivement de 3 millions et 13 millions de paramètres. La deuxième observation est que le nombre de paramètres augmente de manière exponentielle avec le nombre de cartes d'activation. Dans l'optique d'utiliser ce modèle comme codec audio conversationnel sur des appareils avec des puces dédiées ou une mémoire limitée, il peut être intéressant de garder une valeur de N assez faible.

Une piste pour augmenter ce nombre tout en gardant un nombre de paramètres limité pourrait être de mettre en place des algorithmes de *pruning*. Cette optimisation simplifie le modèle une fois entrainé en supprimant les connexions trop peu actives entre les neurones, cela permet de réduire le nombre de paramètres ainsi que le temps de calcul. Cependant dans l'article [Ballé *et al.*, 2017], les auteurs semblent suggérer que l'impact de ce genre d'approche dégrade fortement la qualité de reconstruction.

Tableau 7.2 – Nombre de paramètres et occupation mémoire en fonction du nombre de cartes d'activation N.

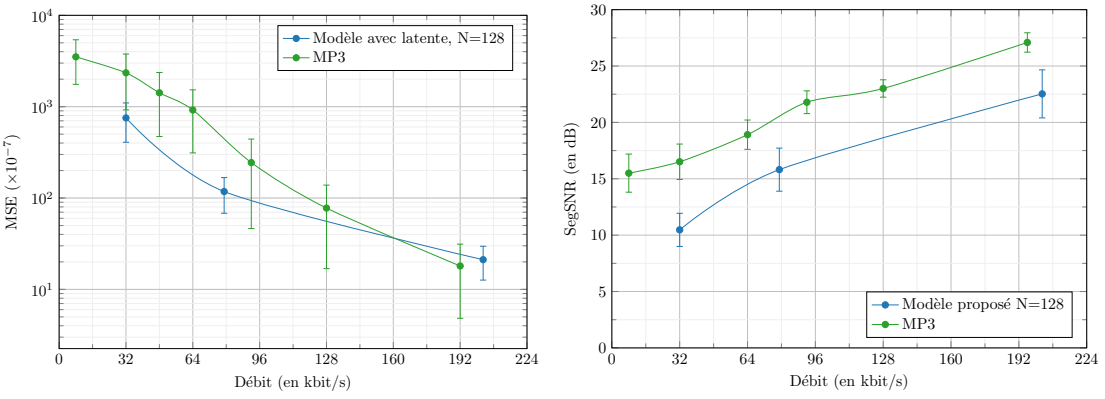
Nombre de cartes d'activation N	**Nombre de paramètres (en milliers)**	**Taille des paramètres (en Mo)**	**Occupation mémoire totale (en Mo)**
32	77,7	0,3	61,0
64	308,9	1,18	93,7
128	1232,0	4,7	185,8
256	4922,2	18,78	270,9

7.2.4.3 Capacité de compression du modèle

Les premières expérimentations ont permis de montrer les capacités du modèle à reproduire un signal audio sans contrainte de débit. Dans ces nouvelles expérimentations, nous allons maintenant étudier les performances du modèle en faisant varier λ, la pondération du débit R par rapport a la distorsion D dans la fonction de coût. Pour ces expérimentations, le modèle sélectionné a été le modèle avec $N = 128$ cartes d'activation car il permet une bonne qualité de reconstruction pour

un nombre de paramètres et un temps d'entrainement du modèle raisonnable. Pour ces tests, une comparaison des résultats a été faite entre notre modèle et le codec MP3. L'implémentation utilisée du codec MP3 est celle de *Lame v3.100* dans son mode Constant Bit-Rate (CBR). Trois modèles ont été entrainés avec chacun une différente valeur de $\lambda = \{10^2, 10^3, 10^4\}$.

Pour comparer la qualité, deux métriques objectives ont été utilisées : la MSE et le SegSNR. Les deux métriques ont été calculées pour chacun des échantillons du corpus de test pour chacune des conditions. La figure 7.4 trace la moyenne et l'écart-type obtenus pour chaque métrique. La figure 7.4(a) montre le score MSE obtenu en fonction du débit. Le modèle par réseaux de neurones, parce qu'il a été entrainé avec cette métrique, arrive à obtenir de meilleurs résultats que le celui du codec MP3. Cependant, lors de l'évaluation des performances des différences conditions selon la métrique SegSNR (figure 7.4(b)), le modèle par réseaux de neurones n'arrive pas à égaler les résultats obtenus par le codec MP3. Même si notre modèle propose des résultats intéressants, ce premier modèle ne permet pas d'égaler le score obtenu par le codec MP3 sur toutes les métriques. Toute fois, ce test a permis de mesurer la différence entre les scores obtenus lors de l'entrainement avec et sans la contrainte de débit présenté dans le tableau 7.1. Le temps d'exécution moyen pour compresser une seconde d'audio a été de $75{,}4$ $(\pm 4{,}6)$ ms.



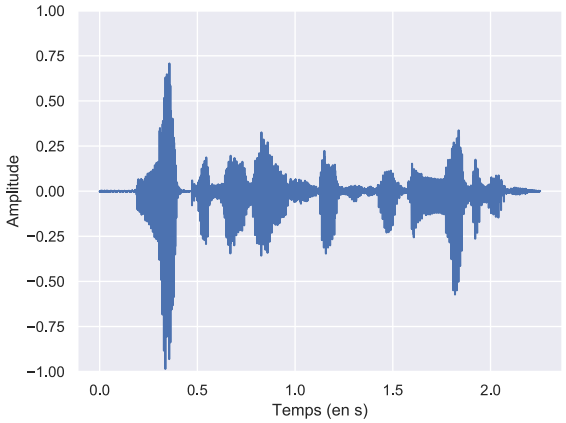
(a) Qualité mesurée selon la métrique MSE.

(b) Qualité mesurée selon la métrique SegSNR.

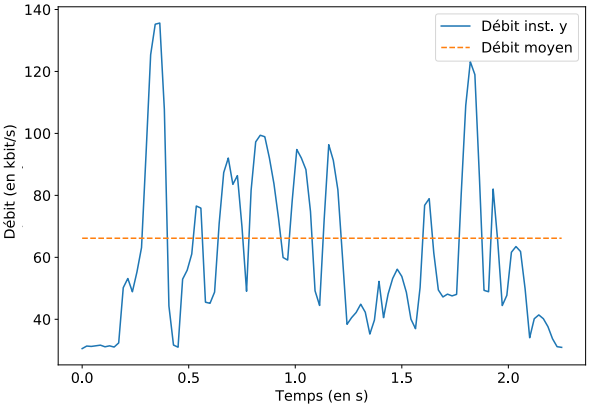
FIGURE 7.4 – Courbe débit-distorsion moyenne calculée sur le corpus de test, pour le codec MP3 et notre modèle en fonction de la pondération λ.

À partir de la fonction de coût, le modèle cherche à optimiser le débit nécessaire pour coder l'espace latent $\mathbf{y}$ du signal d'entrée. Ce débit est un débit moyen, localement l'espace latent $\mathbf{y}$ peut nécessiter plus de débit. Un échantillon donné codé par notre modèle avec $\lambda = 10^3$ aura un débit moyen de $66{,}16$ kbit/s, pourtant les variations de débit à l'intérieur du même échantillon pourront être importantes. La figure 7.5(a) montre la forme d'onde d'un échantillon pris dans le corpus de test. La figure 7.5 trace le débit instantané nécessaire pour coder l'espace latent $\mathbf{y}$ calculé toutes les 50 ms, ainsi que le débit moyen.

Pour plus de lisibilité, les valeurs de débits ont été indiquées en kbit/s. Des variations importantes de débits sont visibles au cours du temps. De plus, des corrélations entre la forme d'onde et le débit nécessaire pour coder le signal semblent présentes.



(a) Forme d'onde du signal étudié.



(b) Débit instantané nécessaire pour coder l'échantillon par notre modèle.

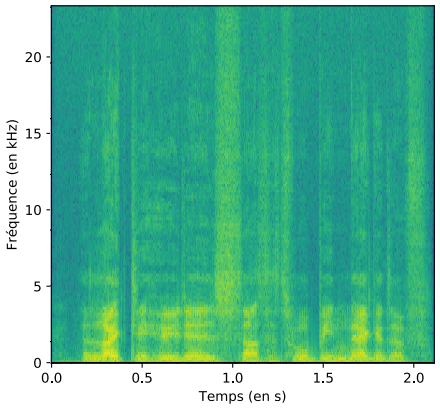
FIGURE 7.5 – Résultats pour un échantillon audio donné.

Le modèle pour la compression d'image [Ballé *et al.*, 2017] sur lequel nous nous sommes basés pour notre modèle permettait d'obtenir des résultats supérieurs au JPEG2000. Cependant, son adaptation directe pour le domaine de l'audio a quant à lui des résultats plus mitigés. Ces premières expérimentations ont permis d'étudier le fonctionnement du modèle et d'en avoir une compréhension plus fine afin de proposer des améliorations du modèle pour la compression audio.

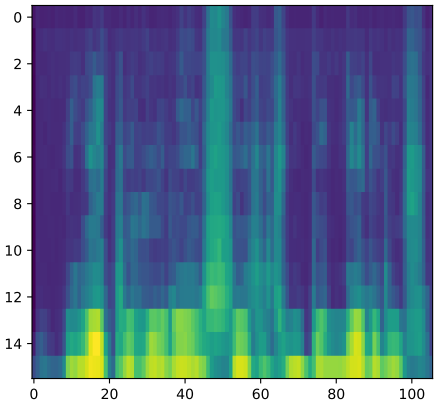
7.3 Compression enrichie par des variables hyper-latentes

Lors des premières expérimentations, l'observation de la distribution à l'intérieur de la même carte d'activation de l'espace latent $p_{\mathbf{y}}$ a montré une certaine dépendance spatiale des activations. La figure 7.6(a) montre la valeur de la transformée MDCT pour le même signal d'exemple utilisé précédemment. Cette transformée MDCT fournie en entrée du modèle produit un ensemble de cartes d'activation latentes $\mathbf{y}$. Il est possible d'extraire une carte $\mathbf{y}_i$ pour observer son activation. La figure 7.6(b) montre une de ces cartes, la structure du signal d'origine est clairement visible.

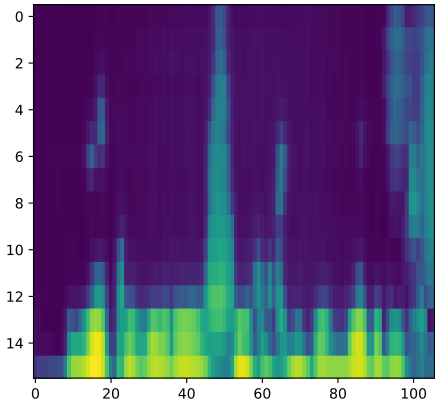
Certaines zones de la carte sont plus activées que d'autres. Toutes ces zones ont donc des distributions différentes de la distribution globale estimée par le modèle pour la carte $\mathbf{y}_i$ de l'espace latent. Cela suggère que la modélisation d'une unique distribution pour l'ensemble d'une carte d'activation est sous-optimale au niveau du codage de la carteL et qu'une autre modélisation de distribution des valeurs de la carte d'activation pourrait être faite.



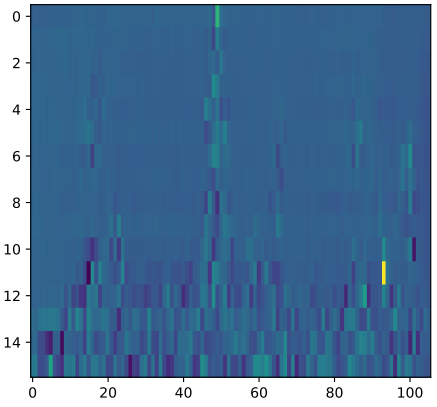
(a) La valeur absolue de la MDCT d'un échantillon audio en échelle log.



(b) Activation d'une couche latente $\mathbf{y}_i$.



(c) La variance σ^2 prédiction par le modèle étendu avec les hyper-latentes.



(d) Activation d'une couche latente $\mathbf{y}_i$ divisée par l'écart-type $\sigma_{i,j,k}$.

FIGURE 7.6 – MDCT et activations de l'espace latent associé pour un échantillon audio.

Les auteurs de l'étude [Minnen *et al.*, 2018] sont arrivés à un constat similaire pour la compression d'images. Ils proposent d'étendre leur modèle en ajoutant un module pour affiner l'estimation de la distribution de l'espace latent $\mathbf{y}$ en introduisant la notion d'un espace hyper-latent. Cet espace hyper-latent permet de modéliser, non pas une distribution globale, mais une distribution propre à chacune des cases j, k de la carte d'activation $\mathbf{y}_i$. Dans notre premier modèle, le réseau devait estimer pour chacune des cartes de l'espace latent $\mathbf{y}_i$, la variance σ^2 de la distribution $p_{\mathbf{y}_i}$, de sorte que :

$$p_{\mathbf{y}_i} \sim \mathcal{N}(0, \sigma_i^2) \tag{7.15}$$

Dans l'article [Minnen *et al.*, 2018], le modèle avec espace hyper-latent doit cette fois estimer la distribution de chaque élément de l'espace latent $\mathbf{y}_{i,j,k}$:

$$p_{\mathbf{y}_{i,j,k}} \sim \mathcal{N}(\mu_{i,j,k}, \sigma_{i,j,k}^2) \tag{7.16}$$

où i l'indice de la couche et j,k les coordonnées de l'élément à l'intérieur de la couche $\mathbf{y}_i$. La figure 7.6(c) montre la variance $\sigma_{i,j,k}^2$ estimée par le modèle hyper-latent pour chaque case de la carte latente $\mathbf{y}_i$. Les cartes hyper-latentes permettent de capturer la structure de l'espace latent par une prédiction de son activation. Cette prédiction peut-être soustraite de la couche latente $pred_{i,j,k} = \mu_{i,j,k}$, seule l'erreur de prédiction $err_{i,j,k} = \mathcal{N}(0, \sigma_{i,j,k}^2)$ reste visible sur la figure 7.6(d). Pour transmettre l'espace latent, seuls l'erreur de prédiction et l'espace hyper-latent doivent être codés par le codeur entropique.

Pour déterminer μ et σ, l'extension du modèle utilise un second VAE. Cet auto-encodeur est composé de deux parties : la partie analyse h_a et la partie synthèse h_s. La partie analyse h_a prend comme donnée d'entrée l'espace latent $\mathbf{y}$ pour le transformer dans une nouvelle représentation : l'espace hyper-latent $\mathbf{z} = h_a(\mathbf{y}, \phi_a)$. Les paramètres ϕ_a seront appris pendant l'entrainement du modèle. Chaque carte hyper-latente $\mathbf{z}_i$ est modélisée par une distribution : $p_{\mathbf{z}_i} \sim \mathcal{N}(0, \sigma_{\mathbf{z}_i}^2)$. De la même manière que pour l'espace latent $\mathbf{y}$, cette représentation hyper-latente $\mathbf{z}$ est quantifiée en arrondissant chaque élément par la valeur entière la plus proche $\hat{\mathbf{z}}_{i,j,k} = \lfloor \mathbf{z}_{i,j,k} \rceil$. L'espace hyper-latent, une fois quantifié, est codé par un codeur arithmétique puis transmis à la partie synthèse. Dans la partie synthèse h_s, la représentation hyper-latente $\hat{\mathbf{z}}$ est utilisée pour estimer $\mu_{i,j,k}, \sigma_{i,j,k} = h_s(\hat{\mathbf{z}}_{i,j,k}, \phi_s)$ pour chaque élément de l'espace latent $\mathbf{y}$. Lors de l'entrainement comme pour $\hat{\mathbf{y}}$, la quantification $\hat{\mathbf{z}}$ est remplacée par l'approximation de la quantification $\tilde{\mathbf{z}} = \mathbf{z} + \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$. Dans ce nouveau modèle, en plus de l'erreur de prédiction de l'espace latent $\hat{\mathbf{y}}$, l'espace hyper-latent $\hat{\mathbf{z}}$ doit également être transmis. Pour prendre cela en compte lors de l'entrainement du modèle, un troisième terme est ajouté à la fonction de coût :

$$\begin{aligned}\mathcal{L}(\lambda) = \lambda\Big(&\underbrace{E_{\mathbf{x}\sim p_{\mathbf{x}}}\left[-\log_2 P_{\hat{\mathbf{y}}}(\hat{\mathbf{y}})\right]}_{\text{Débit de l'espace latent}} + \underbrace{E_{\mathbf{x}\sim p_{\mathbf{x}}}\left[-\log_2 P_{\hat{\mathbf{z}}}(\hat{\mathbf{z}})\right]}_{\text{Débit de l'espace hyper-latent}}\Big) \\ &+ \underbrace{E_{\mathbf{x}\sim p_{\mathbf{x}}}\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|^2}_{\text{Distorsion}}\end{aligned} \tag{7.17}$$

7.3.1 Architecture détaillée du réseau

Pour déterminer μ et σ, l'extension du modèle utilise un second VAE. Pour réaliser cette estimation des paramètres μ et σ, un VAE est ajouté au modèle de la section précédente. La figure 7.7 montre le schéma du codec complet. La partie analyse h_a prend en entrée l'espace latent $\mathbf{y}$. Cette partie est composée d'une couche de convolution 2D avec des filtres 3×3 et d'une *Leaky ReLU*. Cette couche est suivie de 2 couches de convolution 2D avec des filtres 5×5, d'une décimation par 2 et d'une *Leaky ReLU*. La taille des cartes d'activation est divisée par 4 entre

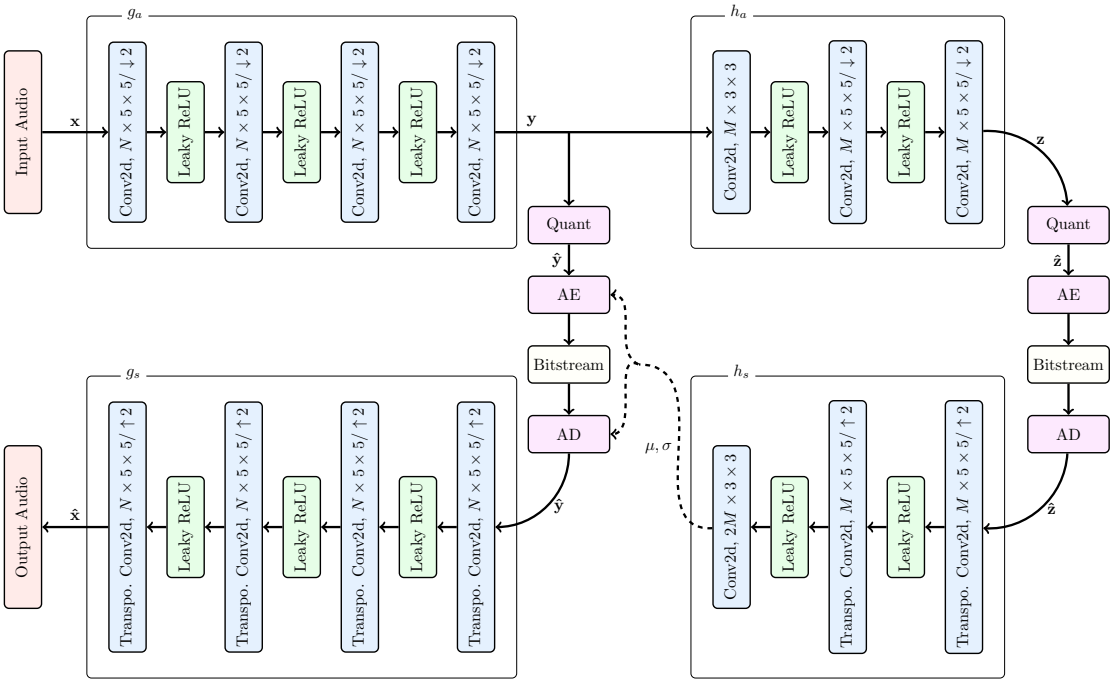


FIGURE 7.7 – Architecture du modèle enrichi avec l'estimation μ et σ grâce au module hyper-latent.

les données d'entrée $\mathbf{y}$ et l'espace hyper-latent $\mathbf{z}$. Le nombre de cartes d'activation pour chaque couche est de M.

À la sortie de la partie analyse, la quantification, ou l'approximation de la quantification, est appliquée à l'espace hyper-latent $\mathbf{z}$ pour obtenir $\hat{\mathbf{z}}$, ou $\tilde{\mathbf{z}}$, selon que le réseau soit en entrainement ou en inférence. La quantification obtenue est traitée par le codeur arithmétique, *Arithmetic Encoder* (*AE*) puis le train binaire est transmis vers le décodeur.

Dans le décodeur, le train binaire est décodé par le décodeur arithmétique, *Arithmetic Decoder* (*AD*) afin d'obtenir $\hat{\mathbf{z}}$. La partie synthèse h_s prend en entrée l'espace hyper-latent $\hat{\mathbf{z}}$ (ou $\tilde{\mathbf{z}}$) pour estimer le $\mu_{i,j,k}$ et $\sigma_{i,j,k}$ pour chaque élément de l'espace latent $\mathbf{y}_{i,j,k}$. Cette partie est composée en miroir par rapport à la partie analyse, avec deux couches de convolution transposée 2D avec des filtres de taille 5×5, suivies d'une couche de convolution 2D de taille 3×3. À partir de la prédiction de la distribution μ et σ et de l'erreur de prédiction, le décodeur peut retrouver $\hat{\mathbf{y}}$ et resynthétiser le signal d'origine $\mathbf{x}$ en utilisant la partie synthèse g_s du VAE.

Comme le modèle avec espace latent simple et le modèle avec espace hyper-latent utilisent une architecture similaire, les conditions d'entrainement et les différentes optimisations utilisées pour le premier modèle ont été reprises pour le second modèle. Ces détails d'implémentation et diverses optimisations peuvent être trouvés à la section 7.2.3.

À titre d'exemple, pour le modèle avec représentation hyper-latente avec $N = 128$, $M = 64$, pour un signal audio de 1 seconde échantillonné à 48 kHz et une transformée $MDCT$ de 2^8 points, la dimension du tenseur d'entrée sera de $(1, 396, 128)$. L'espace latent en entrée de l'analyse h_a sera de $(128, 50, 16)$, ce qui produit un espace hyper-latent de dimension $(64, 13, 4)$. La synthèse

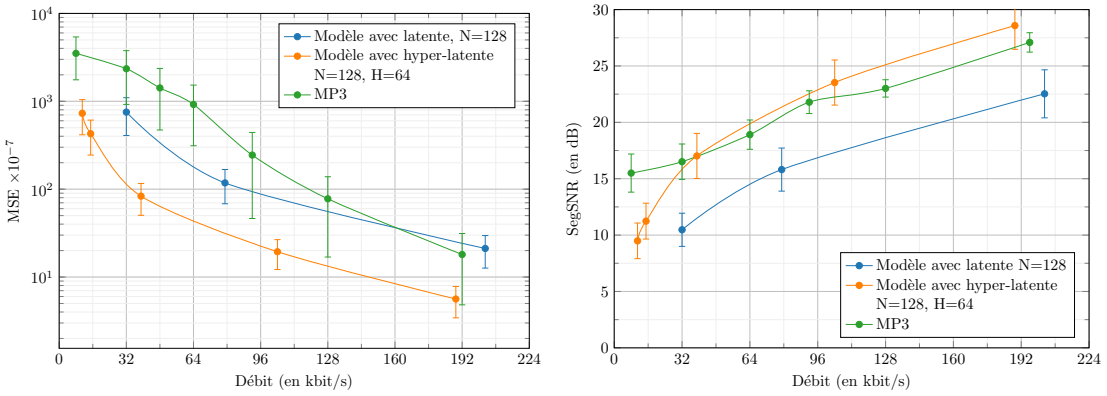
h_s génère un tenseur de dimension $(256, 52, 16)$. Ce tenseur représentant l'estimation μ et σ, il est découpé en deux parties de dimension $(128, 52, 16)$, l'une des parties modélisées μ, l'autre σ. Les deux tenseurs sont tronqués pour avoir la même taille que l'espace latent d'origine $\mathbf{y}$. En sortie du modèle, le tenseur a la dimension $(1, 396, 128)$. Le nombre total de paramètres est de 3 millions et qui représente une taille de 12 Mo.

7.3.2 Expérimentations

7.3.2.1 Comparaison selon la métrique SegSNR

Pour étudier les performances du modèle avec couches hyper-latentes et mesurer le gain ajouté par rapport au premier modèle, un test similaire au test conduit à la section 7.2.4.3 a été réalisé. Ce test cherche à mesurer, selon le débit, la qualité audio selon deux métriques objectives, la MSE et le SegSNR. Dans ce test, trois méthodes ont été évaluées : le modèle avec couches latentes, le modèle enrichi des couches hyper-latentes et le codec MP3.

Pour le modèle avec couches hyper-latentes, le nombre de cartes d'activation pour l'espace latent est $N = 128$ et de $M = 64$ pour l'espace hyper-latent. Le modèle a été entrainé pour différentes valeurs de $\lambda = \{5 \times 10^1, 10^2, 10^3, 10^4, 5 \times 10^5\}$. Comme pour le premier test, l'implémentation utilisée du codec MP3 est celle de *Lame v3.100* dans les mêmes configurations que précédemment.



(a) Qualité mesuré selon la métrique MSE.

(b) Qualité mesuré selon la métrique SegSNR.

FIGURE 7.8 – Qualité audio moyenne pour le codec MP3 ainsi que les modèles avec espace latent et enrichi de l'espace hyper-latent. Le nombre de cartes latentes a été fixé à $N = 128$ et le nombre de cartes hyper-latentes à $M = 64$.

Pour comparer la qualité, la métrique objective utilisée a été le *SegSNR*. Le *SegSNR* a été mesuré pour chacun des échantillons du corpus de test. Puis, le *SegSNR* moyen et son écart-type ont été calculés. La figure 7.8(b) montre la qualité moyenne, ainsi que les écarts-types obtenus pour chacune des méthodes. Le modèle avec espace hyper-latent obtient des performances significativement meilleures que le modèle avec couche latente simple quelle que soit la métrique utilisée.

Pour une qualité équivalente le modèle avec espace hyper-latent demande un débit presque 2 fois moins important que le modèle simple. Pour la métrique MSE, métrique pour laquelle le modèle avec espace hyper-latent a été entrainé, le modèle obtient de bien meilleures performances que le codec MP3. Pour la seconde métrique, ce modèle arrive toutefois à un niveau SegSNR équivalent MP3 pour un débit supérieur à 32 kbit/s.

Pour ce nouveau modèle, le temps d'exécution moyen pour compresser une seconde de signal audio est de $78{,}81$ $(\pm 6{,}55)$ ms. À titre de comparaison, le temps d'exécution moyen pour le modèle avec espace latent simple était de $75{,}4$ $(\pm 4{,}6)$ ms. L'ajout de la modélisation de l'espace hyper-latent ajoute donc un temps d'exécution inférieur à 5 % du temps total, ce qui est raisonnable au vu de l'amélioration de qualité apportée. En terme de nombre de paramètres, ce nouveau modèle nécessite 3 millions de paramètres, contre $2{,}4$ millions pour le modèle sans couche hyper-latente.

7.3.2.2 Comparaison selon la métrique PEAQ

Grâce aux tests conduits à la section précédente, nous avons pu fournir une première évaluation de la capacité de compression de notre nouveau modèle selon des métriques objectives. Cependant, ces métriques objectives prennent en compte uniquement l'erreur de reconstruction du signal sans prendre en considération la perception auditive. Un nouveau test a donc été conduit avec une métrique objective intégrant une modélisation de la perception humaine dans son système de notation de la qualité audio. L'utilisation d'une telle métrique permet une notation plus proche de la qualité subjective que celle obtenue par des métriques objectives basées uniquement sur la reconstruction du signal comme le SegSNR ou la MSE. Pour ce test, c'est la métrique PEAQ [ITU-R BS.1387, 2001] qui a été utilisée. La métrique PEAQ note la dégradation entre un signal de référence et un signal codé, selon une échelle de dégradation Objective Difference Grade (ODG). La dégradation d'un échantillon est faite sur une échelle allant de $[-4, 0]$, plus la note est proche de 0 plus l'échantillon est de bonne qualité. L'implémentation utilisée est celle de Nikolaj Andersson [1].

Comme pour les tests précédents, une première évaluation a été réalisée avec les deux modèles entrainés avec une fonction de coût sans contrainte de débit ($\lambda = 0$). Ce test a permis d'estimer le score maximal que pouvaient obtenir les deux modèles. Le score PEAQ moyen a été calculé sur l'ensemble des échantillons du corpus de test. Le modèle avec espace latent simple a obtenu un score moyen de $-0{,}81$ avec un écart-type de $\pm 0{,}28$, le modèle enrichi de l'espace hyper-latent a obtenu un score de $-0{,}96$ avec un écart-type de $\pm 0{,}32$. Il est intéressant de noter que l'utilisation de la représentation hyper-latente ne semble pas avoir un impact sur la qualité de reconstruction, l'espace hyper-latent semble seulement avoir de l'influence sur le débit nécessaire pour transmettre l'espace latent.

Ce premier test a permis de connaitre la qualité audio maximale que pouvaient produire nos deux modèles. Un second test a été conduit, cette fois-ci avec différentes valeurs de $\lambda \sup 0$. Pour comparer les performances de nos modèles avec des codecs traditionnels, le codec MP3 et le codec Opus ont été utilisés. Le codec Opus a été utilisé dans sa version V1.3.1 dans son mode CBR avec une complexité maximale (10), le choix de la largeur de bandes du signal codé a été fait par le codec en fonction du débit, conformément à la RFC 6716 [Valin *et al.*, 2012]. Pour les méthodes de codage traditionnel, les débits utilisés ont été pour le codec Opus : $8, 16, 24{,}32{,}64, 92$

1. https://github.com/NikolajAndersson/PEAQ

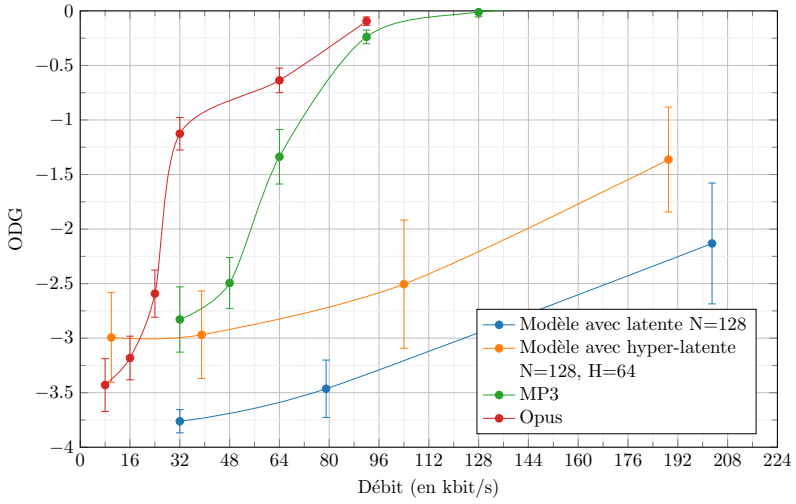


FIGURE 7.9 – Qualité audio PEAQ moyenne en fonction du débit pour le codec MP3, Opus, ainsi que les deux modèles de compression par réseau de neurones, le modèle avec latentes et celui enrichi des hyper-latentes. Pour le modèle avec latentes, le nombre de latentes est de $N = 128$. Pour le modèle avec hyper-latentes, le nombre de latentes $N = 128$ et d'hyper-latentes $M = 64$.

kbit/s, pour le codec MP3 : $16, 48, 64, 92, 128$ kbit/s. Pour les deux méthodes par compression de neurones λ a été fixé à $5 \times 10^1, 10^3, 10^4, 5 \times 10^5$.

La figure 7.9 montre le score ODG moyen et l'écart-type obtenu pour ce second test. La moyenne et l'écart-type sont calculés sur l'ensemble du corpus de test pour les différentes méthodes de codage. Une première observation est que le codec Opus, plus récent que le codec MP3, permet d'obtenir de meilleurs scores que ce dernier. Sur la gamme de débit 32-64 kibt/s, le codec Opus permet une compression de qualité équivalente pour un débit 2 fois moins important. Les méthodes de compression par réseau de neurones ont des scores inférieurs aux deux codecs traditionnels évalués. En observant la courbe, il est possible de remarquer que plus le débit augmente plus la qualité augmente, cependant la forme de la courbe est bien différente de la forme d'un codec traditionnel. Pour les modèles entrainés avec une forte contrainte sur le débit $\lambda = \{5 \times 10^1, 10^3\}$, le modèle avec hyper-latentes arrive toutefois à obtenir des performances proches, voire équivalentes, aux codecs Opus et MP3. Au débit le plus important ($\lambda = 5 \times 10^4$) le score obtenu pour le modèle avec espace hyper-latent est de $-1{,}36$, en comparaison, le score du modèle entrainé sans contrainte de débit avait obtenu un score de $-0{,}97$. Cela semble indiquer que pour des débits autour de 192 kbit/s, le modèle avec contrainte de débit s'approche du score maximal que peut produire le modèle. Plusieurs pistes peuvent expliquer les différences de résultats entre les modèles par réseau de neurones et les codecs traditionnels, ainsi que la forme de courbe.

Dans un but de compresser le signal tout en gardant les caractéristiques les plus importantes pour la perception humaine, les codecs traditionnels intègrent des modèles perceptifs. Ces modèles perceptifs permettent de faire des simplifications du signal et donc d'économiser du débit, sans perte de qualité pour l'auditeur. Il paraît normal que ces codecs traditionnels optimisés, selon des critères subjectifs, obtiennent de meilleurs scores avec une métrique comme PEAQ, là où

les méthodes par réseau de neurones avaient fait l'hypothèse que l'optimisation d'un critère de distance MSE était suffisant pour obtenir de bons résultats sur des métriques intégrant une prise en compte des aspects subjectifs dans sa notation. L'utilisation d'une métrique différente pour l'entrainement et pour l'évaluation pourrait également expliquer la forme de la courbe de débit-qualité. Nos modèles avec une contrainte de débit reproduisent les caractéristiques essentielles du signal important selon sa métrique. Cependant, ces caractéristiques peuvent ne pas être les critères les plus importants du point de la métrique perceptive, d'où une certaine stagnation du score à bas débit. La métrique PEAQ pourrait être utilisée comme métrique d'entrainement des modèles afin d'optimiser les réseaux selon des critères plus proches de la perception. Cependant, le score PEAQ demande un temps de calcul important, ce qui rend la métrique difficilement utilisable en l'état dans un réseau de neurones sans grandement augmenter la durée de l'entrainement. Une solution pourrait être de trouver une métrique reprenant les caractéristiques principales de la métrique PEAQ tout en restant relativement simple à calculer.

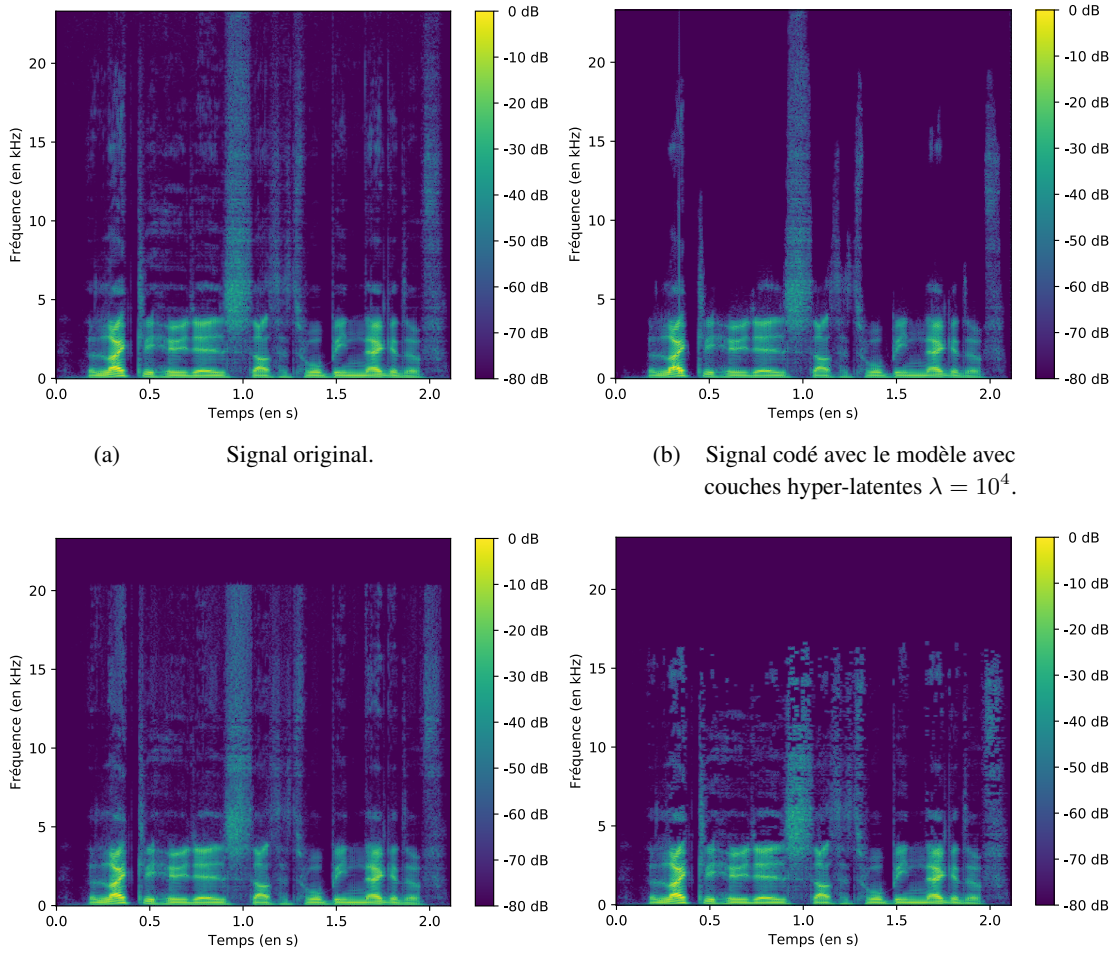
Une seconde raison qui pourrait expliquer les scores de nos modèles pourrait être la difficulté pour le réseau de reproduire les hautes fréquences. Lors d'écoutes informelles des échantillons décodés, selon la contrainte mise sur le débit λ, les hautes fréquences avaient tendance à être mal modélisées. Plus la contrainte est forte, moins les hautes fréquences sont bien modélisées, ce qui produit un son plus sourd que l'original. Pour les échantillons codés sans la contrainte de débit, les hautes fréquences sont correctement représentées.

Ces observations laissent penser que du point de vue de la fonction de coût, il est plus intéressant d'économiser du débit en ne modélisant pas les hautes fréquences, moins énergétiques, plutôt que de modéliser toutes les fréquences, mais en transmettant une grande quantité de données. De plus, plus la fréquence est élevée, plus le signal peut être assimilé à un processus stochastique. Un tel processus peut être difficile à reproduire pour un réseau de neurones sans investir une grande quantité d'informations, il serait donc avantageux pour le modèle de ne pas modéliser cette partie pour économiser du débit.

Une solution pourrait être de ne pas coder l'ensemble du spectre par une approche codage par forme d'onde (*wavform matching*), mais de faire une utilisation combinée d'un codage par forme d'onde et d'un codage paramétrique, comme dans les codecs modernes (Opus, HE-AAC…). La bande basse serait reproduite par un réseau de neurones qui aurait pour objectif de reproduire le signal au plus proche. Quant à la bande haute, un codage par reconstruction de bandes spectrales, ou SBR, pourrait être utilisé.

La figure 7.10 montre un comparatif des signaux codés avec les différentes méthodes. La transformée de Fourier à court terme du signal d'origine est représentée sur la figure 7.10(a). La figure 7.10(b) représente le signal compressé par le modèle avec représentation hyper-latente ($N = 128$, $M = 64$ et $\lambda = 10^4$). Par rapport au signal d'origine, il est possible d'observer que les détails en hautes fréquences ont été effacés, seuls les éléments les plus énergétiques ont été conservés. La figure 7.10(c) représente le signal compressé avec le codec Opus à 64 kbit/s. Contrairement au modèle par réseau de neurones, en hautes fréquences la structure fine du signal a été conservée. À ce débit, le signal est codé avec une représentation pleine bande (*fullband*) par le codec et même si le signal est échantillonné à 48 kHz, seule la bande $0 - 20$ kHz est reproduite. C'est pour cette raison que le signal au-dessus de 20 kHz n'est pas conservé. La figure 7.10(d), représente le signal compressé avec le codec MP3 à 64 kbit/s. Pour la même raison que pour

le codec Opus, seule une partie de la bande de fréquence de $0 - 16$ kHz a été codée. Dans les hautes fréquences, bien que la structure fine soit présente, certains détails ont étés plus fortement dégradés.



(a) Signal original.

(b) Signal codé avec le modèle avec couches hyper-latentes $\lambda = 10^4$.

(c) Signal codé avec Opus à 64 kbit/s

(d) Signal codé avec mp3 à 64 kbit/s

FIGURE 7.10 – Comparatif d'un signal codé par les différentes approches. Le signal est représenté selon la transformée de Fourier à court terme de 2^{10} points.

Comme cela avait été fait pour le modèle avec la représentation latente à la section 7.2.4.3, il est possible d'observer le débit moyen et instantané pour le même échantillon. Cet échantillon a été codé par le modèle avec représentation hyper-latente avec $\lambda = 10^3$. Le débit moyen sur la durée de l'échantillon est de 26,34 kbit/s. La figure 7.11 montre le débit instantané nécessaire pour transmettre la représentation de l'espace latent $R_{\mathbf{y}}$ et l'espace hyper-latent $R_{\mathbf{z}}$. Avec ce nouveau modèle, le débit global nécessaire pour compresser le signal d'entrée est moins important que pour le modèle précédent, 26,34 kbit/s avec le nouveau modèle contre 66,16 kbit/s avec le précédent. Par ailleurs, même si la forme de la courbe est globalement similaire, le débit instantané maximum $R_{\mathbf{y}}$ a été divisé par deux. Le débit instantané $R_{\mathbf{z}}$ quant à lui ne représente que 1 kbit/s. Contrairement au débit de $R_{\mathbf{y}}$, $R_{\mathbf{z}}$ est relativement constant au cours du temps.

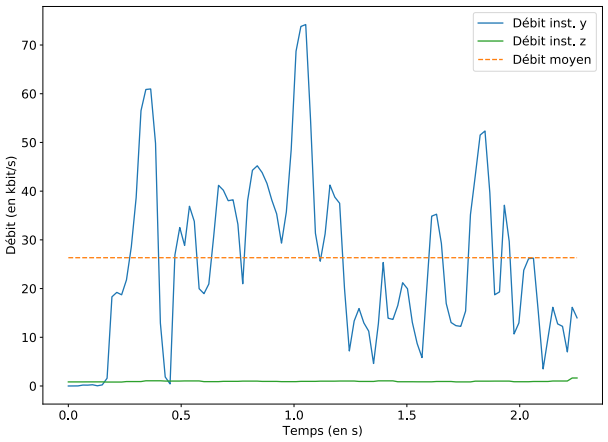


FIGURE 7.11 – Débit instantané nécessaire pour coder les cartes latentes $\mathbf{y}$ et hyper-latentes $\mathbf{z}$ pour un échantillon donné avec le modèle avec espace hyper-latent ($N = 128$, $M = 64$, $\lambda = 10^3$).

En observant l'activation des couches latentes lors de la compression d'un signal, il est possible de constater que toutes les latentes $\mathbf{y}_i$ ne sont pas activées en même temps. En fonction de l'échantillon à coder, seul un sous-ensemble des cartes latentes est activé. L'activation des couches a un mode de fonctionnement proche du tout ou rien : soit la couche a une activation nulle, soit la couche a une activation importante. Le niveau de cette activation est presque identique pour toutes les couches qui sont activées au même moment ce qui semble indiquer que chacune des latentes est spécialisée pour modéliser une caractéristique du signal. L'information transmise par les couches latentes est semblable à une information binaire, présente ou absente de cette caractéristique.

7.3.2.3 Influence du nombre de cartes hyper-latentes

Une analyse a été faite pour étudier l'influence du nombre de cartes d'activation M sur les performances du modèle avec espace hyper-latent. Pour étudier cet impact, nous avons entrainé plusieurs fois un même modèle avec différentes valeurs de M. Le modèle utilisé pour cette évaluation a un nombre de cartes d'activation pour la partie latente de $N = 128$ et une valeur de λ fixée à 10^3. Plusieurs versions de ce modèle ont été entrainées avec un nombre de cartes d'activation M différent, 4 modèles ont été créés $M = \{16, 32, 64, 128\}$. Pour chacun de ces modèles, plusieurs critères ont été évalués : le temps d'exécution, la qualité de compression selon les métriques MSE, SegSNR et PEAQ, le nombre de paramètres. Le tableau 7.3 fait un récapitulatif des résultats obtenus pour chacune des versions du modèle. Les performances indiquées correspondent à la performance moyenne et l'écart-type calculés sur l'ensemble du corpus de test. Le tableau montre que l'augmentation du nombre M de couches pour l'estimation de μ et σ n'a pas d'influence sur la qualité audio. Quelle que soit la valeur de M, les scores de qualité sont similaires pour toutes les métriques. Cela suggère que pour les modèles évalués, la valeur M n'est pas le facteur limitant pour la qualité audio, de plus, cela semble indiquer que la modélisation de l'espace latent peut être réalisée avec un espace hyper-latent d'une relativement petite dimension. Dans l'article [Ballé *et al.*, 2018], pour la compression d'image, les auteurs utilisaient pour leur modèle un nombre de

Tableau 7.3 – Qualité de reconstruction et temps d'exécution en fonction du nombre de cartes d'activation M de l'espace hyper-latent. Le nombre de cartes d'activation pour l'espace latent est fixé à $N = 128$ et $\lambda = 10^3$.

Nombre de cartes hyper-latentes	**M=16**	**M=32**	**M=64**	**M=128**
Nombre de paramètres	$2{,}53 \times 10^6$	$2{,}64 \times 10^6$	$3{,}02 \times 10^6$	$4{,}40 \times 10^6$
***MSE* moyen** ($\times 10^7$)	82,89 ($\pm$32,98)	82,97 ($\pm$33,06)	83,28 ($\pm$32,87)	82,24 ($\pm$32,84)
***SegSNR* moyen (en dB)**	17,01 ($\pm$2,00)	17,25 ($\pm$1,98)	17,09 ($\pm$2,00)	17,35 ($\pm$1,99)
***PEAQ* moyen**	$-3{,}21$ ($\pm$0,35)	$-3{,}43$ ($\pm$0,23)	$-2{,}97$ ($\pm$0,40)	$-3{,}25$ ($\pm$0,35)
Temps moyen d'execution (en ms)	76,53 ($\pm$10,22)	78,52 ($\pm$8,31)	79,84 ($\pm$10,03)	82,32 ($\pm$10,23)

couches M deux fois moins important que le nombre de couches N. Pour la compression audio, il semblerait que le nombre de couches M puisse être 4 fois moins important que le nombre de couches N, sans impacter la qualité.

L'augmentation de M a pour effet d'accroître le temps d'exécution du modèle. Cet accroissement reste, cependant, limité par rapport au temps total d'exécution. L'augmentation de M entraine également l'augmentation du nombre total de paramètres du modèle. La différence de temps d'exécution entre les deux modèles $M = 16$ et $M = 128$ est inférieure à 7,5 %, alors que le nombre de paramètres augmente de 70 % entre les deux. Dans le contexte de la compression, il peut donc être intéressant de favoriser l'utilisation d'une petite valeur de M.

7.4 Résumé et perspectives

Dans ce chapitre, nous avons exploré une nouvelle approche pour la compression audio par réseau de neurones. L'idée de cette nouvelle approche est d'utiliser un auto-encodeur variationnel pour transformer le signal audio vers une représentation qui nécessite un débit moins important pour être transmise.

La partie analyse de l'auto-encodeur transforme le signal d'entrée dans une représentation appelé espace latent. Cette représentation est codée et transmise à la partie synthèse de l'auto-encodeur. À partir de cette représentation latente, la partie synthèse génère le signal de sortie. Lors de l'entrainement, le réseau cherche à optimiser deux critères dans sa fonction de coût : la distorsion entre le signal d'origine et le signal compressé,et le débit théorique nécessaire pour transmettre la représentation latente. Avant d'être transmise, la représentation latente a besoin d'être quantifiée. Cette quantification est une opération continue par morceau, ce qui représente un obstacle pour l'entrainement du réseau car toutes les opérations dans le réseau doivent être dérivables

pour permettre la rétropropagation du gradient. Pour cela, nous nous sommes appuyés sur les travaux présentés dans [Ballé *et al.*, 2017], où les auteurs proposent d'approximer la quantification en ajoutant un bruit uniforme à la représentation de l'espace latent. Cette approximation est substituée à la quantification réelle lors de la phase d'entrainement ce qui permet la rétropropagation du gradient dans le réseau.

Un premier modèle basé sur une architecture de compression d'images [Ballé *et al.*, 2017] a été réalisé. Le réseau prend comme données d'entrée la MDCT du signal. Pour contraindre la plage des valeurs de la MDCT et donner plus d'importance aux parties du spectre les moins énergétiques, une compression de la plage de valeur est effectuée. Des tests ont été réalisés pour évaluer les performances de notre modèle pour la compression de parole. Pour ces tests, notre méthode a été comparé au codec MP3. La qualité de reconstruction a été mesurée selon deux métriques objectives : la MSE et le SegSNR. Les résultats ont montré qu'à débit équivalent notre modèle ne permettait pas d'obtenir une qualité équivalente au codec MP3. Ce test a permis de montrer la faisabilité de la compression par réseau de neurones, que ce soit en terme de temps d'exécution ou d'occupation mémoire.

À partir de ce premier modèle, un second modèle a été réalisé. Ce nouveau modèle propose une version enrichie du premier modèle. Un module est ajouté pour modéliser plus finement la distribution de chaque élément de l'espace latent. Cette modélisation, appelée représentation hyper-latente, permet de faire une prédiction des valeurs de l'espace latent dans le but de réduire le débit nécessaire pour les coder. Une première évaluation de ce nouveau modèle a montré qu'à débit équivalent, la qualité de la reconstruction, selon les deux métriques utilisées, était similaire au codec MP3. De plus, le modèle avec représentation hyper-latente permet d'obtenir le même niveau de reconstruction que le modèle avec latentes simples pour un débit deux fois moins important. Une seconde évaluation de performance a été réalisée selon la métrique PEAQ. Cette métrique permet de prendre en compte l'aspect perceptif du son et donc permet une mesure plus fidèle de la qualité subjective. Ce test a comparé les performances de nos deux modèles avec celles des codecs Opus et MP3. Les résultats du test ont montré que le modèle avec représentation hyper-latente permet d'obtenir une qualité comparable à celle de Opus et MP3 pour un débit inférieur à 32 kbit/s. Cependant, pour les débits plus importants les codecs traditionnels obtiennent une qualité supérieure à nos modèles.

La compression audio par réseau de neurones est un champ de recherche très récent et de nombreuses pistes restent à explorer. Une des premières pistes de recherche serait d'intégrer une prise en compte de l'aspect perceptif dans la fonction de coût. Comme dans [Ballé *et al.*, 2017], nous avons fait l'hypothèse que l'optimisation d'un tel critère de distance serait suffisant pour obtenir une bonne qualité de compression subjective. Cependant, les résultats des différents tests ont montré que cette hypothèse n'était pas suffisante pour produire des signaux d'une qualité équivalente aux codecs actuels. Pourtant, nos modèles ont montré de bons résultats pour reproduire le signal selon une métrique de distance du type SNR, ce qui laisse penser qu'il serait possible d'obtenir de meilleurs résultats subjectifs en ajoutant une prise en compte de critères perceptifs dans la fonction de coût lors de l'entrainement. Récemment, l'article [Zhen *et al.*, 2020] semble avoir réussi à intégrer une partie du modèle perceptif utilisé dans le codec MP3, cela a permis un gain important de qualité audio subjective.

Une seconde piste serait le changement de représentation du signal d'entrée. Dans nos mo-

dèles, le choix de l'utilisation de la MDCT a été fait pour deux raisons. La première raison a été de fournir des données d'entrées de dimension similaire à celle utilisée dans le modèle de compression d'image. La seconde raison a été de permettre une intégration plus facile du module dans un codec traditionnel. Cependant, l'utilisation d'une transformée entraine des limitations [Caracalla et Roebel, 2020]. Dans le cadre de cette thèse, des premières expérimentations ont été réalisées pour adapter le modèle et lui permettre de travailler directement dans le domaine temporel avec le signal brut. Ce modèle temporel a montré de bonnes performances pour modéliser des sons harmoniques mais il se heurte à la difficulté de modéliser les transitoires du signal. Les transitoires sont reproduites avec une forte distorsion harmonique. Cette distorsion laisse penser que les filtres successifs du modèle ont des difficultés pour produire des impulsions précises. C'est actuellement le plus gros frein à l'utilisation de notre modèle avec des signaux temporels. Des recherches pourraient être menées dans le but d'améliorer la modélisation de ces transitoires par notre réseau.

Dans une vision à plus long terme, le modèle devrait être adapté pour pouvoir travailler sur des signaux découpés en trames de 20 ms comme cela est fait pour les codecs audio conversationnels traditionnels. Le découpage en trame du signal peut déjà être utilisé avec le modèle actuel, malheureusement ce découpage s'accompagne d'une perte de qualité significative. En observant le contenu des trames, il est possible de constater que le contenu de deux trames consécutives partage de nombreuses similarités. De plus, un recouvrement est généralement présent entre elles, ce qui augmente la redondance entre deux trames. Coder de manière indépendante chaque trame serait donc une stratégie sous-optimale. Un module pour capter le contexte pourrait être ajouté directement au modèle pour prédire le contenu de la trame suivante. Le domaine du codage vidéo par réseau de neurones pourrait servir de source d'inspiration pour résoudre ce problème. Dans l'approche [Ladune *et al.*, 2021], une partie du réseau est utilisé pour faire une prédiction de l'activation de l'espace latent de l'image courante par rapport à celui de l'image précédente. Seule l'erreur de prédiction doit être transmise d'une trame à l'autre.

Une dernière perspective serait d'avoir différents jeux de poids pour un même modèle. Chaque jeu correspondrait à un entrainement spécifique sur un type de signaux particuliers : musique, voix grave, voix aiguë, bruit... En fonction du signal à coder, il serait envisageable de changer le jeu de poids utilisé dans le modèle. Cet ajustement du réseau selon le type de contenu permettrait une meilleure qualité de codage globale. Un tel système pourrait se rapprocher du choix de coder le signal d'entrée par un codage musique ou parole fait par les codecs traditionnels. Un module de décision sur le principe de la détection d'activité vocale, ou VAD, devrait être ajouté au modèle existant.

Conclusion

Contributions de la thèse

Au cours de cette thèse, nous avons exploré différentes approches pour la compression des signaux ambisoniques de premier ordre (FOA).
Dans la première contribution, décrite au chapitre 4, nous avons étudié le fonctionnement de l'approche multimono. Pour cela, un test subjectif a été mené pour étudier les altérations apportées par la méthode. En plus des artefacts fréquentiels, le test subjectif nous a permis de constater l'apparition d'artefacts spatiaux. La cause de ces artefacts spatiaux semble être les déformations non homogènes des composantes de la base ambisonique par les multiples instances du codec cœur mono. Ces déformations sont différentes pour chaque composante et produisent des artefacts de plusieurs natures : source fantôme, ajout de bruit diffus… À haut débit, les altérations du signal apportées par le codage cœur sont relativement faibles, ce qui semble rendre cette déformation moins perceptible. Cependant à bas débit, ces déformations sont bien plus audibles et semblent impacter la qualité audio plus fortement.
Notre première méthode propose une extension de l'approche multimono pour améliorer la qualité spatiale. Pour cela, notre méthode se base sur la mise en œuvre d'un rematriçage des composantes en amont du codage multimono. Les matrices de transformation sont calculées à partir d'analyse en composante principale (PCA). Pour garantir la continuité du signal entre les trames, une interpolation des matrices de transformation est réalisée. Pour garantir une interpolation à vitesse constante, cette interpolation est réalisée dans le domaine des quaternions. Les tests subjectifs MUSHRA ont montré que notre méthode permet d'améliorer significativement la qualité globale des signaux codés. À l'écoute, il est possible de noter que les artefacts spatiaux ne sont plus audibles.

Dans la seconde partie de nos travaux, nous avons créé deux méthodes de codage utilisant l'information spatiale pour le codage des signaux. L'information spatiale est extraite du signal en calculant la cartographie de l'énergie du signal. Cette cartographie peut être représentée par la matrice de covariance du signal ambisonique. Une méthode de post-traitement, décrite au chapitre 5, a tout d'abord été présentée. Cette méthode propose de corriger la déformation spatiale du signal apportée par l'approche multimono. Des tests subjectifs *RefAB* ont montré que notre méthode permettait d'améliorer en moyenne la qualité globale par rapport à la méthode multimono sans le traitement. Dans ces tests, deux codecs cœur ont été utilisés, *EVS* et *Opus*. Pour analyser plus en détail les résultats des tests et mieux comprendre les intéractions entre les différents facteurs et les résultats, une analyse *ANOVA* a été réalisée. Cette analyse a permis de montrer que les résultats n'étaient pas dépendants du codec utilisé, ce qui semble indiquer que ce post-traitement pourrait être utilisé pour d'autres codecs cœur. Dans la suite de ce chapitre, une expérimentation a été conduite pour utiliser le post-traitement, non pas pour avec une approche multimono mais pour une approche multistéréo.
Par la suite le principe de correction spatiale, étudié dans la méthode de post-traitement, a été adapté pour être utilisé dans un codage paramétrique. Pour cette méthode, décrite au chapitre 6, seule la composante omnidirectionnelle est transmise. À partir de cette composante et de l'information spatiale, la spatialisation du signal d'origine est reproduite. Des tests subjectifs MUSHRA

ont été réalisés, notre méthode a été comparée à deux autres approches : l'approche multimono et l'approche paramétrique DirAC. Les résultats ont montré que notre méthode obtenait des résultats significativement meilleurs que les autres méthodes évaluées sur une gamme de débits allant de 48 et 64 kbit/s. Une étude comparative a été réalisée pour étudier les différences et les points communs de la méthode DirAC. Cette étude a permis de fournir les premiers éléments qui pourraient expliquer les performances obtenues par notre méthode par rapport à la méthode DirAC.

Dans la dernière partie de ce manuscrit, chapitre 7, nous avons tenté de remplacer le codec cœur mono par un réseau de neurones. Le modèle étudié est basé sur un modèle de compression d'image par auto-encodeur variationnel. Ce modèle a été adapté pour permettre de traiter du signal audio mono. Une transformée en cosinus (MDCT) a été utilisée comme représentation temps-fréquence des signaux d'entrée. Pour limiter la dynamique des valeurs et donner une importance plus grande aux fréquences de faible intensité, une compression de l'amplitude des valeurs a été effectuée. Des tests objectifs ont été réalisés pour comparer ce premier modèle avec le codec MP3. Les résultats ont montré que selon la métrique MSE utilisée pour l'entrainement et la métrique SegSNR, notre modèle arrive à des résultats comparables à MP3. L'analyse des cartes d'activation a montré une forte corrélation entre le spectre MDCT du signal d'entrée et l'activation de la couche latente, ce qui indique que le modèle n'arrive pas à capturer la totalité de la structure du signal. Par la suite, une extension du premier modèle a été proposée pour lui permettre de capturer la structure du signal. Pour capturer cette corrélation, le modèle a été étendu par un deuxième VAE chargé de modéliser la distribution des valeurs latentes. Cette amélioration a permis de diminuer grandement le débit nécessaire pour coder le signal. Des tests objectifs ont été réalisés pour évaluer les performances de ce nouveau modèle. Notre modèle a été comparé avec deux codecs mono MP3 et Opus. Pour ces tests, en plus de la MSE et de SegSNR, la métrique objective PEAQ a été utilisée. Cette métrique permet de prendre en compte la perception dans la notation de la qualité. Pour les deux premières métriques, notre modèle a obtenu des résultats supérieurs aux codecs traditionnels. Cependant, pour la métrique PEAQ, nos modèles ont obtenu des résultats inférieurs à ceux des deux autres codecs. Nos modèles quoi que relativement rudimentaires posent les premières fondations pour de nouveaux modèles plus élaborés.

Perspectives de recherche

Les différents travaux présentés dans cette thèse peuvent appeler à divers développements afin de les améliorer et lever une partie des limitations rencontrées.
Pour la première méthode de compression par décorrélation des composantes, l'observation des signaux décorrélés avant codage a montré que les matrices produites par l'analyse PCA pouvaient conduire à des interversions du contenu des canaux. D'une trame à l'autre, un élément présent dans un canal pouvait être déplacé dans un autre canal. Ce déplacement pouvait provoquer des discontinuités du signal et fortement impacter la qualité. Dans notre méthode, un réordonnancement des canaux matricés a été mis en place. Ce réordonnancement est la continuité de l'énergie du signal d'une trame à l'autre. Une piste de recherche pourrait être d'améliorer le mécanisme de réordonnancement en analysant le contenu du signal de chaque canal et non pas seulement sur l'évolution de l'énergie contenue dans les canaux.
Par ailleurs, la méthode présentée se limite aux signaux ambisoniques d'ordre 1. Cette limitation

est présente car la décomposition en doubles quaternions utilisée pour l'interpolation des matrices de transformation n'est possible que pour des matrices de rotation de dimension 4×4. Une seconde piste de recherche pourrait être d'étendre la méthode aux ordres supérieurs HOA, grâce à une adaptation de l'interpolation par quaternion pour permettre une interpolation pour des matrices de plus grande dimension.

Pour la méthode de post-traitement par correction spatiale, l'analyse *ANOVA* a montré que les résultats obtenus étaient indépendants du codec utilisé. Cette analyse a également permis de constater que les performances du post-traitement avaient une forte dépendance au contenu du signal ambisonique codé. Des écoutes informelles semblent suggérer que pour les signaux avec de nombreux sons percussifs, la méthode semble donner une meilleure qualité quand le nombre d'images spatiales par seconde est important. Au contraire, pour les signaux avec de nombreuses sources, la méthode semble donner de meilleurs résultats quand le nombre de découpes en sous-bande est important. Une piste d'amélioration pourrait être d'adapter les paramètres de la méthode, c'est-à-dire le nombre d'images spatiales transmis par seconde et nombre de sous-bandes en fonction du signal d'entrée par le codeur. Une analyse du contenu de la scène sonore d'entrée pourrait permettre de piloter ces paramètres. Le même type d'adaptation des paramètres pourrait également être étendu à la méthode de codage paramétrique par *upmix*.
Lors des expérimentations conduites sur l'usage du post-traitement avec l'approche multistéréo proposé dans le mode ambisonique d'Opus, nous avons pu constater que la méthode n'arrive pas à corriger la déformation spatiale du signal. Le problème semble venir de la manière dont le codec cœur stéréo code la bande haute. Le codec stéréo spatialise le signal par un panoramique d'intensité, ce qui produirait une image spatiale trop éloignée de l'image spatiale d'origine pour être corrigée. Pour retrouver l'image spatiale d'origine, les valeurs de la matrice de correction se retrouvent à amplifier fortement certaines zones de l'espace peu énergétiques dans le signal décodé. Cette forte amplification peut entrainer des dégradations de la qualité perçue (augmentation du bruit de fond, création d'artefacts…). Un raffinement semble nécessaire pour utiliser le post-traitement avec des codecs cœur stéréo. Une première piste pourrait être de limiter le gain apporté par la matrice de correction pour la bande haute.

De manière générale, dans cette thèse, nous nous sommes concentrés sur des approches utilisant des codecs mono ou stéréo comme brique de base, en ajoutant des traitements avant et après ces codecs pour leur permettre de coder des signaux ambisoniques. Un autre axe de recherche, maintenant, pourrait être de se permettre de modifier ces codecs cœur. Nous avons vu que dans l'approche multimono, les artefacts spatiaux étaient, en partie, dus aux altérations apportées par le codage indépendant des composantes par le codec cœur. Chaque codec fait une sélection de paramètres et de mode de codage en fonction de l'analyse d'une composante en entrée. Cette sélection maximise la qualité subjective du signal codé, elle est optimale de manière locale cependant il n'est pas garanti que cela soit la décision optimale de manière globale. En permettant de modifier les codecs cœur, il devient possible d'avoir un mécanisme de décision commun entre tous les codecs. Une décision commune devrait permettre de choisir les paramètres et le mode maximisant la qualité audio globale. De plus, comme chaque codec cœur applique les mêmes modes et paramètres, les altérations apportées sont plus homogènes entre les composantes, ce qui devrait limiter les artefacts spatiaux.

Dans la dernière partie du manuscrit, nous avons présenté une approche de codage mono par réseaux de neurones. À partir de la même approche, deux modèles ont été proposés. Lors de la

comparaison des performances de ces deux modèles par rapport aux codecs Opus et MP3, les résultats obtenus selon la métrique PEAQ sont inférieurs à ceux des codecs traditionnels. Ces résultats semblent pouvoir s'expliquer par le fait que notre modèle n'intègre pas d'aspect perceptuel dans sa fonction de coût. Il semblerait pertinent d'effectuer un raffinement de la fonction de coût pour y intégrer une prise en compte de la perception.
Par ailleurs, les questions de traitement par trame du signal et le retard induit par la méthode n'ont pas été adressés dans ces travaux exploratoires, une seconde piste de recherche pourrait être de trouver comment intégrer ces questions dans le modèle actuel.

Publications

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

Articles de conférence :

— P. Mahé, S. Ragot, S. Marchand et J. Daniel (2021). Ambisonic coding with spatial image correction. In *European Signal Processing Conference (EUSIPCO)*.

— P. Mahé, S. Ragot et S. Marchand (2019). First-Order Ambisonic Coding with Quaternion-based interpolation of PCA rotation matrices. In *EAA Spatial Audio Signal Processing Symposium*.

— P. Mahé, S. Ragot et S. Marchand (2019). First-Order Ambisonic Coding with PCA Matrixing and Quaternion-Based Interpolation. In *International Conference on Digital Audio Effects (DAFx)*.

Présentation en congrès et poster :

— P. Mahé, S. Ragot et S. Marchand (2019). Codage Ambisonique pour les Communications Immersives. In *Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio (JJCASS)*.

Dépôt de demande de brevet :

— S. Ragot et P. Mahé (2020). Codage optimisé d'une information représentative d'une image spatiale d'un signal audio multicanal. *Office européen des brevets*.

— P. Mahé, S. Ragot et J. Daniel (2019). Détermination de corrections à appliquer à un signal audio multicanal, codage et décodage associé. Dépôt de brevet (FR3101741A1), *Office européen des brevets*.

— S. Ragot et P. Mahé (2019). Système de codage audio spatialisé avec interpolation et quantification de la rotation. Dépôt de brevet (EP3706119A1), *Office européen des brevets*.

Annexes

ANNEXE A

Listes des échantillons de test

A.1 Description des échantillons utilisés

Applause - Prise de son réelle

Enregistrement d'applaudissements à la fin d'un concert. La captation a été faite par un micro FOA. Le microphone est placé au milieu du public. L'enregistrement est issu de la base audio Ambisonia [Ambisonia, 2005].

Castanets - Spatialisation de synthèse

Spatialisation d'un enregistrement mono de castagnettes dans plusieurs positions de l'espace. La spatialisation a été faite à l'aide d'un *framework Matlab* interne. Une légère réverbération est présente, cette réverbération est faite par un simulateur de salle type *Shoebox* avec la méthode source image. L'enregistrement de castagnettes est issu de la base audio EBU SQAM [EBU - TECH 3253, 2008].

Drums - Spatialisation de synthèse

Enregistrement multipiste d'une batterie mixée en FOA. Chaque piste correspond à un élément de la batterie (caisse claire, grosse caisse, cymbales...), ces éléments sont spatialisés autour de l'auditeur. Ce mixage comprend un effet de spatialisation ainsi que des effets de champ proche. La spatialisation et les effets de champ proche ont été réalisés à l'aide d'un *framework Matlab* interne. Cet échantillon est une version raccourcie de l'échantillon *H_02_Drums1* généré par Orange Labs dans le cadre de la normalisation de MPEG-H [MPEG, 2013].

Helicopter - Prise de son réelle

Enregistrement du vol d'un hélicoptère. L'enregistrement a été fait par un micro FOA *SoundField SPS200*. L'hélicoptère arrive de la droite avant de s'éloigner par la gauche. L'enregistrement est issu de la base audio Ambisonia [Ambisonia, 2005].

Little-Prince - Prise de son réelle

Enregistrement d'un homme qui lit un extrait du livre "Le Petit Prince" en français, a une position fixe dans une salle de type salon. L'enregistrement se veut au plus proche d'un contexte domestique réel. L'enregistrement a été réalisé par un Microphone *EigenMike*.

Modern - Spatialisation de synthèse

Spatialisation d'une source mono qui passe de gauche à droite. Le déplacement de la source est accompagné d'un effet de champ proche. La source est un extrait de chanson de jazz de *"The Present"* de Laurent de Wilde. La spatialisation et l'effet de champ proche ont été réalisés à l'aide d'un *framework Matlab* interne.

Nature - Spatialisation de synthèse

Simulation d'une scène en pleine nature. Le bruit d'un bourdon se déplace dans la scène sonore. Le bourdon est mixé avec le chant d'un oiseau positionné dans une direction fixe et l'écoulement d'un ruisseau. Les enregistrements du bourdon et de l'oiseau sont des signaux mono. Le ruisseau est un enregistrement FOA. Le mixage a été fait avec le logiciel Reaper. La projection des sources mono dans le domaine ambisonique a été réalisé avec l'aide du *framework Ambix* [Kronlachner, 2014a].

Noise - Spatialisation de synthèse

Bruit rose mono spatialisé artificiellement pour faire le tour de l'auditeur selon le plan horizontal spatialisé de manière synthétique. La rotation complète est faite en 10 secondes. Cet échantillon est *"H_13_NoiseSingle"* généré par Orange Labs dans le cadre de la normalisation de MPEG-H [MPEG, 2013].

Opera - Prise de son réelle

Enregistrement d'une cantatrice accompagnée d'un clavecin et d'un orchestre à cordes dans une salle de concert. La cantatrice est sur la scène, le microphone est placé dans le public devant la scène. L'enregistrement FOA a été réalisé avec un micro *SoundField SPS200*. L'enregistrement est issu de la base audio Ambisonia [Ambisonia, 2005].

Orchestra - Prise de son réelle

Enregistrement d'un orchestre symphonique dans une salle de spectacle. Le microphone *EigenMike* est placé au milieu de l'orchestre. L'enregistrement HOA réalisé par le micro est ensuite tronqué à l'ordre 1. L'enregistrement a été effectué par Orange Labs dans le cadre du projet Bili [Edwige, 2014].

Radio - Prise de son réelle

Enregistrement d'un ensemble de guitares classiques qui joue dans une salle avec réverbérant modéré. Le micro *EigenMike* est placé en face de l'ensemble. Cet échantillon correspond à l'échantillon *"H_12_Radio2"* proposé par Orange Labs dans le cadre la normalisation de MPEG-H [MPEG, 2013].

ShoutingAudience - Prise de son réelle

Enregistrement d'une foule encourageant une équipe sportive, il y a des acclamations ainsi que des applaudissements. Le microphone *EigenMike* est placé au milieu de la foule.

Stadium - Prise de son réelle

Enregistrement d'un public dans un stade durant une manifestation sportive. L'enregistrement est composé d'applaudissements et d'instruments de percussion (type tambour...). Le micro *EigenMike* est placé dans les gradins au milieu du public. Cet échantillon correspond à l'échantillon *"H_04_Stadium2"* proposé par Orange Labs dans le cadre la normalisation de MPEG-H [MPEG, 2013].

Talks - Prise de son réelle

Enregistrement d'un groupe de personnes, femmes et hommes, dispersés dans l'espace. Plusieurs personnes parlent simultanément. Les dimensions de la salle sont importantes, type salle de conférence. L'enregistrement a été réalisé par un microphone *EigenMike*.

Theater - Prise de son réelle

Enregistrement ambisonique de trois acteurs qui jouent une pièce de théâtre dans une salle type salle de spectacles. Le microphone *EigenMike* est devant la scène. Un des acteurs est situé à proximité du microphone (moins d'un mètre du micro). Les deux autres sont situés en champ lointain. L'enregistrement a été fait durant le projet Bili [Edwige, 2014].

Voices-Kids - Spatialisation de synthèse

Spatialisation de deux personnes, une femme et un homme, qui discutent au milieu d'un brouhaha produit par des enfants dans une cour d'école. L'enregistrement de chacune des personnes est un enregistrement mono spatialisé, l'un à gauche l'autre à droite de l'auditeur. Sur la première partie de l'enregistrement, les personnes se parlent l'une après l'autre. Sur la seconde partie, les deux interlocuteurs parlent simultanément. L'enregistrement du brouhaha est une captation réelle FOA provenant de la base Ambisonia [Kronlachner, 2014a].

Water - Prise de son réelle

Enregistrement d'une personne qui verse de l'eau dans plusieurs verres. Les verres sont repartis dans diverses positions de l'espace. Plusieurs personnes parlent au loin. Le microphone *EigenMike* est au centre de la scène. L'enregistrement HOA réalisé par le micro est ensuite tronqué à l'ordre 1.

A.2 Utilisation des échantillons pour les différents tests

La tableau A.1 dresse le récapitulatif des échantillons utilisés pour les différents tests menés dans cette thèse. Les échantillons utilisés pour les tests ont été amenés à évoluer d'un test à l'autre en fonction des retours et des observations faites lors des premiers tests. Certains échantillons jugés redondants ou trop peu discriminant ont été écartés, certains autres possédant des caractéristiques particulières, ont eux été ajoutés.

Tableau A.1 – Échantillons utilisés pour chacun des tests subjectifs.

	Test MUSHRA multimono (section 4.1.1)	**Test MUSHRA PCA (section 4.3)**	**Test RefAB correction spatiale (section 5.3)**	**Test MUSHRA upmix (section 6.2)**
Applause	✗	✗	✗	✗
Castanets			✗	✗
Drums	✗	✗	✗	✗
Helicopter	✗			
Little-Prince		✗	✗	✗
Modern	✗	✗	✗	
Nature	✗	✗	✗	✗
Noise	✗		✗	
Opera	✗	✗		✗
Orchestra	✗	✗	✗	
Radio	✗			
Shouting Audience	✗			
Stadium	✗			
Talks		✗		✗
Theater		✗	✗	✗
Voices-Kids		✗	✗	✗
Water	✗			
Nombre d'échantillons pour le test	12	10	10	9

ANNEXE B

Calcul de la cartographie à partir de la matrice de covariance

---

On pose $s_i[l]$ le signal provenant de la direction i. Cette direction peut être codée dans le domaine ambisonique par un vecteur $\mathbf{d_i}$, où l est l'index de l'échantillon dans une trame de taille $\mathbf{L} = [1, \ldots, L]$. Le signal $s_i[l]$, provenant de la direction i, peut être calculé par :

$$s_i[l] = \mathbf{d}_i^T \cdot \mathbf{b}[l] \tag{B.1}$$

avec $\mathbf{b}[l] = [b_1[l], \ldots, b_N[l]]$ le vecteur contenant les N composantes du signal ambisonique. Et $\mathbf{d}_i$, le vecteur contenant les N coefficients de décodage ambisonique, obtenus par la méthode du réencodage du signal ambisonique [Ward et Abhayapala, 2001]. Il est possible de réécrire l'équation sous la formule :

$$s_i[l] = \sum_{n=1}^{N} d_i[n] b[n][l] \tag{B.2}$$

L'énergie du signal s_i provenant de la direction i peut donc être calculée par :

$$E_i = \|s_i\|^2 = \sum_{l=1}^{L} S_i[l]^2 \tag{B.3}$$

Par ailleurs, la covariance $\mathbf{C}$ du signal ambisonique $\mathbf{B}$, pour une trame donnée, est définie comme :

$$\begin{aligned} \mathbf{C} &= \mathbf{B}\mathbf{B}^T \\ \mathbf{C}[j][k] &= \sum_{l=1}^{L} b[j][l] b[k][l] \end{aligned} \tag{B.4}$$

avec $j = [1, \ldots, N]$ et $k = [1, \ldots, N]$ les deux indices qui permettent d'accéder respectivement aux lignes et aux colonnes de la matrice de covariance $\mathbf{C}$ de dimension $N \times N$.
Dans l'équation (B.3), on pose $n' = n$:

$$\begin{aligned} \|s_i\|^2 &= \sum_{l=1}^{L} \sum_{n=1}^{N} (d_i[n] b_i[l][n]) \sum_{n'=1}^{N} (d_i[n'] b_i[n'][t]) \\ &= \sum_{n=1}^{N} \sum_{n'=1}^{N} d_i[n] d_i[n'] s \underbrace{\sum_{l=1}^{L} b_i[n][l] b_i[n'][l]}_{C[n][n']} \\ &= \mathbf{d}_i \mathbf{C} \mathbf{d}_i \end{aligned} \tag{B.5}$$

Il est donc possible de calculer l'énergie d'une direction quelconque i grâce du vecteur d'encodage $\mathbf{d}_i$ et de la matrice de covariance $\mathbf{C}$ du signal ambisonique. La valeur de $\mathbf{d}_i$ est indépendante du contenu du signal ambisonique, seule la matrice $\mathbf{C}$ varie en fonction du contenu de la trame audio. À partir de cette matrice, l'énergie du signal peut donc être calculée a posteriori pour tout point de l'espace.

ANNEXE C

Calcul du banc de filtres selon l'échelle Mel

L'échelle de Mel est une échelle pour représenter les fréquences selon une échelle logarithmique. Cette représentation permet d'avoir une échelle proche de la perception humaine des hauteurs des sons. La formule de conversion de l'échelle des fréquences à l'échelle de Mel est :

$$M(f) = 1125 \times \ln(1 + \frac{f}{700}) \tag{C.1}$$

où f est la fréquence d'origine et M la fonction qui permet d'obtenir la valeur m de la fréquence selon l'échelle de Mel. L'opération inverse peut être obtenue par la formule :

$$M^{-1}(m) = 700 \times (e^{\frac{m}{1125}} - 1) \tag{C.2}$$

À partir de cette échelle, il est possible de créer un banc de filtres. L'utilisation de l'échelle de Mel permet d'avoir la même différence perceptive de hauteurs entre deux filtres consécutifs du banc de filtres. Pour créer un banc de I filtres, les bornes fréquentielles minimale et maximale doivent être choisies. Pour nos expérimentations, nous avons utilisé $f_1 = 20$ Hz et $f_{I+1} = 24$ kHz. Ces deux fréquences sont converties dans l'échelle de Mel, dans notre cas : $m_{min} = 31{,}69$ et $m_{max} = 4008{,}91$. Selon le nombre de filtre I choisi pour le banc de filtres, les $I - 1$ valeurs sont régulièrement espacées sur l'intervalle $[m_1, m_{I+1}]$. Les deux valeurs extrêmes m_1 et m_{I+1} sont ajoutées aux $I - 1$ valeurs pour former les $I + 1$ bornes des filtres. Grâce à l'équation (C.2), les valeurs des bornes m_i sont converties vers l'échelle fréquentielle pour obtenir les bornes. Les filtres du banc sont calculés dans le domaine de la transformée de Fourier. Les bornes sont converties de l'échelle de Mel vers l'échelle de fréquence dans le domaine de Fourier. Les bornes sont arrondies vers le point le plus proche de la transformée pour obtenir les f_i. A partir des bornes f_i, les filtres H_i pour extraire une bande de fréquence du signal d'origine sont calculés, selon la formule :

$$H_i(k) = \begin{cases} 0 & k < f_{i-1} \\ \frac{k - f_{i-1}}{f_i - f_{i-1}} & f_{i-1} \leq k \leq f_i \\ \frac{f_{i+1} - k}{f_{i+1} - f_i} & f_i \leq k \leq f_{i+1} \\ 0 & k > f_{i+1} \end{cases} \tag{C.3}$$

où $k = [1, \ldots, K]$ représente l'indice d'une raie de la transformée de Fourier.

Bibliographie

[3GPP TR 26.118, 2018] 3GPP TR 26.118 (2018). S4-180975, pCR to 26.118 on Dolby VRStream audio profile candidate. (Cité en page 44.)

[3GPP TR 26.918, 2018] 3GPP TR 26.918 (2018). Virtual Reality (VR) media services over 3GPP. (Cité en page 55.)

[3GPP TR 26.997, 2019] 3GPP TR 26.997 (2019). IVAS codec performance characterization. (Cité en page 1.)

[3GPP TS 26.401, 2007] 3GPP TS 26.401 (2007). General audio codec audio processing functions ; enhanced aacplus general audio codec. (Cité en pages 39 et 40.)

[3GPP TS 26.445, 2019] 3GPP TS 26.445 (2019). Codec for Enhanced Voice Services (EVS) ; Detailed Algorithmic description. (Cité en pages 61 et 81.)

[3GPP TS 26.818, 2018] 3GPP TS 26.818 (2018). Virtual reality (vr) streaming audio ; characterization test results. (Cité en page 52.)

[3GPP TS 26.918, 2018] 3GPP TS 26.918 (2018). Virtual reality media services over 3gpp. (Cité en pages 15, 42, 43, 53 et 58.)

[3GPP TS 26.952, 2019] 3GPP TS 26.952 (2019). Codec for Enhanced Voice Services (EVS) ; Performance Characterization. (Cité en page 66.)

[Adavanne *et al.*, 2018] S. ADAVANNE, A. POLITIS et T. VIRTANEN (2018). Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. *In European Signal Processing Conference (EUSIPCO)*. (Cité en page 74.)

[Ahonen et Pulkki, 2009] J. AHONEN et V. PULKKI (2009). Diffuseness estimation using temporal variation of intensity vectors. *In Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 285–288. IEEE. (Cité en page 45.)

[Ambisonia, 2005] AMBISONIA (2005). www.ambisonia.com. (Cité en pages 153 et 154.)

[Arbogast *et al.*, 2005] T. L. ARBOGAST, C. R. MASON et G. KIDD JR (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America (JASA)*, 117(4):2169–2180. (Cité en page 27.)

[Armstrong *et al.*, 2018] C. ARMSTRONG, L. THRESH, D. MURPHY et G. KEARNEY (2018). A perceptual evaluation of individual and non-individual hrtfs : A case study of the sadie ii database. *Applied Sciences*, 8(11). (Cité en page 13.)

[Ballé *et al.*, 2017] J. BALLÉ, V. LAPARRA et E. P. SIMONCELLI (2017). End-to-end optimized image compression. *In International Conference on Learning Representations (ICLR)*. (Cité en pages 3, 119, 120, 122, 124, 127, 129, 132, 134 et 145.)

[Ballé *et al.*, 2018] J. BALLÉ, D. MINNEN, S. SINGH, S. J. HWANG et N. JOHNSTON (2018). Variational image compression with a scale hyperprior. *In International Conference on Learning Representations (ICLR)*. (Cité en pages 125, 126 et 143.)

[Baque, 2017] M. BAQUE (2017). *Analyse de scène sonore multi-capteurs : un front-end temps-réel pour la manipulation de scène*. Thèse de doctorat, Université du Maine. (Cité en pages 10 et 11.)

[Baqué *et al.*, 2016] M. BAQUÉ, A. GUÉRIN et M. MELON (2016). Separation of direct sounds from early reflections using the entropy rate bound minimization algorithm. *In Audio Engineering Society Conference (AES)*. Audio Engineering Society. (Cité en page 53.)

[Beentjes, 2015] C. H. BEENTJES (2015). Quadrature on a spherical surface. (Cité en page 12.)

[Beerends *et al.*, 2013] J. G. BEERENDS, C. SCHMIDMER, J. BERGER, M. OBERMANN, R. ULLMANN, J. POMY et M. KEYHL (2013). Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment. *Journal of the Audio Engineering Society*, 61(6):366–384. (Cité en page 33.)

[Berge et Barrett, 2010] S. BERGE et N. BARRETT (2010). High angular resolution planewave expansion. *In International Symposium on Ambisonics and Spherical Acoustics*, pages 6–7. (Cité en page 48.)

[Biswas et Jia, 2020] A. BISWAS et D. JIA (2020). Audio codec enhancement with generative adversarial networks. *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 356–360. IEEE. (Cité en page 121.)

[Blauert, 1997] J. BLAUERT (1997). *Spatial hearing : the psychophysics of human sound localization*. MIT press. (Cité en pages 13, 25 et 27.)

[Bleidt *et al.*, 2017] R. L. BLEIDT, D. SEN, A. NIEDERMEIER, B. CZELHAN, S. FÜG, S. DISCH, J. HERRE, J. HILPERT, M. NEUENDORF, H. FUCHS *et al.* (2017). Development of the MPEG-H TV audio system for ATSC 3.0. *IEEE Transactions on broadcasting*, 63(1):202–236. (Cité en pages 47 et 49.)

[Bosca *et al.*, 2021] A. BOSCA, A. GUÉRIN, L. PEROTIN et S. KITIĆ (2021). Dilated U-net based approach for multichannel speech enhancement from First-Order Ambisonics recordings. *In European Signal Processing Conference (EUSIPCO)*, pages 216–220. IEEE. (Cité en page 121.)

[Bosi et Goldberg, 2012] M. BOSI et R. E. GOLDBERG (2012). *Introduction to digital audio coding and standards*, volume 721. Springer Science & Business Media. (Cité en page 24.)

[Bouéri et Kyriakakis, 2004] M. BOUÉRI et C. KYRIAKAKIS (2004). Audio signal decorrelation based on a critical band approach. *In Audio Engineering Society Convention 117*. Audio Engineering Society. (Cité en page 101.)

[Breebaart *et al.*, 2005] J. BREEBAART, S. van de PAR, A. KOHLRAUSCH et E. SCHUIJERS (2005). Parametric coding of stereo audio. *EURASIP Journal on Advances in Signal Processing*, 2005(9):1–18. (Cité en page 39.)

[Brettle et Skoglund, 2016] J. BRETTLE et J. SKOGLUND (2016). Open-source spatial audio compression for VR content. *In SMPTE 2016 Annual Technical Conference and Exhibition*, pages 1–9. SMPTE. (Cité en pages 41, 55 et 58.)

[Briand, 2007] M. BRIAND (2007). *Études d'algorithmes d'extraction des informations de spatialisation sonore : application aux formats multicanaux*. Thèse de doctorat, INPG Grenoble. (Cité en page 63.)

[Briand *et al.*, 2006] M. BRIAND, D. VIRETTE et N. MARTIN (2006). Parametric coding of stereo audio based on principal component analysis. *In International Conference on Digital Audio Effects (DAFx)*. (Cité en page 53.)

[Caracalla et Roebel, 2020] H. CARACALLA et A. ROEBEL (2020). Sound texture synthesis using RI spectrograms. *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 416–420. IEEE. (Cité en pages 128 et 146.)

[Chapman *et al.*, 2009] M. CHAPMAN, W. RITSCH, T. MUSIL, J. ZMÖLNIG, H. POMBERGER, F. ZOTTER et A. SONTACCHI (2009). A Standard for Interchange of Ambisonic Signal Sets. Including a file standard with metadata. *In The Ambisonics Symposium, Graz, Austria*. (Cité en page 7.)

[Cheng *et al.*, 2020] Z. CHENG, H. SUN, M. TAKEUCHI et J. KATTO (2020). Learned image compression with discretized gaussian mixture likelihoods and attention modules. *In Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7939–7948. (Cité en page 123.)

[Chinen *et al.*, 2020] M. CHINEN, F. S. C. LIM, J. SKOGLUND, N. GUREEV, F. O'GORMAN et A. HINES (2020). ViSQOL v3 : An Open Source Production Ready Objective Speech and Audio Metric. *In preprint arXiv*. (Cité en pages 33 et 125.)

[Damaske et Wagener, 1969] P. v. DAMASKE et B. WAGENER (1969). Richtungshörversuche über einen nachgebildeten kopf. *Acustica*, 21(1):30–35. (Cité en page 26.)

[Daniel, 2011] A. DANIEL (2011). *Spatial auditory blurring and applications to multichannel audio coding*. Thèse de doctorat, Université Pierre et Marie Curie-Paris VI. (Cité en pages 26, 105 et 110.)

[Daniel, 2000] J. DANIEL (2000). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. Thèse de doctorat, Université Pierre et Marie Curie, Paris, France. (Cité en pages 6, 7, 11 et 13.)

[Daniel, 2003] J. DANIEL (2003). Spatial sound encoding including near field effect : Introducing distance coding filters and a viable, new ambisonic format. *In Audio Engineering Society Conference*. Audio Engineering Society. (Cité en pages 6, 9 et 115.)

[Daniel et Kitić, 2020] J. DANIEL et S. KITIĆ (2020). Time domain velocity vector for retracing the multipath propagation. *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 421–425. IEEE. (Cité en page 19.)

[Del Galdo *et al.*, 2012] G. DEL GALDO, M. TASESKA, O. THIERGART, J. AHONEN et V. PULKKI (2012). The diffuse sound field in energetic analysis. *The Journal of the Acoustical Society of America (JASA)*, 131(3):2141–2151. (Cité en page 45.)

[Delikaris-Manias et Pulkki, 2013] S. DELIKARIS-MANIAS et V. PULKKI (2013). Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*. (Cité en pages 17 et 78.)

[Dietz *et al.*, 2002] M. DIETZ, L. LILJERYD, K. KJORLING et O. KUNZ (2002). Spectral Band Replication, a novel approach in audio coding. *In Audio Engineering Society Convention 112*. Audio Engineering Society. (Cité en pages 36 et 37.)

[Dietz *et al.*, 2015] M. DIETZ, M. MULTRUS, V. EKSLER, V. MALENOVSKY, E. NORVELL, H. POBLOTH, L. MIAO, Z. WANG, L. LAAKSONEN, A. VASILACHE *et al.* (2015). Overview of the EVS codec architecture. *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5698–5702. IEEE. (Cité en pages 37, 38 et 104.)

[EBU - TECH 3253, 2008] EBU - TECH 3253 (2008). Sound Quality Assessment Material recordings for subjective tests. (Cité en page 153.)

[Edwige, 2014] R. EDWIGE (2014). Compte-rendu de captation (Rapport projet BiLi) – Fiction CNSMDP : 2 femmes pour un fantôme de René de Obaldia. www.bili-project.org. (Cité en pages 154 et 155.)

[Estrella et Plogsties, 2017] J. ESTRELLA et J. PLOGSTIES (2017). Motion-to-Sound Latency Measurement Procedure for VR Sound Reproduction. *In Audio Engineering Society Convention 142*. Audio Engineering Society. (Cité en page 28.)

[Faller et Baumgarte, 2003] C. FALLER et F. BAUMGARTE (2003). Binaural cue coding-part ii : Schemes and applications. *IEEE Transactions on speech and audio processing*, 11(6):520–531. (Cité en page 39.)

[Fletcher, 1940] H. FLETCHER (1940). Auditory patterns. *Reviews of modern physics*, 12(1):47. (Cité en page 23.)

[Fliege et Maier, 1996] J. FLIEGE et U. MAIER (1996). A two-stage approach for computing cubature formulae for the sphere. *In Mathematik 139T, Universitat Dortmund, Fachbereich Mathematik, Universitat Dortmund, 44221*. (Cité en page 42.)

[Gerzon, 1985] M. A. GERZON (1985). Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871. (Cité en pages 6 et 11.)

[Glorot et Bengio, 2010] X. GLOROT et Y. BENGIO (2010). Understanding the difficulty of training deep feedforward neural networks. *In International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256. JMLR Workshop and Conference Proceedings. (Cité en page 130.)

[Golub et Van Loan, 1983] G. H. GOLUB et C. F. VAN LOAN (1983). *Matrix computations, 3rd Ed.* The Johns Hopkins University Press, Baltimore. (Cité en page 12.)

[Gorlow et Marchand, 2012] S. GORLOW et S. MARCHAND (2012). Informed audio source separation using linearly constrained spatial filters. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):3–13. (Cité en page 16.)

[Gorzel *et al.*, 2019] M. GORZEL, A. ALLEN, I. KELLY, J. KAMMERL, A. GUNGORMUSLER, H. YEH et F. BOLAND (2019). Efficient encoding and decoding of binaural sound with resonance audio. *In Audio Engineering Society Conference : Immersive and Interactive Audio*. Audio Engineering Society. (Cité en pages 69, 85 et 111.)

[Hamilton, 1840] W. R. HAMILTON (1840). On a new species of imaginary quantities connected with a theory of quaternions. *Proceedings of the Royal Irish Academy*, 2:424–434. (Cité en page 63.)

[Hanson, 2006] A. HANSON (2006). *Visualizing Quaternions*. Morgan Kaufmann Publishers. (Cité en page 64.)

[Heller *et al.*, 2008] A. HELLER, R. LEE et E. BENJAMIN (2008). Is my decoder ambisonic ? *In Audio Engineering Society Convention 125*. Audio Engineering Society. (Cité en page 13.)

[Herre *et al.*, 1994] J. HERRE, K. BRANDENBURG et D. LEDERER (1994). Intensity stereo coding. *In Audio Engineering Society Convention*. Audio Engineering Society. (Cité en page 39.)

[Herre *et al.*, 2015] J. HERRE, J. HILPERT, A. KUNTZ et J. PLOGSTIES (2015). Mpeg-h 3d audio—the new standard for coding of immersive spatial audio. *IEEE Journal of selected topics in signal processing*, 9(5):770–779. (Cité en pages 47 et 54.)

[Hines *et al.*, 2015] A. HINES, J. SKOGLUND, A. C. KOKARAM et N. HARTE (2015). ViSQOL : an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–18. (Cité en page 33.)

[Hirvonen *et al.*, 2009] T. HIRVONEN, J. AHONEN et V. PULKKI (2009). Perceptual compression methods for metadata in directional audio coding applied to audiovisual teleconference. *In Audio Engineering Society Convention*. Audio Engineering Society. (Cité en page 109.)

[Hoffman *et al.*, 1972] D. K. HOFFMAN, R. C. RAFFENETTI et K. RUEDENBERG (1972). Generalization of Euler Angles to N-Dimensional Orthogonal Matrices. *Journal of Mathematical Physics*, 13(4):528–533. (Cité en page 63.)

[Honda *et al.*, 2007] A. HONDA, H. SHIBATA, J. GYOBA, K. SAITOU, Y. IWAYA et Y. SUZUKI (2007). Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game. *Applied Acoustics*. (Cité en page 25.)

[ISO 226, 2003] ISO 226 (2003). Acoustique - Lignes isosoniques normales. (Cité en page 23.)

[ITU-R BS. 1284, 2019] ITU-R BS. 1284 (2019). General methods for the subjective assessment of sound quality. (Cité en page 30.)

[ITU-R BS.1116, 2015] ITU-R BS.1116 (2015). Methods for the subjective assessment of small impairments in audio systems including Multichannel Sound Systems. (Cité en pages 29 et 55.)

[ITU-R BS.1387, 2001] ITU-R BS.1387 (2001). Method for objective measurments of perceived audio quality. (Cité en page 139.)

[ITU-R BS.1534, 2015] ITU-R BS.1534 (2015). Method for the subjective assessment of intermediate quality level of coding systems. (Cité en pages 29, 68 et 69.)

[ITU-T G.711, 1988] ITU-T G.711 (1988). Pulse Code Modulation (PCM) Of Voice Frequencies. (Cité en page 83.)

[ITU-T P.1401, 2020] ITU-T P.1401 (2020). Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. (Cité en page 28.)

[ITU-T P.862, 2005] ITU-T P.862 (2005). Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs. (Cité en page 32.)

[ITU-T P.863, 2018] ITU-T P.863 (2018). Perceptual objective listening quality prediction. (Cité en pages 32 et 125.)

[Jarrett *et al.*, 2010] D. P. JARRETT, E. A. HABETS et P. A. NAYLOR (2010). 3D source localization in the spherical harmonic domain using a pseudointensity vector. *In European Signal Processing Conference (EUSIPCO)*. (Cité en pages 17 et 78.)

[Jayant et Noll, 1984] N. S. JAYANT et P. NOLL (1984). Digital coding of waveforms : principles and applications to speech and video. *Englewood Cliffs, NJ*, pages 115–251. (Cité en page 36.)

[Johnston et Ferreira, 1992] J. D. JOHNSTON et A. J. FERREIRA (1992). Sum-difference stereo transform coding. *In International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE Computer Society. (Cité en page 38.)

[Kingma et Ba, 2015] D. P. KINGMA et J. BA (2015). Adam : A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*. (Cité en page 129.)

[Kleijn et Paliwal, 1995] W. B. KLEIJN et K. K. PALIWAL (1995). *Speech coding and synthesis*. Elsevier Science Inc. (Cité en page 36.)

[Kronlachner, 2014a] M. KRONLACHNER (2014a). Plug-in suite for mastering the production and playback in surround sound and ambisonics. *In AES Student Design Competition, AES Convention*. (Cité en pages 154 et 155.)

[Kronlachner, 2014b] M. KRONLACHNER (2014b). *Spatial transformations for the alteration of ambisonic recordings*. Thèse de doctorat, University of Music and Performing Arts, Graz, Institute of Electronic Music and Acoustics. (Cité en page 16.)

[Kumar, 2017] S. K. KUMAR (2017). On weight initialization in deep neural networks. *preprint arXiv*. (Cité en page 130.)

[Ladune *et al.*, 2021] T. LADUNE, P. PHILIPPE, W. HAMIDOUCHE, L. ZHANG et O. DÉFORGES (2021). Conditional coding for flexible learned video compression. *International Conference on Learning Representations (ICLR)*. (Cité en pages 123 et 146.)

[Laitinen *et al.*, 2012] M.-V. LAITINEN, T. PIHLAJAMÄKI, C. ERKUT et V. PULKKI (2012). Parametric time-frequency representation of spatial sound in virtual worlds. *ACM Transactions on Applied Perception*, 9(2):1–20. (Cité en page 109.)

[Lindblom *et al.*, 2005] J. LINDBLOM, J. H. PLASBERG et R. VAFIN (2005). Flexible sum-difference stereo coding based on time-aligned signal components. *In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 255–258. IEEE. (Cité en page 39.)

[Lossius et Anderson, 2014] T. LOSSIUS et J. ANDERSON (2014). ATK Reaper : The Ambisonic Toolkit as JSFX plugins. *In International Computer Music Conference (ICMC)*. (Cité en page 101.)

[Mahé *et al.*, 2019a] P. MAHÉ, S. RAGOT et S. MARCHAND (2019a). First-Order Ambisonic Coding with PCA Matrixing and Quaternion-Based Interpolation. *In International Conference on Digital Audio Effects (DAFx)*. (Cité en pages 41 et 75.)

[Mahé *et al.*, 2019b] P. MAHÉ, S. RAGOT et S. MARCHAND (2019b). First-Order Ambisonic Coding with Quaternion-Based Interpolation of PCA Rotation Matrices. *In EAA Spatial Audio Signal Processing Symposium*, pages 7–12. (Cité en page 75.)

[Mahé *et al.*, 2021] P. MAHÉ, S. RAGOT, S. MARCHAND et J. DANIEL (2021). Ambisonic coding with spatial image correction. *In European Signal Processing Conference (EUSIPCO)*. (Cité en page 96.)

[Mahé *et al.*, 2019] P. MAHÉ, S. RAGOT et J. DANIEL (2019). Détermination de corrections à appliquer à un signal audio multicanal, codage et décodage associé. Dépôt de brevet (FR3101741A1), Office européen des brevets. (Cité en page 96.)

[Malham, 1999] D. G. MALHAM (1999). Higher Order Ambisonic Systems for the Spatialisation of Sound. *In The International Computer Music Conference (ICMC), Beijing, China*. Michigan Publishing. (Cité en page 7.)

[Marple, 1980] L. MARPLE (1980). A new autoregressive spectrum analysis algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):441–454. (Cité en page 36.)

[Mccormack et Delikaris-Manias, 2019] L. MCCORMACK et S. DELIKARIS-MANIAS (2019). Parametric first-order ambisonic decoding for headphones utilising the cross-pattern coherence algorithm. *In EAA Spatial Audio Signal Processing Symposium*, pages 173–178. (Cité en page 13.)

[McCormack et Politis, 2019] L. MCCORMACK et A. POLITIS (2019). SPARTA & COMPASS : Real-time implementations of linear and parametric spatial audio reproduction and processing methods. *In Audio Engineering Society Conference on Immersive and Interactive Audio*. (Cité en pages 58, 85 et 111.)

[McCormack *et al.*, 2019] L. MCCORMACK, A. POLITIS et V. PULKKI (2019). Sharpening of Angular Spectra Based on a Directional Re-assignment Approach for Ambisonic Sound-field Visualisation. *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (Cité en pages 78 et 91.)

[McGrath *et al.*, 2019] D. MCGRATH, S. BRUHN, H. PURNHAGEN, M. ECKERT, J. TORRES, S. BROWN et D. DARCY (2019). Immersive audio coding for virtual reality using a metadata-assisted extension of the 3gpp EVS codec. *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 730–734. IEEE. (Cité en page 50.)

[Mh-Acoustics, 2013] MH-ACOUSTICS (2013). EM32 Eigenmike microphone array release notes (v17. 0). *Technical report*. (Cité en page 9.)

[Miller Jr, 1997] R. G. MILLER JR (1997). *Beyond ANOVA : basics of applied statistics*. Chapman & Hall / CRC press. (Cité en page 90.)

[Minnen *et al.*, 2018] D. MINNEN, J. BALLÉ et G. TODERICI (2018). Joint autoregressive and hierarchical priors for learned image compression. *In Neural Information Processing Systems (NIPS)*. (Cité en pages 3, 119, 120, 129, 135 et 136.)

[Miron et Davies, 2018] M. MIRON et M. DAVIES (2018). High frequency magnitude spectrogram reconstruction for music mixtures using convolutional autoencoders. *In International Conference on Digital Audio Effects (DAFx)*, pages 173–180. (Cité en page 121.)

[Moore et Glasberg, 1996] B. C. MOORE et B. R. GLASBERG (1996). A revision of zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345. (Cité en page 25.)

[Moreau, 2006] S. MOREAU (2006). *Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de distance*. Thèse de doctorat, École doctorale de l'université du Maine, Le Mans, France. (Cité en pages 10, 12 et 13.)

[Morgado *et al.*, 2018] P. MORGADO, N. VASCONCELOS, T. LANGLOIS et O. WANG (2018). Self-supervised generation of spatial audio for 360 video. *Neural Information Processing Systems (NeurIPS)*. (Cité en page 123.)

[MPEG, 2013] MPEG (2013). MPEG-H, Submission and Evaluation Procedures for 3D Audio. ISO/IEC JTC1/SC29/WG11/N13633. (Cité en pages 1, 55, 153, 154 et 155.)

[Narbutt *et al.*, 2018] M. NARBUTT, A. ALLEN, J. SKOGLUND, M. CHINEN et A. HINES (2018). AMBIQUAL-a full reference objective quality metric for ambisonic spatial audio. *In International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE. (Cité en pages 33 et 34.)

[Oord *et al.*, 2016] A. v. d. OORD, S. DIELEMAN, H. ZEN, K. SIMONYAN, O. VINYALS, A. GRAVES, N. KALCHBRENNER, A. SENIOR et K. KAVUKCUOGLU (2016). WaveNet : A generative model for raw audio. *In ISCA Speech Synthesis Workshop*. (Cité en page 121.)

[Painter et Spanias, 2000] T. PAINTER et A. SPANIAS (2000). Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4):451–515. (Cité en page 36.)

[Perez-Gracia et Thomas, 2017] A. PEREZ-GRACIA et F. THOMAS (2017). On Cayley's factorization of 4D rotations and applications. *Advances in Applied Clifford Algebras*, 27(1):523–538. (Cité en page 64.)

[Perotin, 2019] L. PEROTIN (2019). *Localisation et rehaussement de sources de parole au format Ambisonique*. Thèse de doctorat, Université de Lorraine. (Cité en page 16.)

[Perotin *et al.*, 2019] L. PEROTIN, R. SERIZEL, E. VINCENT et A. GUERIN (2019). CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):22–33. (Cité en page 19.)

[Perrott et Saberi, 1990] D. R. PERROTT et K. SABERI (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America (JASA)*, 87(4):1728–1731. (Cité en page 26.)

[Politis *et al.*, 2015] A. POLITIS, J. VILKAMO et V. PULKKI (2015). Sector-based parametric sound field reproduction in the spherical harmonic domain. *Journal of Selected Topics in Signal*, 9(5):852–866. (Cité en pages 46 et 47.)

[Pulkki, 1997] V. PULKKI (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466. (Cité en pages 46 et 109.)

[Pulkki, 2007] V. PULKKI (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516. (Cité en page 44.)

[Pulkki *et al.*, 2018] V. PULKKI, S. DELIKARIS-MANIAS et A. POLITIS (2018). *Parametric time-frequency domain spatial audio*. Wiley Online Library. (Cité en pages 3, 18, 19, 44, 46, 47, 107, 109 et 110.)

[Rafaely, 2019] B. RAFAELY (2019). *Fundamentals of spherical array processing*. Springer. (Cité en pages 12, 16, 17 et 78.)

[Ragot et Mahé, 2019] S. RAGOT et P. MAHÉ (2019). Système de codage audio spatialisé avec interpolation et quantification de la rotation. Dépôt de brevet (EP3706119A1), Office européen des brevets. (Cité en page 75.)

[Ragot et Mahé, 2020] S. RAGOT et P. MAHÉ (2020). Codage optimisé d'une information représentative d'une image spatiale d'un signal audio multicanal. Dépôt de brevet, Office européen des brevets. (Cité en page 118.)

[Rämö et Toukomaa, 2011] A. RÄMÖ et H. TOUKOMAA (2011). Voice quality characterization of IETF Opus codec. *In Conference of the International Speech Communication Association (INTERSPEECH)*. (Cité en page 66.)

[Rayleigh, 1907] L. RAYLEIGH (1907). On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232. (Cité en page 26.)

[Rix *et al.*, 2001] A. W. RIX, J. G. BEERENDS, M. P. HOLLIER et A. P. HEKSTRA (2001). Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. *In International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE. (Cité en page 32.)

[Rudzki *et al.*, 2019] T. RUDZKI, I. GOMEZ-LANZACO, P. HENING, J. SKOGLUND, T. MCKENZIE, J. STUBBS, D. MURPHY et G. KEARNEY (2019). Perceptual evaluation of bitrate compressed ambisonic scenes in loudspeaker based reproduction. *In Audio Engineering Society Conference on Immersive and Interactive Audio*. (Cité en page 72.)

[Schroeder *et al.*, 1979] M. R. SCHROEDER, B. S. ATAL et J. HALL (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, 66(6):1647–1652. (Cité en page 36.)

[Shen *et al.*, 2018] J. SHEN, R. PANG, R. J. WEISS, M. SCHUSTER, N. JAITLY, Z. YANG, Z. CHEN, Y. ZHANG, Y. WANG, R. SKERRV-RYAN *et al.* (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *In nternational Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE. (Cité en page 132.)

[Shinn-Cunningham, 2005] B. G. SHINN-CUNNINGHAM (2005). Influences of spatial cues on grouping and understanding sound. *In Forum Acusticum*, volume 29. (Cité en page 27.)

[Shinn-Cunningham *et al.*, 2001] B. G. SHINN-CUNNINGHAM, J. SCHICKLER, N. KOPČO et R. LITOVSKY (2001). Spatial unmasking of nearby speech sources in a simulated anechoic environment. *The Journal of the Acoustical Society of America (JASA)*, 110(2):1118–1129. (Cité en page 27.)

[Shoemake, 1985] K. SHOEMAKE (1985). Animating rotation with quaternion curves. *In Computer graphics and interactive techniques*, pages 245–254. (Cité en pages 63 et 64.)

[Skoglund, 2018] J. SKOGLUND (2018). Ambisonics in an Ogg Opus Container. *IETF RFC 8486*. (Cité en pages 15, 43, 53 et 107.)

[Skoglund et Valin, 2020] J. SKOGLUND et J.-M. VALIN (2020). Improving Opus low bit rate quality with neural speech synthesis. *Conference of the International Speech Communication Association (INTERSPEECH)*. (Cité en page 122.)

[Sloan *et al.*, 2017] C. SLOAN, N. HARTE, D. KELLY, A. C. KOKARAM et A. HINES (2017). Objective assessment of perceptual audio quality using ViSQOLAudio. *IEEE Transactions on Broadcasting*, 63(4):693–705. (Cité en page 33.)

[Stevens *et al.*, 1937] S. S. STEVENS, J. VOLKMANN et E. B. NEWMAN (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America (JASA)*, 8(3):185–190. (Cité en page 25.)

[Suzuki *et al.*, 2003] Y. SUZUKI, V. MELLERT, U. RICHTER, H. MØLLER, L. NIELSEN, R. HELLMAN, K. ASHIHARA, K. OZAWA et H. TAKESHIMA (2003). Precise and full-range determination of two-dimensional equal loudness contours. *Tohoku University, Japan*. (Cité en page 23.)

[Tancerel *et al.*, 2000] L. TANCEREL, S. RAGOT, V. T. RUOPPILA et R. LEFEBVRE (2000). Combined speech and audio coding by discrimination. *In 2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No. 00EX421)*, pages 154–156. IEEE. (Cité en page 37.)

[Thiede *et al.*, 2000] T. THIEDE, W. C. TREURNIET, R. BITTO, C. SCHMIDMER, T. SPORER, J. G. BEERENDS et C. COLOMES (2000). PEAQ - The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29. (Cité en page 33.)

[Umesh *et al.*, 1999] S. UMESH, L. COHEN et D. NELSON (1999). Fitting the Mel scale. *In International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 217–220. IEEE. (Cité en page 103.)

[Valin *et al.*, 2016] J.-M. VALIN, G. MAXWELL, T. B. TERRIBERRY et K. VOS (2016). High-quality, low-delay music coding in the opus codec. *journal of the Audio Engineering Society*. (Cité en page 120.)

[Valin et Skoglund, 2019] J.-M. VALIN et J. SKOGLUND (2019). A real-time wideband neural vocoder at 1.6 kb/s using LPCNet. *Conference of the International Speech Communication Association (INTERSPEECH)*. (Cité en pages 122, 130 et 132.)

[Valin *et al.*, 2012] J.-M. VALIN, K. VOS et T. TERRIBERRY (2012). Definition of the Opus Audio Codec. *IETF RFC 6716*. (Cité en page 139.)

[Van Trees, 2002] H. L. VAN TREES (2002). *Optimum array processing*. John Wiley & Sons. (Cité en page 15.)

[Veaux *et al.*, 2017] C. VEAUX, J. YAMAGISHI, K. MACDONALD *et al.* (2017). Superseded-CSTR VCTK corpus : English multi-speaker corpus for cstr voice cloning toolkit. (Cité en pages 120 et 130.)

[Wang *et al.*, 2003] Z. WANG, E. P. SIMONCELLI et A. C. BOVIK (2003). Multiscale structural similarity for image quality assessment. *In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee. (Cité en page 125.)

[Ward et Abhayapala, 2001] D. B. WARD et T. D. ABHAYAPALA (2001). Reproduction of a plane-wave sound field using an array of loudspeakers. *IEEE Transactions on speech and audio processing*, 9(6):697–707. (Cité en page 157.)

[Wolters *et al.*, 2003] M. WOLTERS, K. KJORLING, D. HOMM et H. PURNHAGEN (2003). A closer look into MPEG-4 High Efficiency AAC. *In Audio Engineering Society Convention 115*. Audio Engineering Society. (Cité en page 37.)

[Zamani *et al.*, 2017] S. ZAMANI, T. NANJUNDASWAMY et K. ROSE (2017). Frequency domain singular value decomposition for efficient spatial audio coding. *In Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 126–130. (Cité en pages 48, 49, 50, 54 et 62.)

[Zen *et al.*, 2019] H. ZEN, V. DANG, R. CLARK, Y. ZHANG, R. J. WEISS, Y. JIA, Z. CHEN et Y. WU (2019). LibriTTS : A corpus derived from LibriSpeech for text-to-speech. *Conference of the International Speech Communication Association (INTERSPEECH)*. (Cité en page 120.)

[Zhen *et al.*, 2020] K. ZHEN, M. S. LEE, J. SUNG, S. BEACK et M. KIM (2020). Psychoacoustic Calibration of Loss Functions for Efficient End-to-End Neural Audio Coding. *IEEE Signal Processing Letters*, 27:2159–2163. (Cité en page 145.)

[Zotter et Frank, 2012] F. ZOTTER et M. FRANK (2012). All-round ambisonic panning and decoding. *Journal of the Audio Engineering Society*, 60(10):807–820. (Cité en page 12.)

[Zotter *et al.*, 2014] F. ZOTTER, M. FRANK, M. KRONLACHNER et J.-W. CHOI (2014). Efficient phantom source widening and diffuseness in ambisonics. *EAA Joint Symposium on Auralization and Ambisonics*. (Cité en page 101.)

[Zotter *et al.*, 2011] F. ZOTTER, M. FRANK, G. MARENTAKIS et A. SONTACCHI (2011). Phantom source widening with deterministic frequency dependent time delays. *In International Conference on Digital Audio Effects (DAFx)*. (Cité en page 101.)

[Zwicker et Fastl, 2013] E. ZWICKER et H. FASTL (2013). *Psychoacoustics : Facts and Models*, volume 22. Springer Science & Business Media. (Cité en pages 22, 23 et 24.)

Codage Ambisonique pour les Communications Immersives

Résumé :

Cette thèse s'inscrit dans le contexte de l'essor des contenus immersifs. Depuis quelques années, les technologies de captation et de restitution sonore immersive se sont développées de manière importante. Ce nouveau contenu a fait naître le besoin de créer de nouvelles méthodes dédiées à la compression audio spatialisée, notamment dans le domaine de la téléphonie et des services conversationnels. Il existe plusieurs manières de représenter l'audio spatialisé, dans cette thèse nous sommes intéressés à l'ambisonie d'ordre 1. Dans un premier temps, nos travaux ont porté sur la recherche d'une solution pour améliorer le codage multimono. Cette solution consiste en un traitement en amont du codec multimono pour décorréler les signaux des composantes ambisoniques. Une attention particulière a été portée à la garantie de continuité du signal entre les trames et à la quantification des métadonnées spatiales. Dans un second temps, nous avons étudié comment utiliser la connaissance de la répartition de l'énergie du signal dans l'espace, aussi appelée image spatiale, pour créer de nouvelles méthodes de codage. L'utilisation de cette image spatiale a permis d'élaborer deux méthodes de compression. La première approche proposée est basée sur la correction spatiale du signal décodés. Cette correction se base sur la différence entre les images spatiales du signal d'origine et du signal décodés pour atténuer les altérations spatiales. Ce principe a été étendu dans une seconde approche à une méthode de codage paramétrique. Dans une dernière partie de cette thèse, plus exploratoire, nous avons étudié une approche de compression par réseaux de neurones en nous inspirant de modèles de compression d'images par auto-encodeur variationnel.

Mots clés : ambisonie, codage audio multicanal, ambisonie de premier ordre, correction de l'image spatiale, codage paramétrique, compression par réseau de neurones, auto-encodeur variationnel.

Ambisonic Coding for Immersive Communications

Summary :

This thesis takes place in the context of the spread of immersive content. For the last couple of years, immersive audio recording and playback technologies have gained momentum and have become more and more popular. New codecs are needed to handle those spatial audio formats, especially for communication applications. There are several ways to represent spatial audio scenes. In this thesis, we focused on First Order Ambisonic. The first part of our research focused on improving multi-mono coding by decorrelated each ambisonic signal component before the multi-mono coding. To guarantee signal continuity between frames, efficient quantization new mechanisms are proposed. In the second part of this thesis, we proposed a new coding concept using a powermap to recreate the original spatial image. With this concept, we proposed two compressing methods. The first one is a post-processing focused on limiting the spatial distortion of the decoded signal. The spatial correction is based on the difference between the original and the decoded spatial image. This post-processing is later extended to a parametric coding method. The last part of this thesis presents a more exploratory method. This method studied audio signal compression by neural networks inspired by image compression models using variational autoencoders.

Keywords : ambisonics, first-order ambisonics, multichannel audio coding, spatial correction, parametric coding, neural network compression, variational autoencoder.

L3i - Laboratoire Informatique Image et Interaction
Avenue Michel Crépeau
17042 LA ROCHELLE Cedex 1 - France